Pairwise AUC Optimization Needs Corrective Power: A Unified View

Jia Chen[1] **Zhiyong Yang**[1,*] **Shilong Bao**[2] **Qianqian Xu**[2] **Qingming Huang**[1,*]
[1]School of Computer Science and Technology, University of Chinese Academy of Sciences
[2]Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
chenjia26@mails.ucas.ac.cn, {yangzhiyong21,qmhuang}@ucas.ac.cn
{baoshilong,xuqianqian}@ict.ac.cn

Abstract

Direct pairwise AUC-surrogate training is attractive because it trains toward the ranking metric used at test time, especially under class imbalance. This creates a direct alignment between the training objective and the evaluation metric, but such metric alignment alone does not guarantee that stochastic training remains corrective. We study this gap through *Corrective Power*: whether pairwise mini-batch updates produce stronger and more reliable improvements for more severe positive-negative ranking errors. Our main observation is that direct pairwise training can enter regions where severe errors remain common but no longer produce useful corrective motion. In such regions, even the mini-batch noise that would normally help SGD move along the steepest severe-error correction direction can become much weaker. In this paper, we seek a theoretical explanation for this by exploring the degenerate property of the U-statistics formed from the pairwise gradients therein. Our results provide a mechanistic explanation for several empirical patterns: (a) squared-hinge loss is often an effective AUC surrogate; (b) cross-entropy warm-up can help models that enter the AUC phase from a poor starting region; and (c) parameter-efficient fine-tuning from a pretrained model can reduce that need by starting closer to a favorable ranking geometry. Experiments on 8 CV tasks and 8 NLP tasks, using ResNet50, DenseNet121, DistilBERT-base, and Qwen2.5-1.5B-Instruct, support these predictions.

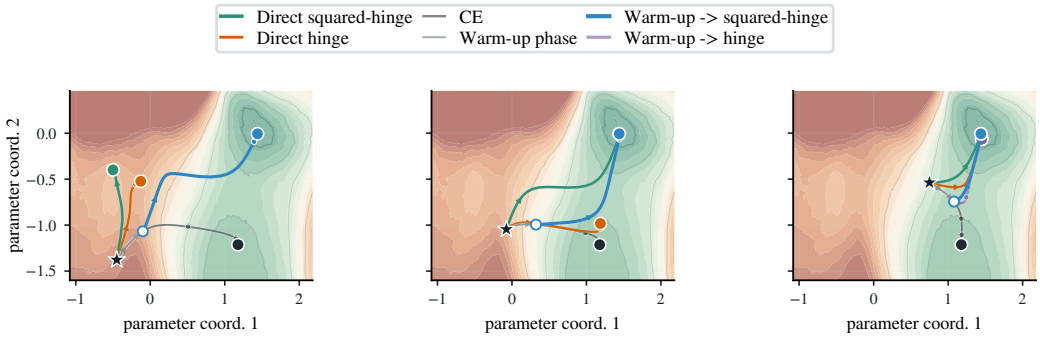
1 Introduction

AUC has a simple but important appeal: it is a threshold-free measure of ranking quality, evaluating whether positives receive higher scores than negatives rather than fixing a single operating point [1–3]. This perspective is especially important under class imbalance, where error rate and pointwise classification losses can be poor proxies for the positive-negative ordering that AUC measures [4–6]. A natural response is to optimize pairwise AUC surrogates directly, as studied in both classical and modern AUC maximization [7–11]. The difficulty we study is that the same pairwise structure that aligns training with AUC can also make the stochastic optimization dynamics fragile.

Several empirical patterns indicate that pairwise AUC training cannot be achieved simply by replacing pointwise empirical risk minimization (ERM). First, direct pairwise AUC-surrogate training from an ordinary starting point can be slow or unstable. Deep-AUC pipelines therefore often use cross-entropy (CE) warm-up or staged optimization before switching to an AUC-family objective [10–15]. Second, the choice of a pairwise surrogate can determine whether training makes progress once the AUC phase begins. On difficult tasks, squared-hinge-style AUC margins have shown strong

*: corresponding authors

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Bad initialization **(b)** Slightly better initialization **(c)** Good initialization

Figure 1: **AUC dynamics comparison under different initialization.** The background color shows AUC on a fixed two-dimensional parameter slice. Curves are simulated training paths; stars and large circles mark starts and final iterates. Panels **(a)**, **(b)**, and **(c)** use bad, slightly better, and good initializations. Details are in Appendix B.1.

empirical performance. A deep AUC system built around this margin achieved first place in the Stanford CheXpert competition [11]. Third, the dependence on warm-up becomes much weaker when performing parameter-efficient fine-tuning (*e.g.*, using LoRA or adapters) for LLMs [16–18]. Taken together, these patterns highlight the difficulty of training: the model may already contain many ranking errors, but the next pairwise update might not focus on the worst ones. A useful explanation must therefore look beyond objective alignment and ask whether the stochastic update still allocates its corrective effort to severely misranked pairs. Rather than treating the observations as separate heuristics or implementation choices, we argue that they are symptoms of one mechanism.

Figure 1 checks the training dynamics with the AUC landscape fixed while the initialization and update rule vary. In this controlled simulation, from a bad start, direct pairwise training moves weakly, and CE warm-up reaches a basin where squared-hinge succeeds. From a slightly better start, direct squared-hinge can already make corrective progress while hinge remains weak. From a good start, both hinge and squared-hinge can work without a warm-up. This motivates *Corrective Power*: reliable direct pairwise training requires severe ranking errors not merely to incur pairwise loss, but to generate finite-batch updates that actively push the model toward correcting them.

This paper examines under what conditions such corrective power becomes weak. The main clue is the useful motion along a severe-error correction direction, meaning a direction that locally decreases the loss contributed by severely misranked pairs. In a **bad region**, many severely misranked pairs do not automatically create such motion. Their gradient contributions may cancel, point away from the correction direction, or receive too little surrogate weight. Along this direction, SGD has a projected mean drift and finite-batch noise. In a bad region, severe errors remain common, but the projected mean drift is weak. Then progress depends on the noise in the same direction. For pairwise AUC training, the U-statistic structure can reduce this useful projected variance from order B^{-1} to order B^{-2}, changing the projected escape-time scale from order B to order B^2.

Based on this, we figure out two ways to enter or leave the bad regions. The first is the choice of surrogate losses. A useful pairwise loss should put enough update mass on severe active violations while ignoring already separated pairs. Squared-hinge has this profile within the standard surrogate family. The second is changing where the pairwise dynamics begin. Warm-up and pretrained parameter-efficient fine-tuning exactly serve this purpose. They can therefore help avoid the poor-start issue.

In short, this paper uses *Corrective Power* to explain when direct pairwise AUC-surrogate training works and when it loses corrective strength. The contributions are:

1. **A mechanism for loss of Corrective Power.** We define bad regions through two observable ingredients: many severe positive-negative misrankings and weak projected motion along the steepest severe-error correction direction (Section 2).
2. **A weight-based explanation of surrogate behavior.** We analyze standard pairwise losses by examining the weights they assign to margins during updates. Squared-hinge is well matched

to difficult active pairs within this family: it increases correction with violation severity while turning off easy pairs (Section 3).

3. **Understanding warm-up and pretraining.** The corrective-power view predicts that CE warm-up can help when the AUC-surrogate phase starts in a poor region, while pretrained low-dimensional trainable subspaces can make a separate warm-up less necessary. We test these predictions on computer vision and NLP tasks with ResNet, DenseNet, DistilBERT, and Qwen models (Sections 4 and 5).

2 Bad Regions in the Stochastic Dynamics of Pairwise AUC Optimization

We now formalize the failure mode from the introduction. Let $f_\theta : \mathcal{X} \to \mathbb{R}$ be the scoring function. Let $X^+ \sim P^+$ and $X^- \sim P^-$ be independent positive and negative examples. We use the positive-margin convention

$$m_\theta(x^+, x^-) = f_\theta(x^+) - f_\theta(x^-).$$

For a pairwise surrogate ℓ, define

$$R_\ell(\theta) = \mathbb{E}[\ell(m_\theta(X^+, X^-))], \qquad w_\ell(x) = -\ell'(x).$$

Although training minimizes the surrogate risk R_ℓ, the quantity we ultimately seek to reduce is the hard pairwise error $\mathbb{P}(m_\theta(X^+, X^-) \le 0)$, with particular attention to its severe-error subset. For sampled examples, let x_i^+ and x_j^- denote the i-th sampled positive example and the j-th sampled negative example, respectively, and write

$$m_{ij} = f_\theta(x_i^+) - f_\theta(x_j^-), \quad g_i^+ = \nabla_\theta f_\theta(x_i^+), \quad g_j^- = \nabla_\theta f_\theta(x_j^-).$$

Given B_+ positives and B_- negatives in a batch, the pairwise mini-batch update is

$$\widehat{g}^{\ell}_{B_+,B_-}(\theta) = \frac{1}{B_+ B_-} \sum_{i=1}^{B_+} \sum_{j=1}^{B_-} w_\ell(m_{ij})(g_j^- - g_i^+), \tag{1}$$

With fresh mini-batches, the estimator is unbiased in the sense that $\mathbb{E}[\widehat{g}^{\ell}_{B_+,B_-}(\theta)] = \nabla R_\ell(\theta)$.

Pairwise SGD updates the parameters by

$$\theta_{t+1} = \theta_t - \eta \widehat{g}^{\ell}_{B_+,B_-}(\theta_t),$$

where $\widehat{g}^{\ell}_{B_+,B_-}(\theta_t)$ is a stochastic estimate of $\nabla R_\ell(\theta_t)$. Thus the update direction has two parts: the mean descent direction $-\nabla R_\ell(\theta_t)$ and the finite-batch noise around it. Here, the finite-batch noise is the mini-batch gradient fluctuation $\widehat{g}^{\ell}_{B_+,B_-}(\theta_t) - \nabla R_\ell(\theta_t)$ relative to the full-sample gradient. We ask whether these two parts still give useful motion in a direction that reduces severe ranking errors.

Here, *local correction direction* is terminology introduced in this paper for a direction that locally reduces the loss contributed by severely misranked pairs. We are now ready to measure the useful motion along the local correction direction $\boldsymbol{v}$. Along this direction, the mean drift is

$$d_{\boldsymbol{v}}(\theta) = \langle -\nabla R_\ell(\theta), \boldsymbol{v} \rangle.$$

If $d_{\boldsymbol{v}}(\theta)$ is large, SGD can progress even with small noise. Many severe misranked pairs do not by themselves make $d_{\boldsymbol{v}}(\theta)$ large. The mean drift is a weighted directional average, and these terms may not add up along $\boldsymbol{v}$. The hard case is when severe errors remain common but $d_{\boldsymbol{v}}(\theta)$ is small. Then progress depends on the finite-batch noise along the same direction.

The next question is which direction to measure. We focus on hard, severely misranked pairs. To isolate this part, we introduce an intermediate loss $\Phi_{M,\ell}(\theta)$. For $M > 0$, define

$$\Phi_{M,\ell}(\theta) = \mathbb{E}\big[\mathbf{1}\{m_\theta(X^+, X^-) \le -M\}\, \ell(m_\theta(X^+, X^-))\big].$$

This is the severe-error part of the current pairwise objective. At a reference point θ_0, if $\nabla \Phi_{M,\ell}(\theta_0) \neq 0$, set

$$\boldsymbol{v}_{\mathrm{sev}}(\theta_0) = -\frac{\nabla \Phi_{M,\ell}(\theta_0)}{\|\nabla \Phi_{M,\ell}(\theta_0)\|_2}.$$

The direction is fixed before projected drift or variance is measured. It is the steepest local direction for reducing the hard pairs, which makes perturbations along $\boldsymbol{v}_{\mathrm{sev}}$ a targeted probe of local corrective motion. **Motion in other directions may change easy pairs or other parts of the objective, but it does not directly reduce the loss from hard examples.** We therefore measure corrective motion along $\boldsymbol{v}_{\mathrm{sev}}$ and interpret finite interventions in light of the step size, local curvature, and subsequent trajectory. See Appendix B.5 for the implementation details.

We now study the finite-batch noise along the fixed correction direction $\boldsymbol{v}$. For this direction, define the scalar pairwise-gradient kernel

$$\psi_{\theta,\boldsymbol{v}}(x^+, x^-) = \big\langle w_\ell(m_\theta(x^+, x^-))\big(\nabla_\theta f_\theta(x^-) - \nabla_\theta f_\theta(x^+)\big), \boldsymbol{v}\big\rangle .$$

For the sampled positive-negative pair indexed by (i, j), define

$$\phi_{\theta,\boldsymbol{v}}(i, j) = \psi_{\theta,\boldsymbol{v}}(x_i^+, x_j^-) = \big\langle w_\ell(m_{ij})(g_j^- - g_i^+), \boldsymbol{v}\big\rangle .$$

A two-sample U-statistic averages a kernel over all sampled positive-negative pairs; its Hoeffding decomposition separates the resulting fluctuation into positive-example, negative-example, and residual pair-interaction components. In the equal-size case, write $\widehat{g}_B^\ell(\theta) = \widehat{g}_{B,B}^\ell(\theta)$. Then $\langle \widehat{g}_B^\ell(\theta), \boldsymbol{v}\rangle$ is the average of $\phi_{\theta,\boldsymbol{v}}(i, j)$, equivalently a two-sample U-statistic built from $\psi_{\theta,\boldsymbol{v}}$. The unequal-size case only changes the normalizing factors: the two first-order terms scale as B_+^{-1} and B_-^{-1}, and the residual interaction term scales as $(B_+B_-)^{-1}$. All conclusions below remain the same.

The next result identifies the variance structure by decomposing the first-order Hoeffding projections and the residual pair-interaction term [19, 20]. The first two components are the leading B^{-1} stochastic term. They are the only Hoeffding terms that contribute at order B^{-1} below.

Proposition 1 (Projected Hoeffding variance split)

Fix θ and a unit direction $\boldsymbol{v}$. Assume $\mathbb{E}[\psi_{\theta,\boldsymbol{v}}(X^+, X^-)^2] < \infty$. Let $\psi_{\theta,\boldsymbol{v}}^{(1,0)}$, $\psi_{\theta,\boldsymbol{v}}^{(0,1)}$, and $\psi_{\theta,\boldsymbol{v}}^{(1,1)}$ be the positive-example, negative-example, and residual Hoeffding components of $\psi_{\theta,\boldsymbol{v}}$. Then, with B positives and B negatives,

$$\mathrm{Var}\big(\langle \widehat{g}_B^\ell(\theta), \boldsymbol{v}\rangle\big) = \frac{\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,0)}(X^+)) + \mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(0,1)}(X^-))}{B} + \frac{\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,1)}(X^+, X^-))}{B^2}.$$

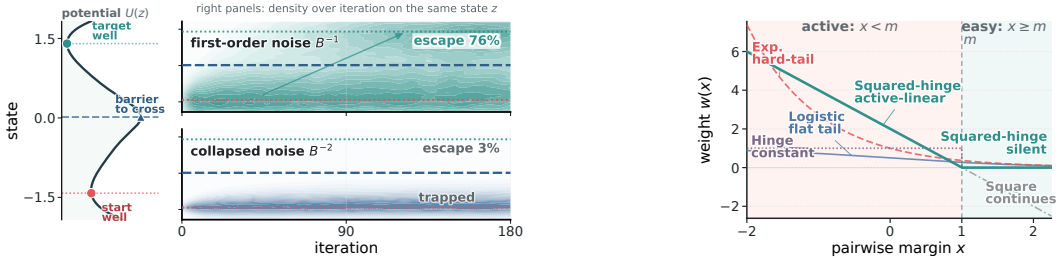
Thus the projected variance has an order-B^{-1} term unless the first-order components vanish; exact order-B^{-2} scaling additionally requires a nonzero residual pair-interaction variance.

See Appendix D for the proof. In ordinary mini-batch behavior, replacing one example changes the projected update at the order B^{-1} variance scale. Pairwise AUC mini-batches have a second possibility: the two first-order projections can vanish, leaving only the weaker pair-interaction randomness at order B^{-2}. Proposition 9 gives a sufficient condition: the severe-truncated row and column effects are nearly constant across examples; a large severe-misranking mass alone does not imply collapse. Partial compression leaves a small but nonzero order-B^{-1} coefficient, so real checkpoints can move toward without attaining exact B^{-2} behavior. Thus the update may remain noisy at the raw pair level while being weak in the parameter direction needed for hard example correction. Appendix D gives the local argument, and Appendix G gives Transformer realizations with the same projected bad-component structure.

Recall that if the projected descent drift $-\mathbb{E}\big[\langle \widehat{g}_B^\ell(\theta), \boldsymbol{v}\rangle\big]$ is weak, then **the remaining mini-batch variance is what can move the iterate along that direction**. When the projected conditional second moment changes from order B^{-1} to order B^{-2}, the time needed to move a fixed projected distance should increase substantially. The next corollary makes this projected escape-time calculation explicit.

Corollary 1 (Projected escape-time scale)

Fix a unit direction $\boldsymbol{v}$, and let $\tau_{\boldsymbol{v},r} = \inf\{t \geq 0 : |\langle \theta_t - \theta_0, \boldsymbol{v}\rangle| \geq r\}$ be the first time the stopped SGD trajectory moves distance r along $\boldsymbol{v}$. In the bad-region application, $\boldsymbol{v}$ is the optimistic steepest severe-error correction direction at the reference point. Under the weak-drift and bounded-second-moment conditions stated precisely in Appendix H, if the projected



(a) Escape-time simulation

(b) Canonical surrogate weights

Figure 2: **Noise scale and surrogate weights illustrate two ingredients of corrective motion. (a)** A one-dimensional projected escape simulation comparing a first-order B^{-1} variance regime with a collapsed B^{-2} variance regime. The left potential and the right heatmaps use the same state axis z. **(b)** Update weights for five standard pairwise surrogates as functions of the pairwise margin x. Simulation details are in Appendix B.2.

increment has conditional second-moment scale ν_B^2, then $\mathbb{E}[\tau_{\boldsymbol{v},r}] = \Omega(r^2/\nu_B^2)$. With fixed step size η, Proposition 1 makes the projected gradient variance contribute at scale η^2/B when the first-order term is present and at scale η^2/B^2 after first-order projection collapse. Thus the optimistic projected escape-time lower-bound scale changes from order B to order B^2.

Appendix H proves the stopped-process escape-time bound. Figure 2(a) illustrates it with a one-dimensional potential $U(z)$: the left curve marks the start well, barrier, and target well, while the right heatmaps show where simulated iterates concentrate on the same state axis. With first-order B^{-1} noise, 76% of runs cross the barrier by the plotted horizon; with collapsed B^{-2} noise, only 3% do. This echoes the main claim: a bad region is not merely high-error; it is a region where severe errors remain but the current update loses projected motion along a severe-error correction direction.

Above all, we can now identify a bad region in the following.

A **bad region** is not merely a high-error region and need not be a local minimum. It is a high-severe-error region where SGD loses useful motion along the severe-error correction direction $\boldsymbol{v}_{\mathrm{sev}}$: **the projected mean drift is weak, and the leading projected noise term is small or, in the exact case, vanishes**. Because these ingredients are defined independently of escape time, the result above quantifies their local consequence: under the stated conditions, variance contraction slows motion over a fixed distance along $\boldsymbol{v}_{\mathrm{sev}}$, even though the full trajectory may still progress through other directions.

We next ask how the update can be changed to enter or leave a bad region.

3 Surrogate Losses and Their Weight Geometry

The first way to change the update is to change the surrogate loss. The surrogate loss choice changes the weight function in the pairwise gradient. For a differentiable surrogate, the key object is the scalar pairwise weight $w_\ell(x) = -\ell'(x)$. More severe ranking mistakes have more negative x. For a margin $m > 0$, active pairs satisfy $x < m$, and easy pairs satisfy $x \geq m$. A corrective weight profile emphasizes severe active mistakes while stopping already separated pairs from consuming update mass.

Figure 2(b) compares popular surrogates in this sense. The relevant quantity is the allocation of update mass across pairs. The comparison is therefore organized by update weight, because the weight determines which pairs enter the mini-batch gradient and how strongly they enter. The five canonical losses lead to visibly different structures. Hinge is constant on active pairs, logistic flattens in the severe tail, exponential over-amplifies the hard tail, and square keeps nonzero easy-side weight. Within this five-loss comparison, squared-hinge realizes the desired combination of active-side severity sensitivity and easy-side silence: it grows linearly on active pairs while becoming exactly silent for pairs that already satisfy the margin. For the active-batch comparison below, a *fully active*

mini-batch means that every sampled positive-negative pair violates the margin. Appendix E gives the exact weights, the within-family characterization, and the active-batch decomposition used below.

Theorem 1 (Informal surrogate-weight geometry)

Among logistic, exponential, square, hinge, and squared-hinge, squared-hinge is the only one in this canonical set that both ignores easy pairs and upweights active violations linearly by severity. On a fully active mini-batch, hinge contains only the class-mean score-gradient direction. Squared-hinge contains this direction plus a severity-dependent within-class correction. Thus squared-hinge, and more generally any surrogate with the same severity-increasing, easy-pair-suppressing weight structure, can retain corrective motion outside the hinge direction.

The theorem gives a loss-level reason for the empirical gap between hinge and squared-hinge. Hinge ignores easy pairs, but on a fully active mini-batch it treats all active violations equally, so the update is driven only by the two class-mean score-gradient averages. If that class-mean direction is weak or poorly aligned with the direction that would correct the severe misranking pairs, hinge has no additional severity-dependent term to use. Squared-hinge also ignores easy pairs, but its active-side weight grows with violation size, letting the mini-batch update react to the distribution of active margins inside the positive and negative classes. Appendix E gives the calculation showing where this within-class correction enters.

The other four canonical losses considered here miss either the active-side severity weighting or the easy-pair silence that make squared-hinge useful in this comparison. Logistic has bounded hard-tail weights, so increasingly severe violations eventually receive almost the same weight. Exponential has the opposite problem: it can concentrate the update on the extreme tail. Square has the same affine active-side response as squared-hinge, but it continues to spend update mass on already separated pairs. Appendix E gives the hard-tail, easy-side, active-batch, and projected-variance calculations behind these comparisons. In a fully active setting, hinge can have zero projected mini-batch variance, while squared-hinge keeps a nonzero order-B^{-1} term under class-score variation. Thus the surrogate can change both deterministic correction and projected stochastic motion along the steepest correction direction.

4 Changing Where Pairwise AUC Optimization Begins

Section 3 treated surrogate choice as a way to change the pairwise update. **A second intervention is to change where the AUC phase begins.** This is the role of CE warm-up and pretrained parameter-efficient fine-tuning. CE warm-up can move an ordinary model away from many severe misranking pairs before pairwise training starts. A pretrained representation with a low-dimensional trainable module can instead start in a region where severe misrankings are already controlled and first-order stochastic motion remains present. This starting-region effect is the second lesson of Figure 1: warm-up helps in the bad-start panel because it changes the starting region of the pairwise phase, while the good-start panel shows that the same pairwise objective can succeed without a separate warm-up when training already begins on the favorable side of the bad-region mechanism.

Define the balanced cross-entropy risk, obtained by averaging the positive and negative classes separately, as $R_{\mathrm{CE}}(\theta) = \mathbb{E}[\log(1 + e^{-f_\theta(X^+)})] + \mathbb{E}[\log(1 + e^{f_\theta(X^-)})]$. For $M > 0$, let $\pi_M(\theta) = \mathbb{P}(m_\theta(X^+, X^-) \le -M)$ be the **misranking mass**. The following result links the pointwise warm-up and the severe-misranking mass.

Proposition 2 (CE risk controls severe misrankings)

For every $M > 0$, $\pi_M(\theta) \le R_{\mathrm{CE}}(\theta)/\log(1 + e^{M/2})$. Consequently, if a CE warm-up endpoint θ_{wu} satisfies $R_{\mathrm{CE}}(\theta_{\mathrm{wu}}) \le \delta \cdot \log(1 + e^{M/2})$, then $\pi_M(\theta_{\mathrm{wu}}) \le \delta$.

The proof appears in Appendix F.1: if a positive is scored far below a negative by at least M, then either the positive score is low or the negative score is high, and CE penalizes these events. Thus CE warm-up can improve the later pairwise AUC-surrogate phase even though CE itself does not optimize a pairwise AUC surrogate. The role of CE warm-up is to reduce the mass of severe misranking pairs before the pairwise phase starts.

This gives a precise condition under which warm-up improves the pairwise starting state. The bound controls the severe-misranking mass at the AUC starting point. When the subsequent AUC trajectory remains near the warm-up endpoint, local continuity of the margins transfers this low severe-misranking mass to the early pairwise phase; Appendix F.1 gives the finite-horizon statement. A poor warm-up endpoint, or a pairwise trajectory that quickly leaves that neighborhood, falls outside this local argument. We therefore treat warm-up as a starting-region intervention.

Pretrained parameter-efficient fine-tuning addresses the same question about starting regions in a different way. Let $\boldsymbol{\xi}$ denote the trainable coordinates of a head, adapter, or LoRA module, with the pretrained base absorbed into $f_{\boldsymbol{\xi}}$. The pairwise gradient identity and Hoeffding split remain valid after replacing θ by $\boldsymbol{\xi}$, because they use only the score map and its gradient with respect to the trainable variable. The relevant question is therefore whether the pretrained starting neighborhood is already favorable for AUC optimization. In this subspace, write $\pi_M(\boldsymbol{\xi}) = \mathbb{P}(m_{\boldsymbol{\xi}}(X^+, X^-) \le -M)$, where $m_{\boldsymbol{\xi}}(x^+, x^-) = f_{\boldsymbol{\xi}}(x^+) - f_{\boldsymbol{\xi}}(x^-)$.

Proposition 3 (Finite-horizon transfer from a favorable pretrained subspace)

Let $\psi_{\boldsymbol{\xi},\boldsymbol{v}}$ be the scalar projected pairwise-gradient kernel from Section 2, with θ replaced by the trainable coordinate $\boldsymbol{\xi}$. Suppose a trainable-subspace neighborhood $\mathcal{U}$ satisfies $\sup_{\boldsymbol{\xi}\in\mathcal{U}} \pi_M(\boldsymbol{\xi}) \le \delta_M$. Suppose also that, for some unit direction $\boldsymbol{v}$, $\mathrm{Var}(\psi^{(1,0)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^-)) \ge \sigma^2_{1,\boldsymbol{v}} > 0$ for $\boldsymbol{\xi} \in \mathcal{U}$. If the direct pairwise AUC-surrogate trajectory remains in $\mathcal{U}$ for the finite horizon of interest, then every iterate on that horizon satisfies $\pi_M(\boldsymbol{\xi}_t) \le \delta_M$ and $\mathrm{Var}(\langle \widehat{\boldsymbol{g}}^{\ell}_B(\boldsymbol{\xi}_t), \boldsymbol{v}\rangle) \ge \sigma^2_{1,\boldsymbol{v}}/B$.

The proposition is a local finite-horizon transfer statement conditional on a verified favorable trainable-subspace neighborhood and trajectory containment; Appendix F gives sufficient score-perturbation, confinement, and local descent checks for this verification. Along such a trajectory, severe misrankings stay controlled and the first-order projected stochastic component remains present, making a separate warm-up stage less necessary.

Under the same LoRA-style fine-tuning protocol, a pretrained base with more favorable starting geometry can place the trainable trajectory closer to a favorable ranking region, reducing the marginal value of a separate CE warm-up; a weaker starting geometry leaves more room for warm-up to help. The experiments test this starting-state explanation across CV and NLP tasks.

5 Experiments

The experiments focus on a specific question: when the evaluation metric is held fixed as AUC, which training choices preserve Corrective Power in finite-batch deep optimization? We compare CE, direct pairwise AUC-surrogate training, and CE-warm-started pairwise AUC-surrogate training on binary CV and NLP ranking tasks using ResNet50 [21], DenseNet121 [22], DistilBERT-base [23], and Qwen2.5-1.5B-Instruct [24]. Table 1 reports the compact DenseNet121 and Qwen direct and warm-up rows with sample standard deviations, while Table 2 gives the corresponding ResNet50 and DistilBERT-base blocks. Figure 3 gives trajectory-level mechanism measurements, and Figure 4 gives aggregate warm-up and weight-allocation summaries. Appendix B gives the data construction, trainable-module details, optimization protocol, and full reporting rules. We use the tables and trajectory measurements to examine four questions about when pairwise AUC training works reliably.

RQ1. Is metric-aligned pairwise training enough by itself?
No. Directly optimizing a pairwise AUC surrogate can improve held-out AUC, but it is not automatically reliable. Table 1 should be read task by task: the CE column is the pointwise baseline, the direct row starts pairwise training immediately, and the warm-up row first runs CE before switching to the specified pairwise surrogate. Bold and underlined entries mark the best and second-best validation-selected test AUC values within each task. This comparison shows that pairwise losses can beat CE, but it also shows failures in which a metric-aligned pairwise objective falls below CE or becomes much worse than other pairwise surrogates. The trajectory evidence gives the same message from inside training: Figure 3(c) plots the logged-batch misranking ratio, namely the fraction of logged positive-negative training pairs with $f_\theta(x^+) - f_\theta(x^-) < 0$. Direct hinge training can keep this ratio high for many pairwise updates, meaning that the loss is aligned with AUC but the trajectory

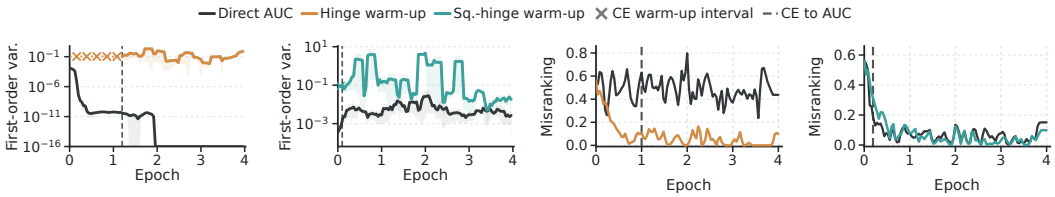
Table 1: **Held-out AUC (%) for direct and CE-warm-started pairwise training.** Representative DenseNet121 and Qwen2.5-1.5B-Instruct results are shown; ResNet50 and DistilBERT-base are reported in Table 2. Direct starts pairwise training immediately; Warm-up first runs CE and then switches to the indicated pairwise loss. Values are mean AUC (%) with sample standard deviations as subscripts; bold and underline mark the best and second-best results within each task.

Task	Mode	CE	Square	Hinge	Exp.	Logistic	Sq.-hinge
DenseNet121							
Apple Disease	Direct	$60.64_{\pm 2.63}$	$61.10_{\pm 1.85}$	$58.81_{\pm 2.46}$	$60.18_{\pm 1.21}$	$60.46_{\pm 3.09}$	**62.69**$_{\pm 4.86}$
	Warm-up	–	$61.10_{\pm 1.85}$	$59.91_{\pm 2.60}$	$61.03_{\pm 1.92}$	$60.54_{\pm 2.02}$	$\underline{62.69}_{\pm 4.86}$
Rice Disease	Direct	$72.45_{\pm 6.57}$	$71.43_{\pm 7.23}$	$70.92_{\pm 7.66}$	$70.61_{\pm 6.81}$	$71.63_{\pm 7.30}$	$71.53_{\pm 7.30}$
	Warm-up	–	$\underline{74.90}_{\pm 6.31}$	$73.98_{\pm 6.26}$	$74.90_{\pm 6.41}$	$71.63_{\pm 7.30}$	**75.71**$_{\pm 8.61}$
APTOS Sev./Mod.	Direct	$82.75_{\pm 0.95}$	$83.83_{\pm 1.41}$	**85.10**$_{\pm 0.76}$	$83.68_{\pm 1.66}$	$84.18_{\pm 1.37}$	$\underline{84.20}_{\pm 1.12}$
	Warm-up	–	$83.96_{\pm 1.04}$	$83.86_{\pm 0.99}$	$83.93_{\pm 1.25}$	$83.76_{\pm 1.47}$	$83.08_{\pm 1.49}$
HAM10000 M/N	Direct	$96.50_{\pm 0.58}$	$96.88_{\pm 0.57}$	$94.13_{\pm 0.81}$	$95.97_{\pm 0.50}$	$95.35_{\pm 0.56}$	$\underline{97.32}_{\pm 0.42}$
	Warm-up	–	$96.81_{\pm 1.02}$	$96.08_{\pm 0.64}$	$96.29_{\pm 0.46}$	$96.51_{\pm 0.74}$	**97.70**$_{\pm 0.21}$
Qwen2.5-1.5B-Instruct							
GoEmotions Annoy.	Direct	$82.97_{\pm 1.27}$	$83.74_{\pm 1.45}$	$83.44_{\pm 1.13}$	$81.16_{\pm 1.63}$	$80.71_{\pm 1.87}$	$\underline{84.89}_{\pm 0.65}$
	Warm-up	–	$84.34_{\pm 0.69}$	$83.27_{\pm 1.03}$	$83.36_{\pm 0.62}$	$83.07_{\pm 0.72}$	**84.92**$_{\pm 0.66}$
GoEmotions Fear	Direct	$95.19_{\pm 1.22}$	$94.93_{\pm 1.02}$	$93.42_{\pm 0.77}$	$93.75_{\pm 1.28}$	$92.68_{\pm 0.76}$	$96.33_{\pm 0.91}$
	Warm-up	–	$\underline{96.69}_{\pm 0.54}$	$94.88_{\pm 0.99}$	$94.28_{\pm 1.26}$	$94.51_{\pm 0.60}$	**96.98**$_{\pm 0.62}$
ANLI R1 E/N	Direct	$86.01_{\pm 0.74}$	$84.87_{\pm 1.87}$	$70.88_{\pm 12.94}$	$79.01_{\pm 1.53}$	$77.82_{\pm 2.72}$	**86.85**$_{\pm 0.73}$
	Warm-up	–	$85.41_{\pm 1.40}$	$83.23_{\pm 1.70}$	$83.44_{\pm 1.42}$	$83.09_{\pm 1.72}$	$\underline{86.45}_{\pm 0.84}$
ANLI R3 C/N	Direct	**72.39**$_{\pm 1.49}$	$72.04_{\pm 0.97}$	$52.02_{\pm 2.09}$	$69.12_{\pm 2.11}$	$69.33_{\pm 1.22}$	$71.99_{\pm 0.94}$
	Warm-up	–	$71.71_{\pm 0.70}$	$70.37_{\pm 1.65}$	$70.80_{\pm 1.21}$	$70.47_{\pm 0.97}$	$\underline{72.15}_{\pm 0.64}$
ANLI R3 E/N	Direct	$\underline{79.98}_{\pm 0.94}$	$78.27_{\pm 0.87}$	$63.97_{\pm 12.49}$	$74.46_{\pm 1.69}$	$75.03_{\pm 1.68}$	$79.72_{\pm 0.44}$
	Warm-up	–	$79.40_{\pm 0.62}$	$77.03_{\pm 1.45}$	$77.86_{\pm 0.88}$	$78.35_{\pm 0.79}$	**80.60**$_{\pm 1.27}$
SNLI E/N	Direct	$96.69_{\pm 0.27}$	$\underline{96.79}_{\pm 0.44}$	$80.58_{\pm 14.39}$	$90.06_{\pm 4.27}$	$91.21_{\pm 2.01}$	$96.64_{\pm 0.33}$
	Warm-up	–	**96.93**$_{\pm 0.31}$	$94.73_{\pm 2.30}$	$94.11_{\pm 1.75}$	$94.60_{\pm 0.71}$	$96.65_{\pm 0.36}$
WANLI C/N	Direct	$\underline{89.81}_{\pm 0.46}$	$89.35_{\pm 0.56}$	$67.57_{\pm 14.73}$	$79.72_{\pm 15.49}$	$85.46_{\pm 2.52}$	$89.77_{\pm 0.60}$
	Warm-up	–	$89.47_{\pm 0.39}$	$85.87_{\pm 3.05}$	$87.48_{\pm 1.11}$	$88.12_{\pm 0.75}$	**89.83**$_{\pm 0.43}$
WANLI E/N	Direct	$86.50_{\pm 0.25}$	**86.61**$_{\pm 0.17}$	$62.20_{\pm 5.34}$	$75.41_{\pm 14.83}$	$71.09_{\pm 16.10}$	$86.40_{\pm 0.75}$
	Warm-up	–	$\underline{86.50}_{\pm 0.20}$	$82.66_{\pm 0.85}$	$83.88_{\pm 1.25}$	$83.90_{\pm 1.38}$	$86.29_{\pm 0.41}$

is not quickly correcting the wrong pairs. The matched intervention in Appendix B.5 provides a direct mechanism check: perturbations confined to $\widehat{\boldsymbol{v}}_{\text{sev}}$ produce markedly more finite-horizon changes than equal-energy isotropic controls. This identifies $\widehat{\boldsymbol{v}}_{\text{sev}}$ as a mechanism-relevant direction while showing that the realized effect depends on the perturbation sequence and subsequent trajectory. The experiments therefore support the paper's central claim that objective–metric alignment must be paired with Corrective Power.

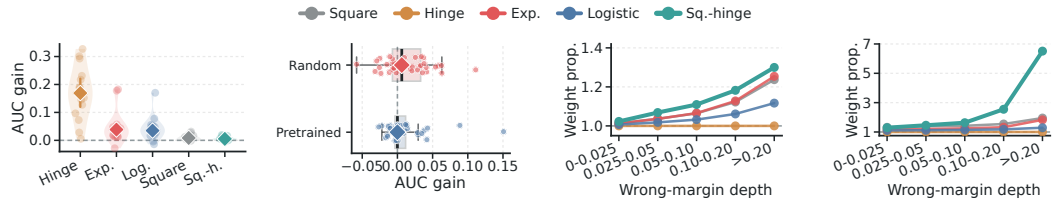
RQ2. Which surrogate most consistently preserves useful correction?
Squared-hinge gives the highest AUC values in the representative summary. Among the displayed DenseNet121 and Qwen rows in Table 1, the best result across CE and all direct/warm-started pairwise settings is usually obtained by squared-hinge, while the second-best result varies across CE and square. The weight plots in Figure 4(c,d) explain how to read this beyond the final AUC numbers. For each logged positive-negative pair, we compute the current margin $m_\theta = f_\theta(x^+) - f_\theta(x^-)$; wrong pairs have $m_\theta < 0$, and more negative intervals correspond to more severe misranking. For an interval I, P_I^{rel} is the proportion of total gradient weight assigned to pairs in I, divided by the frequency of pairs in I. Thus $P_I^{\text{rel}} > 1$ means that the interval receives more update weight than its frequency alone would imply, while $P_I^{\text{rel}} < 1$ means that it is relatively down-weighted. The useful visual signal is therefore not just whether a loss has many wrong pairs, but whether the leftmost wrong-margin intervals receive extra corrective weight. Squared-hinge assigns larger relative weight to more severe wrong-margin intervals, while hinge treats active violations uniformly; this is the empirical weight-profile counterpart of the surrogate analysis. Appendix B.5 gives the finite-batch formula, row/column projection definitions, and logging protocol, and Table 2 reports the corresponding ResNet50 and DistilBERT-base splits.

RQ3. When is CE warm-up helpful?
Warm-up helps most when it changes a poor pairwise starting state. The direct/warm-up rows in Table 1 show that warm-up can give large gains for fragile pairwise settings, especially hinge on harder NLP tasks, but the gains are not uniform across losses or models. Figure 4(a) summarizes this



(a) Hinge variance term **(b)** Sq.-hinge variance term **(c)** Hinge trajectory **(d)** Sq.-hinge trajectory

Figure 3: **Real-run trajectory measures support the Corrective Power explanation.** Four training-trajectory quantities computed from experiment runs. Panels (a,b) show the scaled first-order Hoeffding variance term on SNLI E/N along the steepest severe-error correction direction; shaded bands show one standard error after the same smoothing used for the displayed mean. x marks denote the CE warm-up interval before the pairwise variance term is recorded. Panels (c,d) show ANLI R1 E/N trajectories and mark the CE-to-AUC switch. Definitions are in Appendix B.5.



(a) Warm-up gain by loss **(b)** Warm-up gain by initialization **(c)** Direct weight distribution **(d)** Warm-up weight distribution

Figure 4: **Aggregate real-run measures support the Corrective Power explanation.** Four aggregate quantities computed from the same experiment runs. Panels (a,b) report warm-up gains by loss and initialization, respectively. Panels (c,d) report relative weight proportions over wrong-margin intervals. Definitions are in Appendix B.5.

with warm-up gain, defined as validation-selected test AUC after CE-warm-started pairwise training minus validation-selected test AUC after direct pairwise training. Positive values mean that CE improves the starting point for the later pairwise phase; values near zero mean that the direct pairwise run already starts from, or quickly reaches, a usable region. The largest and most variable gains appear for hinge, matching the view that warm-up is most valuable when direct pairwise optimization enters a poor high-misranking state. Figure 3(a,b) then checks the stochastic side of the same explanation: it records the scaled first-order Hoeffding variance term along the steepest severe-error correction direction. This term measures whether the mini-batch update still has first-order stochastic motion in the favorable projected direction; when it is small, the leading mini-batch component along that direction is weak. Figure 3(c,d) records the logged-batch misranking ratio, which measures how many logged training pairs are still ordered incorrectly. In these trajectory panels, the black curve is direct pairwise AUC-surrogate training, the colored curve is CE warm-up followed by pairwise training, and the vertical dashed line marks the CE-to-AUC switch. The hinge trajectory shows the intended warm-up effect: CE lowers the misranking state before the pairwise phase begins, after which pairwise training starts from an easier region. The squared-hinge trajectory shows a different regime: direct squared-hinge can already reduce misranking quickly and preserve stronger corrective motion, so a separate warm-up has smaller marginal value. Appendix B.5 gives the finite-batch construction and independent logging-batch protocol, and Figure 5 in Appendix B.5 reports the all-loss trajectory counterpart.

Fixed-checkpoint batch-size scaling provides a complementary check of this stochastic explanation. Across APTOS, SNLI, and ANLI-R1, weakly corrective regions show more attenuated severe-component variance scaling than their matched corrective regions, matching the predicted shift toward the residual pair-interaction component as first-order projected motion weakens. Appendix B.5 gives the protocol and complete results.

The two trajectory quantities should be interpreted together. A low misranking ratio alone would show that the run is improving the ordering, but it would not explain whether the pairwise mini-batch update still has the stochastic component emphasized by the theory. The warm-up panels therefore

support the mechanism when both signals move in the expected direction: CE changes the state before the AUC phase, the later pairwise updates operate from a lower-misranking region, and the projected first-order term remains more visible along the steepest correction direction.

RQ4. What changes under pretrained initialization and parameter-efficient tuning? A pretrained start can reduce the need for a separate warm-up stage when it places the trainable trajectory in a favorable region. The Qwen block in Table 1 often has small squared-hinge warm-up gains, and the direct squared-hinge entries are already close to the corresponding warm-up entries on several NLP tasks. Figure 4(b) provides a matched same-backbone comparison between random and ImageNet-pretrained initialization. The aggregate shift and reduced dispersion of warm-up gains under pretrained initialization support that pretraining places optimization closer to a favorable ranking state and reduces the average marginal benefit of an additional warm-up stage. This is consistent with the starting-state explanation: when the pretrained representation and trainable subspace place AUC training closer to a favorable ranking geometry, warm-up has smaller marginal value, although it remains useful in some settings.

Together, the experiments support a conditional prescription: pairwise AUC-surrogate training is most dependable when the surrogate keeps corrective gradient weight on severe wrong pairs, and warm-up or pretraining is most useful when it moves the trajectory into a state where such corrective updates remain available.

6 Conclusion

This paper studies why direct pairwise AUC-surrogate training can still be unreliable despite matching the evaluation metric. The key issue is *Corrective Power*: whether SGD keeps useful motion on severe ranking errors. We identify bad regions where severe misranked pairs remain common, but the projected mean drift along the severe-error correction direction is weak. In these regions, progress depends on the finite-batch noise in the same direction. The pairwise mini-batch gradient is a two-sample U-statistic, so this projected noise has a leading order-B^{-1} term and a residual order-B^{-2} interaction term in the equal-size case. When the leading term vanishes, the useful noise drops to order B^{-2}, and the projected escape-time scale changes from order B to order B^2. This view explains two ways to improve corrective power. First, the surrogate loss controls the gradient weight assigned to severe wrong pairs; squared-hinge is effective because it gives larger weight to severe active violations while ignoring easy pairs. Second, warm-up and PEFT change where pairwise training begins; CE warm-up can move a model away from a poor start, while pretrained PEFT can start closer to a favorable ranking region. The CV and NLP experiments are consistent with these two predictions.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China: 62525212, U23B2051, 62236008, 62441232, 62521007, U21B2038, 62502496, and 62576332, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0680201, in part by Beijing Natural Science Foundation–Xiaomi Innovation Joint Fund (Grant No. 26L030072), in part by the project ZR2025ZD01 supported by Shandong Provincial Natural Science Foundation, and in part by the Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan No. 20250687. The authors declare no competing interests.

References

[1] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[2] Andrew P Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[3] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[4] Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

[5] Aditya Krishna Menon and Robert C Williamson. Bipartite ranking: a risk-theoretic perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016.

[6] Tianbao Yang and Yiming Ying. AUC maximization in the era of big data and AI: A survey. *ACM computing surveys*, 55(8):1–37, 2022.

[7] Lian Yan, Robert H Dodier, Michael Mozer, and Richard H Wolniewicz. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 848–855, 2003.

[8] Alan Herschtal and Bhavani Raskutti. Optimising area under the ROC curve using gradient descent. In *Proceedings of the twenty-first international conference on Machine learning*, page 49, 2004.

[9] Alain Rakotomamonjy. Optimizing Area Under Roc Curve with SVMs. In *ROCAI*, pages 71–80, 2004.

[10] Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic AUC maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*, 2019.

[11] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3020–3029. IEEE, 2021.

[12] Zhuoning Yuan, Zhishuai Guo, Nitesh Chawla, and Tianbao Yang. Compositional Training for End-to-End Deep AUC Maximization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gPvB4pdu_Z.

[13] Tianbao Yang. Deep AUC maximization for medical image classification: Challenges and opportunities. *arXiv preprint arXiv:2111.02400*, 2021.

[14] Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao, and Qingming Huang. When all we need is a piece of the pie: A generic framework for optimizing two-way partial AUC. In *International Conference on Machine Learning*, pages 11820–11829. PMLR, 2021.

[15] Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao, and Qingming Huang. Optimizing Two-Way Partial AUC With an End-to-End Framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10228–10246, August 2023. doi: 10.1109/TPAMI.2022.3185311.

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[18] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[19] Alan J Lee. *U-statistics: Theory and Practice*. Routledge, 2019.

[20] Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2261–2269. IEEE, 2017.

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[24] Qwen:, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2025. URL https://arxiv.org/abs/2412.15115.

[25] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384, 2005.

[26] Harikrishna Narasimhan, Purushottam Kar, and Prateek Jain. Optimizing non-decomposable performance measures: A tale of two classes. In *International Conference on Machine Learning*, pages 199–208. PMLR, 2015.

[27] Tino Werner. A review on instance ranking problems in statistical learning. *Machine Learning*, 111(2):415–463, 2022.

[28] Zhiyong Yang, Qianqian Xu, Shilong Bao, Xiaochun Cao, and Qingming Huang. Learning With Multiclass AUC: Theory and Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7747–7763, November 2022. doi: 10.1109/TPAMI.2021.3101125.

[29] Huiyang Shao, Qianqian Xu, Zhiyong Yang, Peisong Wen, Peifeng Gao, and Qingming Huang. Weighted ROC curve in cost space: Extending AUC to cost-sensitive learning. *Advances in Neural Information Processing Systems*, 36:17357–17368, 2023.

[30] Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. OpenAUC: Towards AUC-oriented open-set recognition. *Advances in Neural Information Processing Systems*, 35:25033–25045, 2022.

[31] Boyu Han, Qianqian Xu, Zhiyong Yang, Shilong Bao, Peisong Wen, Yangbangyan Jiang, and Qingming Huang. AUCSeg: AUC-oriented pixel-level long-tail semantic segmentation. *Advances in Neural Information Processing Systems*, 37:126863–126907, 2024.

[32] Zhaoda Ye, Xiangteng He, and Yuxin Peng. Unsupervised Cross-Media Hashing Learning via Knowledge Graph. *Chinese Journal of Electronics*, 31(6):1081–1091, 2022. doi: 10.1049/cje.2021.00.455.

[33] Ye Tian, Ying Fu, and Jun Zhang. Transformer-Based Under-sampled Single-Pixel Imaging. *Chinese Journal of Electronics*, 32(5):1151–1159, 2023. doi: 10.23919/cje.2022.00.284.

[34] Tao Zhang, Ying Fu, and Jun Zhang. Deep Guided Attention Network for Joint Denoising and Demosaicing in Real Image. *Chinese Journal of Electronics*, 33(1):303–312, 2024. doi: 10.23919/cje.2022.00.414.

[35] Peilin Zhao, Steven C. H. Hoi, Rong Jin, and Tianbao Yang. Online AUC maximization. In *Proceedings of the 28th International Conference on Machine Learning ICML 2011: Bellevue, WA, June 28 - July 2*, pages 233–240. International Machine Learning Society, 2011. URL https://ink.library.smu.edu.sg/sis_research/2351.

[36] Purushottam Kar, Bharath Sriperumbudur, Prateek Jain, and Harish Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In *International Conference on Machine Learning*, pages 441–449. PMLR, 2013.

[37] Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass AUC optimization. In *International conference on machine learning*, pages 906–914. PMLR, 2013.

[38] Stephan Clémençon, Igor Colin, and Aurélien Bellet. Scaling-up empirical risk minimization: optimization of incomplete U-statistics. *Journal of Machine Learning Research*, 17(76):1–36, 2016.

[39] Michael Natole, Yiming Ying, and Siwei Lyu. Stochastic proximal algorithms for AUC maximization. In *International Conference on Machine Learning*, pages 3710–3719. PMLR, 2018.

[40] Yunwen Lei and Yiming Ying. Stochastic proximal AUC maximization. *Journal of Machine Learning Research*, 22(61):1–45, 2021.

[41] Zhishuai Guo, Mingrui Liu, Zhuoning Yuan, Li Shen, Wei Liu, and Tianbao Yang. Communication-efficient distributed stochastic AUC maximization with deep neural networks. In *International conference on machine learning*, pages 3864–3874. PMLR, 2020.

[42] Zhiyuan Dang, Xiang Li, Bin Gu, Cheng Deng, and Heng Huang. Large-Scale Nonlinear AUC Maximization via Triply Stochastic Gradients. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1385–1398, 2022. doi: 10.1109/TPAMI.2020.3024987.

[43] Wenzheng Hou, Qianqian Xu, Zhiyong Yang, Shilong Bao, Yuan He, and Qingming Huang. AdAUC: End-to-end adversarial AUC optimization against long-tail problems. In *International Conference on Machine Learning*, pages 8903–8925. PMLR, 2022.

[44] Siran Dai, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. DRAUC: an instance-wise distributionally robust AUC optimization framework. *Advances in Neural Information Processing Systems*, 36:44658–44670, 2023.

[45] Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed Diffusion Models with Oriented Calibration. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NW2s5XXwXU.

[46] JIAAN LUO, Feng Hong, Qiang Hu, Xiaofeng Cao, Feng Liu, and Jiangchao Yao. Long-tailed Recognition with Model Rebalancing. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 135103–135130. Curran Associates, Inc., 2025. doi: 10.52202/085713-4508. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c4889bd7f7ce643003746526da2c2fc4-Paper-Conference.pdf.

[47] Feng Hong, Jiangchao Yao, Yifei Shen, Dongsheng Li, Ya Zhang, and Yanfeng Wang. Improving Diffusion Models for Class-imbalanced Training Data via Capacity Manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wSGle6ag5I.

[48] Yu Huang, Zihua Zhao, Zhaoxin Huan, Wanli Gu, Feng Hong, Xinmu Ge, Lin Yuan, Weichang Wu, Qiang Hu, Xiaolu Zhang, Jun Zhou, and Jiangchao Yao. Focal Reward: Balanced Reinforcement Learning under Rubric-Based Rewards, 2026. URL https://arxiv.org/abs/2605.26579.

[49] Aditya Krishna Menon and Robert C Williamson. Bayes-optimal scorers for bipartite ranking. In *Conference on Learning Theory*, pages 68–106. PMLR, 2014.

[50] Wojciech Kotlowski, Krzysztof J Dembczynski, and Eyke Huellermeier. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1113–1120, 2011.

[51] Shivani Agarwal. Surrogate regret bounds for the area under the ROC curve via strongly proper losses. In *Conference on Learning Theory*, pages 338–353. PMLR, 2013.

[52] Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *The Journal of Machine Learning Research*, 15(1):1653–1674, 2014.

[53] Wei Gao and Zhi-Hua Zhou. On the Consistency of AUC Pairwise Optimization. In Qiang Yang and Michael Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-15*, pages 939–944. International Joint Conferences on Artificial Intelligence, 7 2015. URL https://ijcai.org/proceedings/2015/137. Main Track.

[54] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online AUC maximization. *Advances in neural information processing systems*, 29, 2016.

[55] Mengdi Wang, Ethan X. Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161:419–449, 2017. doi: 10.1007/s10107-016-1017-3.

[56] Bokun Wang and Tianbao Yang. Finite-Sum Coupled Compositional Stochastic Optimization: Theory and Applications. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23292–23317. PMLR, 2022. URL https://proceedings.mlr.press/v162/wang22ak.html.

[57] Harikrishna Narasimhan and Shivani Agarwal. A Structural SVM Based Approach for Optimizing Partial AUC. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 516–524. PMLR, 2013. URL https://proceedings.mlr.press/v28/narasimhan13.html.

[58] Huiyang Shao, Qianqian Xu, Zhiyong Yang, Shilong Bao, and Qingming Huang. Asymptotically unbiased instance-wise regularized partial AUC optimization: Theory and algorithm. *Advances in Neural Information Processing Systems*, 35:38667–38679, 2022.

[59] Dixian Zhu, Gang Li, Bokun Wang, Xiaodong Wu, and Tianbao Yang. When AUC meets DRO: Optimizing partial AUC for deep learning with non-convex convergence guarantee. In *International Conference on Machine Learning*, pages 27548–27573. PMLR, 2022.

[60] Linli Zhou, Bokun Wang, My T Thai, and Tianbao Yang. Stochastic Primal-Dual Double Block-Coordinate for Two-way Partial AUC Maximization. *arXiv preprint arXiv:2505.21944*, 2025.

[61] Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

[62] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3045–3059, 2021.

[63] Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo Yin, and Hulingxiao He. A Survey on Fine-Grained Multimodal Large Language Models. *Chinese Journal of Electronics*, 35(2): 771–803, 2026. doi: 10.23919/cje.2025.00.336.

[64] Linhui Xiao, Guiping Cao, Mingyue Guo, Xianchao Guan, Fan Yang, Ming Tao, Xin Li, Yuxin Peng, and Yaowei Wang. A Survey on the Green Development of Large Models: From Resource-Efficient Architectures to Hardware-Software Co-Design. *Chinese Journal of Electronics*, 35 (5):1–24, 2026. doi: 10.23919/cje.2025.00.438.

[65] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

[66] Emmanuel Abbe, Elisabetta Cornacchia, Jan Hazla, and Christopher Marquis. An initial alignment between neural network and target is needed for gradient descent to learn. In *International Conference on Machine Learning*, pages 33–52. PMLR, 2022.

[67] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. doi: 10.1214/aoms/1177729586.

[68] Harold J Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

[69] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

[70] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019. URL http://jmlr.org/papers/v20/17-526.html.

[71] Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for SGD and its continuous-time counterpart. In *Conference on Learning Theory*, pages 1965–2058. PMLR, 2021.

[72] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2532–2541. IEEE, 2019.

[73] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[74] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel Dataset for Fine-Grained Image Categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[75] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[76] Asia Pacific Tele-Ophthalmology Society. APTOS 2019 Blindness Detection. Kaggle Competition Dataset, 2019. URL https://www.kaggle.com/c/aptos2019-blindness-detection.

[77] Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. DeepWeeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1):2058, 2019.

[78] Himanshu Sharma, Devanand Padha, and Nahida Bashir. D-KAP: A deep learning-based Kashmiri apple plant disease prediction framework. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 576–581. IEEE, 2022.

[79] Harshadkumar B Prajapati, Jitesh P Shah, and Vipul K Dabhi. Detection and classification of rice plant diseases. *Intelligent Decision Technologies*, 11(3):357–373, 2017.

[80] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 180161, 2018.

[81] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. RandAugment: Practical automated data augmentation with a reduced search space. *Advances in neural information processing systems*, 33:18613–18624, 2020.

[82] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images, 2009. URL https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

[83] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pages 8769–8778. IEEE, 2018.

[84] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[85] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4040–4054, 2020.

[86] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4885–4901, 2020.

[87] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642, 2015.

[88] Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, 2022.

[89] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[90] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

[91] Zhuoning Yuan, Dixian Zhu, Zi-Hao Qiu, Gang Li, Xuanhui Wang, and Tianbao Yang. LibAUC: A deep learning library for X-Risk optimization. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5487–5499, 2023.

Appendix

Table of Contents

A Related Work . . . 18
B Experimental Details . . . 19
B.1 Details of the Controlled Two-Dimensional Simulation . . . 19
B.2 Details of the Noise-Scale Escape Simulation . . . 21
B.3 Main Experimental Protocol . . . 21
B.4 Data Sources and Task Construction . . . 22
B.5 Models, Optimization, and Reporting Protocol . . . 24
B.6 Multiclass, Multilabel, and Deep-AUC Extensions . . . 30
B.7 Assets, Licenses, and Terms . . . 32
B.8 Ethics, Broader Impacts, and Safeguards . . . 33
C Direct Pairwise SGD as a Stochastic Dynamics . . . 33
C.1 Extensions Beyond Binary AUC . . . 37
D Bad Regions and Hoeffding Projection Collapse . . . 37
E Surrogate Weight Calculations . . . 50
E.1 Canonical Weights . . . 50
E.2 Active-Side Correction Versus Class-Mean Collapse . . . 50
E.3 Easy-Side Silence Versus Overshoot . . . 52
E.4 Hard-Tail Alignment . . . 53
E.5 Projected Stochastic Separation . . . 54
F Changing Where Pairwise AUC Optimization Begins . . . 55
F.1 CE Warm-Up and Severe-Misranking Mass . . . 55
F.2 Parameter-Efficient Fine-Tuning in a Stable Trainable Subspace . . . 57
F.3 Local Pairwise AUC-Surrogate Descent in a Favorable Subspace . . . 60
G Model-Class Realizations of Projected Bad Components . . . 61
G.1 Data-Dependent One-Layer Construction . . . 61
G.2 Projected Bad Components in Trainable Transformer Neighborhoods . . . 64
H Localized Escape-Time Bounds for Stopped SGD . . . 69
I Limitations . . . 73

How to read the appendix. Appendix A gives the extended related work. Appendix B records the controlled simulations, the full protocol behind the reported CV and NLP experiments, and the asset, license, and ethics information relevant to the experiments. The remaining appendices give the proofs: Appendix C fixes the direct pairwise SGD process and its local Langevin interface; Appendix D proves the bad-region and Hoeffding-projection statements used in Section 2; Appendix E proves the surrogate-weight claims used in Section 3; Appendix F proves the warm-up and parameter-efficient fine-tuning statements behind Section 4; Appendix G shows that the projected bad-component conditions can be realized inside Transformer-style model classes; Appendix H gives the stopped-SGD escape-time tool that converts directional drift and second-moment bounds into the projected B versus B^2 scale contrast; and Appendix I states the scope and limitations of the theory.

A Related Work

This appendix organizes the literature behind the paper's main question: when direct pairwise AUC-surrogate training uses finite mini-batch SGD, do severe ranking mistakes still generate useful parameter motion? We group the discussion by statistical target, scalable AUC algorithms, surrogate validity, U-statistic structure, related auxiliary-state methods, partial-AUC variants, warm-up and pretraining practice, and local stochastic dynamics.

AUC, ranking, and non-decomposable learning. AUC is a bipartite-ranking objective rather than a decomposable classification loss: it evaluates whether positives receive larger scores than negatives across opposite-class pairs. Classical and structural views of ROC area and multivariate performance measures formalize this non-decomposable nature [25, 26]. Broader surveys place AUC beside bipartite ranking, partial AUC, top-focused ranking, and precision-style criteria [27]. Related work also adapts the AUC target itself to multiclass learning, cost-sensitive decision making, open-set recognition, and pixel-level long-tail segmentation [28–31]. Cross-media retrieval provides another ranking-oriented application context, including unsupervised cross-media hashing based on knowledge graphs [32]. Adjacent learning-based vision problems include under-sampled single-pixel reconstruction and joint denoising and demosaicing [33, 34], but these are reconstruction and restoration tasks rather than the binary-ranking problem studied here. This literature fixes the target of interest. Our analysis focuses on the training dynamics induced by that target: a pairwise SGD update must convert opposite-class ranking errors into corrective motion of the scorer.

AUC optimization algorithms. The quadratic coupling between positives and negatives makes AUC maximization structurally different from pointwise ERM. This has motivated online AUC optimization, pairwise stochastic methods, one-pass and incomplete-pair methods, proximal methods, triply stochastic methods, distributed methods, and nonlinear extensions [35–42]. Deep AUC training adds representation learning, robust surrogates, warm-up, and task-dependent optimization design [10, 12]. AUC-oriented algorithms have also been developed for long-tail adversarial training and domain-robust or distributionally robust settings [43, 44]. These works make pairwise AUC computable at scale. Adjacent long-tailed learning research addresses class imbalance beyond AUC through oriented calibration for diffusion models, parameter-space rebalancing for recognition, and explicit capacity manipulation for class-imbalanced diffusion training [45–47]. Related imbalance-aware objective design also appears in rubric-based LLM reinforcement learning, where criterion-level rewards are dynamically rebalanced according to their saturation [48]. Our contribution is complementary: even when the pairwise update is computable, its local drift and covariance can stop turning severe misrankings into effective motion.

Surrogate validity versus surrogate dynamics. Population-level theory explains why several losses are meaningful for ranking. Bayes-optimal bipartite ranking, univariate surrogate control, proper-loss regret transfers, and pairwise consistency results address calibration, regret, and statistical validity [5, 49–53]. The present paper studies a different property of the same losses. Logistic, exponential, square, hinge, and squared-hinge losses assign different weights to pairwise margins; those weights determine whether severe mistakes receive stronger correction, whether easy pairs continue to consume update mass, and whether useful stochastic variation remains available along difficult trajectories.

U-statistics and finite-batch stochastic structure. Empirical AUC risks are two-sample U-statistics. Classical U-statistic theory gives the Hoeffding projection decomposition and the corresponding first-order and canonical components [19, 20]. We use this decomposition as a stochastic-dynamics tool. For pairwise mini-batches with B_+ positives and B_- negatives, the first-order projection scales are B_+^{-1} and B_-^{-1}, while the residual pair-interaction scale is $(B_+B_-)^{-1}$. When the first-order projections vanish in a bad region, the remaining canonical term has scale $(B_+B_-)^{-1}$, which becomes B^{-2} in the equal-size case. This variance split is the technical source of the escape-time contrast used in the main text.

Auxiliary-state AUC methods. Several scalable AUC methods reformulate the update process rather than optimize a raw pairwise objective in the most direct mini-batch form. Stochastic saddle-point AUC maximization [54], stochastic proximal AUC methods [39], deep AUC [10], robust deep-AUC surrogates [11], compositional AUC training [12, 55, 56], and adversarial or distributionally

robust AUC formulations [43, 44] introduce dual variables, proximal structure, auxiliary trackers, nested stochastic updates, or robustness-oriented inner problems. These methods are related because they redesign the stochastic process used for AUC training. The main text studies direct pairwise mini-batch SGD; auxiliary-state methods provide broader algorithmic context.

Partial AUC and top-focused variants. Partial AUC and top-focused ranking sharpen the same non-decomposable structure by emphasizing selected portions of the ROC curve or the highest-ranked examples. Structural partial-AUC optimization [57], generic and end-to-end two-way partial-AUC frameworks [14, 15], asymptotically unbiased instance-wise partial-AUC optimization [58], distributionally robust deep partial-AUC methods [59], and stochastic primal-dual algorithms for two-way partial AUC [60] introduce selection, sorting, distributional robustness, min-max reformulations, or primal-dual state. We use full pairwise AUC as the clean setting for the proofs, while these variants reinforce the need for optimization analyses that go beyond pointwise ERM.

Warm-up, pretraining, and parameter-efficient fine-tuning. Practical deep-AUC pipelines often use pointwise warm-up, pretrained backbones, or low-dimensional trainable modules before or during AUC optimization [10–12]. In large-model settings, adapters, prefix tuning, prompt tuning, LoRA, and QLoRA are standard forms of parameter-efficient fine-tuning from pretrained models [16–18, 61, 62]. These questions become increasingly relevant as pretrained systems extend to fine-grained multimodal large models and as resource-efficient architectures and hardware–software co-design receive growing attention [63, 64]. Initialization and alignment results further show that the starting point can change learnability and optimization behavior [65, 66]. Our theory separates the roles of these practices: CE warm-up changes the starting region of the AUC phase, while the pretrained base and trainable subspace can place the pairwise phase closer to a favorable starting region.

Local stochastic dynamics and escape. The bridge from discrete SGD updates to local stochastic dynamics uses standard stochastic-approximation and Markov-process tools [67–71]. In our setting, the drift is governed by severity-weighted pairwise correction, while the covariance is governed by the Hoeffding projection structure of the mini-batch U-statistic. Appendix H turns the B^{-1} versus B^{-2} directional second-moment regimes into the projected escape-time statement used in the main text.

B Experimental Details

B.1 Details of the Controlled Two-Dimensional Simulation

This subsection gives the full configuration for Figure 1. The simulation makes the local dynamical mechanism in the introduction reproducible on a fixed two-dimensional parameter plane. All curves in the figure are generated by running the stated update rules; no trajectory is hand drawn or smoothed.

Data generation. The controlled simulation uses a fixed synthetic binary ranking dataset generated from the following configuration. The random seed is 1136848004. Each group parameter is $n_{\text{per group}} = 12$, giving four groups with counts

$$\max(2, \lfloor n_{\text{per group}}/3 \rfloor), \quad 2 \cdot n_{\text{per group}}, \quad n_{\text{per group}}, \quad 4 \cdot n_{\text{per group}},$$

where groups 0 and 2 are positive and groups 1 and 3 are negative. The data-generation parameters are

$$\alpha = 0.012, \qquad \beta = 0.64, \qquad \texttt{hard_offset} = 0.16, \qquad \texttt{nonlinear} = 0.16,$$

$$\texttt{bump} = 0.46, \qquad \texttt{rbf_scale} = 0.08, \qquad \texttt{fourier_scale} = 0.15.$$

For a point $\theta = (\theta_1, \theta_2)$, the sample score has the form

$$f_\theta(x_i) = a_i \cdot \theta_1 + b_i \cdot \theta_2 + c_{1i} \cdot h_{1i}(\theta_1, \theta_2) + c_{2i} \cdot h_{2i}(\theta_1, \theta_2) + c_{3i} \cdot q_i(\theta_1, \theta_2) + r_i(\theta_1, \theta_2) + s_i(\theta_1, \theta_2) + o_i.$$

Here h_{1i} is a tanh feature, h_{2i} is a sinusoidal feature, q_i is a Gaussian bump, r_i is a sum of 14 radial-basis features, s_i is a sum of 8 Fourier features, and o_i is a sample offset. Their coefficients are generated once from the seed above and then held fixed for all panels and all update rules.

Landscape definition. The displayed landscape is the empirical AUC of the same fixed sample set on a two-dimensional parameter slice. For each grid point (θ_1, θ_2), the positive-margin convention is

$$m_\theta(x_i^+, x_j^-) = f_\theta(x_i^+) - f_\theta(x_j^-),$$

and the plotted value is

$$\widehat{\mathrm{AUC}}(\theta) = \frac{1}{N_+ \cdot N_-} \sum_{i:y_i=+1} \sum_{j:y_j=-1} \left\{ \mathbf{1}_{\{m_\theta(x_i,x_j)>0\}} + \frac{1}{2} \cdot \mathbf{1}_{\{m_\theta(x_i,x_j)=0\}} \right\}.$$

The contour is evaluated on the fixed window

$$\theta_1 \in [-1.08, 2.18], \qquad \theta_2 \in [-1.60, 0.46],$$

using a 300×250 grid. The filled contour levels are evenly spaced between 0.50 and 1.00.

Surrogate losses and update rules. Figure 1 displays direct hinge and direct squared-hinge AUC trajectories, together with CE warm-up trajectories that switch to hinge or squared-hinge after the warm-up phase. The pairwise losses used in the displayed AUC phases are

$$\ell_{\text{hinge}}(x) = (m - x)_+, \qquad \ell_{\text{squared-hinge}}(x) = (m - x)_+^2,$$

with margin $m = 0.8$. The AUC vector field uses the full finite-sample pair average

$$\frac{1}{N_+ \cdot N_-} \sum_{i:y_i=+1} \sum_{j:y_j=-1} w_\ell(m_\theta(x_i, x_j)) \cdot \nabla_\theta m_\theta(x_i, x_j), \qquad w_\ell(x) = -\ell'(x).$$

The CE vector field is the full finite-sample gradient direction of the binary cross-entropy objective on the same samples. Direct pairwise AUC-surrogate curves repeatedly apply the corresponding pairwise surrogate vector field. Warm-up curves apply the CE field first and then switch to the chosen pairwise surrogate field.

Optimization settings. The AUC step size is

$$\eta_{\text{auc}} = 0.15,$$

the CE step size is

$$\eta_{\text{ce}} = 0.105,$$

the warm-up length is 26 steps, and the maximum trajectory length is 360 steps. The severe-misranking threshold used by the scan is $M = 0.2$, and the severe-badness exit threshold is 0.0015, used only to detect when a trajectory leaves the severe-start region. The trajectories are clamped to the numerical box $[-8, 8]^2$ only as a stability guard; the displayed trajectories stay in the plotted window.

Panel initializations. All three panels use the same generated sample set, the same landscape, and the same optimization hyperparameters. They differ only in the initialization:

$$\begin{aligned} \text{panel (a)} &: (\theta_{1,0}, \theta_{2,0}) = (-0.450, -1.3775), \\ \text{panel (b)} &: (\theta_{1,0}, \theta_{2,0}) = (-0.075, -1.0450), \\ \text{panel (c)} &: (\theta_{1,0}, \theta_{2,0}) = (0.750, -0.5375). \end{aligned}$$

The corresponding initial severe-misranking fractions are 0.3993, 0.2752, and 0.1372, respectively. Final empirical AUC values for the trajectories emphasized in the main figure are:

panel	CE	hinge	squared-hinge	warm → hinge	warm → squared-hinge
(a)	0.9106	0.6016	0.6207	0.9028	0.9766
(b)	0.9106	0.9028	0.9766	0.9028	0.9766
(c)	0.9106	0.9757	0.9766	0.9757	0.9766

Plotting choices. The plotted paths are raw simulated iterates. Stars mark initializations, large circles mark final iterates, and small dots mark every 20 update steps. The contour is evaluated once on the fixed grid above and reused across all panels. The displayed runs are fixed reproducible simulations corresponding to the three regimes discussed in the introduction: weak correction from a severe start, recovery from a milder start, and favorable-start behavior. All reported trajectories are raw iterates from the corresponding update rules.

B.2 Details of the Noise-Scale Escape Simulation

This subsection gives the configuration for panel (a) of Figure 2. The simulation serves as a controlled visualization of the escape-time mechanism in Corollary 1. It keeps the potential, initialization, drift, and time horizon fixed, and changes only the variance scale from the first-order B^{-1} regime to the collapsed B^{-2} regime.

Potential and landscape. The one-dimensional state is denoted by z. The unnormalized potential is

$$\begin{aligned} U_{\mathrm{raw}}(z) = {} & 0.105 \cdot (z + 1.45)^2 \cdot (z - 1.34)^2 \\ & + 0.18 \cdot \exp\left(- \left(\frac{z - 0.05}{0.37} \right)^2 \right) - 0.075 \cdot z \\ & + 0.018 \cdot \sin(2.6 \cdot z + 0.2). \end{aligned}$$

For plotting and simulation, this potential is rescaled on the grid $z \in [-2.10, 2.02]$ by

$$U(z) = 0.04 + 0.90 \cdot \frac{U_{\mathrm{raw}}(z) - \min_{\tilde{z}} U_{\mathrm{raw}}(\tilde{z})}{\max_{\tilde{z}} U_{\mathrm{raw}}(\tilde{z}) - \min_{\tilde{z}} U_{\mathrm{raw}}(\tilde{z})}.$$

The grid has 3000 points. The displayed state window is $z \in [-1.92, 1.82]$. The left well minimum, barrier, right well minimum, and escape threshold are

$$z_{\mathrm{left}} = -1.3691, \qquad z_{\mathrm{barrier}} = 0.0170, \qquad z_{\mathrm{right}} = 1.4032, \qquad z_{\mathrm{esc}} = 0.5370.$$

The initialization is $z_0 = -1.42$.

Stochastic update. The simulation uses an Euler discretization of overdamped motion in the potential:

$$z_{t+1} = z_t - \Delta t \cdot U'(z_t) + \sigma_B \cdot \xi_t, \qquad \xi_t \sim \mathcal{N}(0, 1),$$

with independent ξ_t's across time and replicates. The derivative U' is computed by finite differences on the same 3000-point grid and linearly interpolated at the current state. States are clipped to $[-2.05, 1.95]$ as a numerical guard.

Noise regimes and hyperparameters. Both regimes use

$$B = 8, \qquad \Delta t = 0.025, \qquad \texttt{noise_base} = 4.5, \qquad T = 180$$

iterations. The first-order regime uses

$$\sigma_B = \texttt{noise_base} \cdot \sqrt{\Delta t / B},$$

while the collapsed regime uses

$$\sigma_B = \texttt{noise_base} \cdot \sqrt{\Delta t} / B.$$

The empirical density in Figure 2(a) is computed from 12000 independent replicates per regime and sampled every 2 iterations. The random seeds are 20260524 for the B^{-1} regime and 20260512 for the B^{-2} regime. A replicate is counted as escaped once $z_t \geq z_{\mathrm{esc}}$ for some $t \leq T$. With the above configuration, the final escape fractions are 0.76275 for the first-order regime and 0.0265 for the collapsed regime, shown in the figure as 76% and 3%.

Rendering choices. The left curve in panel (a) is the same potential U, rotated so that the vertical axis is the state variable z. The two right subpanels show smoothed empirical state-density heatmaps over iteration and state. Horizontal guide lines mark the initialization, barrier, and target well. The density smoothing is used only for rendering; the escape fractions are computed from the raw simulated paths.

B.3 Main Experimental Protocol

This subsection fixes the reporting protocol behind Table 1 and the expanded tables below. Each experiment is a binary ranking task evaluated by held-out AUC. For every model–task pair, the protocol compares CE with five pairwise surrogate families: square, hinge, exponential, logistic, and squared-hinge. Pairwise AUC training is run either directly from the beginning or after a CE warm-up phase.

Reported quantity. Every number in the result tables is the held-out test AUC of the validation-best checkpoint, averaged over matching runs. The main table and the expanded appendix tables report mean held-out AUC values with sample standard deviations. When a dataset does not provide fixed train, validation, and test splits, the same seed convention is used to create stratified splits before training.

Direct and warm-started rows. For CE, the main table reports the CE run. For each pairwise surrogate family, we run both direct pairwise AUC-surrogate training and CE-warm-started pairwise AUC-surrogate training when both are part of the grid. Table 1 reports these two modes as separate direct and warm-up rows for each task, with bold and underline marking the best and second-best results within each task. The expanded appendix tables keep the same direct/warm-started separation for the remaining model-family blocks.

Expanded results corresponding to Table 1. Table 1 displays DenseNet121 and Qwen2.5-1.5B-Instruct with direct and CE-warm-started rows for all five pairwise surrogate families. Table 2 gives the corresponding ResNet50 and DistilBERT-base results: each task is split into a direct row and a CE-warm-started row, and each cell reports the mean held-out AUC together with the corresponding sample standard deviation.

Table 2: **Expanded direct and CE-warm-started results (AUC %) for ResNet50 and DistilBERT-base.** This table reports the model-family blocks not shown in Table 1. Each task has a direct row and a warm-up row. Values are mean held-out AUC (%), with the $\pm$ sample standard deviation shown as a subscript. Within each task, the best and second-best configurations are bolded with darker shading and underlined with lighter shading, respectively.

Task	Mode	CE	Square	Hinge	Exp.	Logistic	Sq.-hinge
ResNet50							
ImageNet-LT Dogs	Direct	$91.34_{\pm 1.00}$	$93.56_{\pm 0.38}$	$79.50_{\pm 3.75}$	$91.35_{\pm 0.86}$	$93.67_{\pm 0.47}$	**95.83**$_{\pm 0.14}$
	Warm-up	–	$93.84_{\pm 0.34}$	$82.33_{\pm 2.75}$	$91.00_{\pm 1.48}$	$93.51_{\pm 0.54}$	$\underline{94.42}_{\pm 0.41}$
Dog Breeds	Direct	$86.06_{\pm 1.09}$	$85.09_{\pm 1.08}$	$84.63_{\pm 1.28}$	$84.68_{\pm 0.84}$	$84.82_{\pm 1.28}$	$85.65_{\pm 1.04}$
	Warm-up	–	$\underline{86.52}_{\pm 1.03}$	$83.72_{\pm 1.74}$	$85.87_{\pm 0.91}$	$85.30_{\pm 1.79}$	**87.00**$_{\pm 1.44}$
CRJ700/CRJ900	Direct	$53.57_{\pm 7.75}$	$54.99_{\pm 8.32}$	$54.77_{\pm 8.25}$	$54.93_{\pm 8.27}$	$55.54_{\pm 7.40}$	$\underline{55.57}_{\pm 8.12}$
	Warm-up	–	$54.66_{\pm 7.30}$	$54.77_{\pm 8.25}$	$54.73_{\pm 7.27}$	$54.51_{\pm 7.30}$	**56.62**$_{\pm 6.45}$
APTOS Sev./Mod.	Direct	$81.02_{\pm 1.47}$	$\underline{82.70}_{\pm 0.63}$	$81.54_{\pm 0.84}$	$82.01_{\pm 0.34}$	$81.79_{\pm 1.04}$	**83.92**$_{\pm 1.09}$
	Warm-up	–	$80.49_{\pm 0.95}$	$80.62_{\pm 1.61}$	$80.70_{\pm 1.22}$	$80.38_{\pm 0.78}$	$81.26_{\pm 1.34}$
DeepWeeds P/S	Direct	$96.39_{\pm 1.37}$	$96.98_{\pm 2.35}$	$96.99_{\pm 2.36}$	$96.98_{\pm 2.37}$	$96.90_{\pm 2.41}$	$\underline{97.04}_{\pm 2.38}$
	Warm-up	–	$96.98_{\pm 2.35}$	$96.15_{\pm 2.76}$	$96.98_{\pm 2.37}$	$96.90_{\pm 2.41}$	**97.12**$_{\pm 2.61}$
DistilBERT-base							
GoEmotions Annoy.	Direct	$83.97_{\pm 0.35}$	$82.73_{\pm 1.61}$	$80.39_{\pm 3.96}$	$79.91_{\pm 4.16}$	$79.20_{\pm 4.16}$	$83.87_{\pm 0.50}$
	Warm-up	–	$83.97_{\pm 0.35}$	$82.75_{\pm 1.16}$	$\underline{83.98}_{\pm 0.35}$	$83.97_{\pm 0.35}$	**84.31**$_{\pm 0.38}$
GoEmotions Fear	Direct	$94.74_{\pm 0.58}$	$94.05_{\pm 1.65}$	$93.50_{\pm 0.73}$	$93.07_{\pm 0.80}$	$94.67_{\pm 1.48}$	$\underline{95.62}_{\pm 0.82}$
	Warm-up	–	$93.89_{\pm 2.03}$	$93.84_{\pm 1.74}$	$94.25_{\pm 1.22}$	$93.26_{\pm 0.60}$	**96.48**$_{\pm 0.31}$
ANLI R1 E/N	Direct	**70.21**$_{\pm 1.21}$	$66.37_{\pm 2.07}$	$61.51_{\pm 5.60}$	$64.44_{\pm 1.84}$	$63.32_{\pm 2.28}$	$67.84_{\pm 1.63}$
	Warm-up	–	$69.07_{\pm 1.52}$	$\underline{69.86}_{\pm 1.40}$	$69.09_{\pm 1.10}$	$68.48_{\pm 2.16}$	$69.77_{\pm 1.35}$
ANLI R3 C/N	Direct	$\underline{62.09}_{\pm 0.84}$	$61.48_{\pm 0.63}$	$57.40_{\pm 2.26}$	**62.47**$_{\pm 2.17}$	$61.76_{\pm 1.62}$	$61.22_{\pm 1.30}$
	Warm-up	–	$61.88_{\pm 0.64}$	$61.09_{\pm 0.77}$	$59.64_{\pm 1.35}$	$61.31_{\pm 0.98}$	$61.81_{\pm 0.69}$
ANLI R3 E/N	Direct	$\underline{63.49}_{\pm 0.62}$	$62.40_{\pm 1.23}$	$58.85_{\pm 0.79}$	$58.49_{\pm 2.04}$	$60.11_{\pm 0.85}$	$62.24_{\pm 0.86}$
	Warm-up	–	$62.59_{\pm 1.60}$	$62.86_{\pm 0.59}$	$63.00_{\pm 0.67}$	$63.14_{\pm 0.57}$	**63.87**$_{\pm 0.51}$
SNLI E/N	Direct	$\underline{92.56}_{\pm 0.34}$	$89.07_{\pm 2.98}$	$86.21_{\pm 1.25}$	$86.69_{\pm 1.06}$	$84.42_{\pm 3.97}$	$92.52_{\pm 0.43}$
	Warm-up	–	$92.21_{\pm 0.44}$	$92.21_{\pm 0.49}$	$92.37_{\pm 0.20}$	$92.21_{\pm 0.44}$	**92.98**$_{\pm 0.22}$
WANLI C/N	Direct	$\underline{82.39}_{\pm 0.49}$	$80.58_{\pm 0.35}$	$75.51_{\pm 3.21}$	$79.54_{\pm 0.61}$	$77.89_{\pm 0.61}$	$82.08_{\pm 0.36}$
	Warm-up	–	$81.74_{\pm 0.43}$	$81.79_{\pm 0.48}$	$81.90_{\pm 0.34}$	$81.84_{\pm 0.40}$	**82.41**$_{\pm 0.32}$
WANLI E/N	Direct	**75.44**$_{\pm 0.50}$	$74.43_{\pm 0.27}$	$65.37_{\pm 4.09}$	$73.42_{\pm 0.35}$	$71.60_{\pm 1.88}$	$74.76_{\pm 0.67}$
	Warm-up	–	$74.72_{\pm 0.27}$	$73.99_{\pm 0.68}$	$74.63_{\pm 0.36}$	$74.24_{\pm 0.56}$	$\underline{75.41}_{\pm 0.29}$

B.4 Data Sources and Task Construction

The main experiments use binary ranking tasks. Appendix B.6 further reports long-tailed multiclass, multilabel, and large-scale multiclass settings. For the binary tasks, each row defines a positive class and a negative class or negative class set, and all models are evaluated by held-out AUC on the resulting binary task. When a source dataset provides official train, validation, and test splits, those

Table 3: CV task construction. Each row gives the source dataset and the binary ranking task used in the experiments.

Task	Source	Binary construction
ImageNet-LT Dogs	ImageNet-LT split files with ImageNet images	Tail dog classes versus sibling dog classes.
Dog Breeds	Stanford Dogs	Spitz, husky, and malamute versus Eskimo dog and Samoyed.
CRJ700/CRJ900	FGVC-Aircraft	Bombardier CRJ700 versus Bombardier CRJ900.
APTOS Sev./Mod.	APTOS diabetic-retinopathy fundus images	Severe-or-worse retinopathy versus moderate retinopathy.
DeepWeeds P/S	DeepWeeds	Parthenium weed versus Siam weed, with low-shot training construction.
Apple Disease	D-KAP apple leaf disease archive	Apple scab versus apple rot, with low-shot training construction.
Rice Disease	Rice Leaf Diseases	Brown spot versus leaf smut, with low-shot training construction.
HAM10000 M/N	HAM10000	Melanoma versus melanocytic nevus.

Table 4: NLP task construction. DistilBERT and Qwen use the same binary task definitions.

Task	Source	Binary construction
GoEmotions Annoyance	GoEmotions, simplified	Annoyance label versus all other examples.
GoEmotions Fear	GoEmotions, simplified	Fear label versus all other examples.
ANLI R1 Ent./Neu.	ANLI Round 1 prepared split	Entailment versus neutral.
ANLI R3 Con./Neu.	ANLI Round 3 prepared split	Contradiction versus neutral.
ANLI R3 Ent./Neu.	ANLI Round 3 prepared split	Entailment versus neutral.
SNLI High-Overlap	SNLI prepared high-overlap split	Entailment versus neutral on a high-overlap slice.
WANLI Con./Neu.	WANLI prepared split	Contradiction versus neutral.
WANLI Ent./Neu.	WANLI prepared split	Entailment versus neutral.

splits are used. Otherwise, stratified splits follow the reporting protocol. Text fields are converted into a single model input for NLP tasks, and labels are binarized before training.

CV tasks. The CV tasks are constructed from public image datasets and trained with ImageNet-normalized image pipelines. The CV rows use the following sources and binary definitions. The public CV sources are ImageNet-LT split files derived from ImageNet [72, 73], Stanford Dogs [74], FGVC-Aircraft [75], APTOS 2019 [76], DeepWeeds [77], D-KAP [78], Rice Leaf Diseases [79], and HAM10000 [80]. Public releases are available for Stanford Dogs, FGVC-Aircraft, APTOS 2019, DeepWeeds, D-KAP, Rice Leaf Diseases, and HAM10000. The image preprocessing uses random resized crop to 224, random horizontal flip, RandAugment [81] with 2 operations and magnitude 9, conversion to tensor, and ImageNet normalization for training; validation and test use resize, center crop, tensor conversion, and ImageNet normalization.

The extension experiments additionally use CIFAR-100 [82], iNaturalist 2018 [83], and CheXpert [84]. The CIFAR-100 experiments construct long-tailed medium-mammals and vehicles tasks from the official classes; iNaturalist 2018 uses the official training and validation sets for an 8,142-class long-tailed task; and CheXpert uses patient-disjoint training and internal-validation data, with the official validation set reserved for final testing. The task compositions, split sizes, and evaluation metrics are specified in Appendix B.6.

NLP tasks. The DistilBERT and Qwen experiments use the same NLP task definitions. GoEmotions tasks are one-versus-rest emotion detection tasks; the official test split is used and a stratified validation split is carved from the official training split. ANLI, SNLI, and WANLI rows are prepared binary natural-language-inference contrasts with train, validation, and test splits fixed before model training. The NLP sources are Hugging Face `go_emotions` with the simplified configuration [85], `anli` [86],

Table 5: CV split sizes. In the split columns, N is the total number of examples and pos./neg. gives the positive/negative counts.

CV task	Train (N; pos./neg.)	Val. (N; pos./neg.)	Test (N; pos./neg.)
ImageNet-LT Dogs	6788 (202/6586)	1080 (360/720)	2700 (900/1800)
Dog Breeds	517 (259/258)	111 (56/55)	110 (55/55)
CRJ700/CRJ900	50 (25/25)	30 (15/15)	66 (33/33)
APTOS Sev./Mod.	1041 (342/699)	223 (73/150)	223 (73/150)
DeepWeeds P/S	50 (25/25)	30 (15/15)	80 (40/40)
Apple Disease	50 (25/25)	30 (15/15)	80 (40/40)
Rice Disease	36 (18/18)	16 (8/8)	28 (14/14)
HAM10000 M/N	5473 (779/4694)	1173 (167/1006)	1172 (167/1005)

Table 6: NLP split sizes. In the split columns, N is the total number of examples and pos./neg. gives the positive/negative counts.

NLP task	Train (N; pos./neg.)	Val. (N; pos./neg.)	Test (N; pos./neg.)
GoEmotions Annoy.	39069 (2223/36846)	4341 (247/4094)	5427 (320/5107)
GoEmotions Fear	39069 (536/38533)	4341 (60/4281)	5427 (78/5349)
ANLI R1 E/N	12423 (5371/7052)	667 (334/333)	667 (334/333)
ANLI R3 C/N	20000 (10000/10000)	798 (396/402)	798 (396/402)
ANLI R3 E/N	20000 (10000/10000)	804 (402/402)	804 (402/402)
SNLI High-Overlap	20000 (10000/10000)	4000 (2000/2000)	4000 (2000/2000)
WANLI C/N	20000 (10000/10000)	3540 (1540/2000)	3142 (745/2397)
WANLI E/N	20000 (10000/10000)	4000 (2000/2000)	3858 (1858/2000)

`snli` [87], and `alisawuffles/WANLI` [88]. GoEmotions uses the official text field. NLI examples are formatted as premise–hypothesis inputs. The Qwen2.5-1.5B-Instruct experiments use the same prepared binary NLP tasks as the DistilBERT experiments, so the comparison changes the pretrained model family, tokenizer, scale, and resulting LoRA trainable-subspace geometry rather than the task distribution.

Split sizes. The following tables report the exact sample counts used by the result tables. Counts are written as total examples with positive/negative counts in parentheses.

B.5 Models, Optimization, and Reporting Protocol

All deep-learning experiments were run on a server with four NVIDIA RTX 3090 GPUs, each with 24GB of memory. Each training run uses one GPU; runs are scheduled independently across available devices rather than by distributed data-parallel training. Device assignment is not used as a statistical unit. Based on the available experiment logs, all experiments involved in this paper required approximately 1.8×10^3 GPU-hours in total.

Models. The CV experiments use ImageNet-pretrained ResNet50 [21] and DenseNet121 [22] backbones with a binary classification head. The NLP experiments use DistilBERT-base-uncased [23] and Qwen2.5-1.5B-Instruct [24] sequence-classification models. The Qwen experiments use LoRA adapters [17] with rank 16, LoRA alpha 32, dropout 0.05, and bfloat16 computation. The DistilBERT runs use the same binary sequence-classification pipeline with LoRA rank 16, LoRA alpha 32, dropout 0.05, and float32 computation.

Losses. The pairwise losses are square, hinge, exponential, logistic, and squared-hinge, using the positive margin convention

$$m_\theta(x^+, x^-) = f_\theta(x^+) - f_\theta(x^-).$$

Table 7: Base training settings for the experimental regimes.

Regime	Models/tasks	Optimization settings
CV, ImageNet-LT	ResNet50 on ImageNet-LT Dogs	20 epochs, batch/eval batch $128/256$, SGD with momentum 0.9, learning rate 0.05, weight decay 10^{-4}, cosine annealing, positive batch fraction 0.25, CE-to-AUC switch LR scale 1.0.
CV, other ResNet50	Dog Breeds, CRJ700/CRJ900, APTOS Sev./Mod., DeepWeeds P/S	10 epochs, batch/eval batch $64/128$, AdamW [89], learning rate 10^{-4}, weight decay 10^{-4}, cosine annealing, positive batch fraction 0.25, CE-to-AUC switch LR scale 0.5.
CV, DenseNet121	Apple Disease, Rice Disease, APTOS Sev./Mod., HAM10000 M/N	10 epochs, batch/eval batch $48/128$, AdamW, learning rate 10^{-4}, weight decay 10^{-4}, cosine annealing, positive batch fraction 0.25, CE-to-AUC switch LR scale 0.5.
DistilBERT NLP	Eight NLP tasks in Appendix B.4	4 epochs, batch/eval batch $32/64$, gradient accumulation 1, AdamW, learning rate $2 \cdot 10^{-4}$, weight decay 0.01, cosine schedule with warm-up steps equal to 5% of total optimizer updates, maximum length 128, positive batch fraction 0.25, max gradient norm 1.0, CE-to-AUC switch LR scale 0.5.
Qwen NLP	Same eight NLP tasks	4 epochs, batch/eval batch $24/48$, gradient accumulation 1, AdamW, learning rate $8 \cdot 10^{-5}$, weight decay 0.01, cosine schedule with warm-up steps equal to 5% of total optimizer updates, maximum length 256, positive batch fraction 0.25, max gradient norm 1.0, CE-to-AUC switch LR scale 0.5.

Table 8: Warm-up-ratio and pairwise-margin grids used for pairwise AUC-surrogate runs.

Regime	Grid
CV	Warm-up ratios $0, 0.01, 0.025, 0.03, 0.05, 0.075, 0.10, 0.15, 0.20, 0.25, 0.30$. Margins $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.
DistilBERT	Warm-up ratios $0, 0.01, 0.025, 0.05, 0.075, 0.10, 0.15, 0.20, 0.25, 0.30$. Margins $0.1, 0.2, 0.4, 0.6, 0.7$.
Qwen	Warm-up ratios $0, 0.02, 0.04, 0.06, 0.08, 0.10$. Margins $0.1, 0.2, 0.4, 0.6$.

Direct pairwise AUC-surrogate training starts immediately with the specified pairwise loss. CE-warm-started AUC training uses CE for an initial fraction of the update budget and then switches to the specified pairwise loss.

Base training settings. The following settings define the training configurations used in the final tables. ImageNet-LT uses a separate ResNet50 configuration because the training split is much larger than the low-shot and medium-scale CV tasks. APTOS Sev./Mod. is the only CV task that appears under both ResNet50 and DenseNet121; it uses the same binary construction and split protocol in both rows. The positive batch fraction fixes the class-batch ratio used in the experiments. Thus the implementation uses unequal class counts B_+ and B_-, while the B-positive and B-negative notation in the main text gives the same B^{-1} versus B^{-2} Hoeffding scaling up to constants.

Warm-up and margin grids. The result grids tune the CE warm-up ratio and the pairwise margin. Direct pairwise AUC-surrogate training is the zero-warm-up case.

Configuration selection and statistics. Within every individual run, train, validation, and test AUC are evaluated once per epoch, so candidate checkpoints are the end-of-epoch states and no intra-epoch checkpoint is used for table selection. The test AUC reported for that run is taken from the epoch with the largest validation AUC. The table-generation step groups runs with identical model, task, loss, mode, warm-up ratio, margin, epoch count, batch size, and learning rate. Each configuration is

summarized by the arithmetic mean and sample standard deviation of the validation-best test AUC values across matching runs. Hyperparameters are selected by mean validation AUC across seeds; the reported value is the corresponding held-out test AUC. For the expanded direct/warm-up tables, ties are broken by smaller sample standard deviation.

Matched initialization control. To isolate initialization from backbone architecture, we compare random and ImageNet-pretrained initialization over 45 matched settings, holding the backbone, task, surrogate, direct/warm-up configuration, and model-selection rule fixed. Warm-up gain is computed within each setting as the validation-selected test AUC after CE-warm-started pairwise training minus that after direct pairwise training. All matched settings are retained, and the mean and interquartile range are reported. The comparison is treated as descriptive because no paired inferential test is added.

Splits, seeds, and reproducibility. The reported seed set is $\{123, 124, 125, 126, 127\}$ for all experiments. Means, sample standard deviations, and validation-selected hyperparameters are computed over the corresponding seed set for each model family. Seeds are used for model initialization, data-loader shuffling, stratified batch sampling, and any stratified split construction when the source task does not provide the exact split used by the pipeline. Thus the split construction is reproducible from the stated seed and task script, while the split-size tables in Appendix B.4 give the resulting positive and negative counts. Prepared NLI splits are written once before model training; GoEmotions one-versus-rest runs use the official test split and a stratified validation split carved from the official training split by seed.

Implementation environment. The experiments are implemented in Python with PyTorch, torchvision, Hugging Face Transformers, Hugging Face Datasets, PEFT, NumPy, pandas, and scikit-learn ROC-AUC evaluation. The recorded Python environment uses Python $3.12.10$, PyTorch $2.6.0+\mathrm{cu124}$, torchvision $0.21.0+\mathrm{cu124}$, Transformers $4.51.3$, tokenizers $0.21.1$, huggingface-hub $0.30.2$, safetensors $0.5.3$, NumPy $2.1.2$, pandas $2.3.3$, and scikit-learn $1.7.2$. CV preprocessing uses torchvision image transforms, and NLP preprocessing uses Hugging Face tokenizers with dynamic padding.

Evaluation, checkpointing, and mechanism logs. Validation and test AUC are computed with the standard ROC-AUC estimator. The same training pipeline also records the pairwise quantities used for the mechanism measurements in Figures 3 and 4. These mechanism measurements are computed on logged training mini-batches rather than on the held-out test set. Scores are converted to sigmoid scores before pairwise margin statistics are computed. This logged sigmoid-margin scale is used for cross-model comparability; it only changes the empirical logging scale and does not change the score-margin convention or theoretical results. Thus the logged severe-misranking thresholds are fixed thresholds in sigmoid-score difference units. The primary unthresholded statistic is the logged-batch misranking ratio

$$\mathbb{P}\{m_\theta(X^+, X^-) < 0\},$$

which measures the fraction of logged positive-negative training pairs whose score margin has the wrong sign. We additionally log thresholded severe-misranking ratios at thresholds -0.01, -0.025, -0.05, -0.075, -0.1, -0.15, and -0.2. At the start of pairwise training, namely the first pairwise update for direct pairwise training and the first update after the CE-to-AUC switch for warm-started training, logged positive-negative pairs are binned by their current margin. Figure 4(c,d) uses disjoint wrong-pair intervals I and reports the relative weight proportion

$$P_I^{\mathrm{rel}} = \frac{\sum_{(i,j):\, m_{ij} \in I} q_{ij} \Big/ \sum_{i=1}^{B_+} \sum_{j=1}^{B_-} q_{ij}}{\#\{(i,j) : m_{ij} \in I\}/(B_+ B_-)}, \qquad q_{ij} = \left\| w_\ell(m_{ij}) \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+) \right\|_2.$$

Thus the numerator is the proportion of weighted pair-gradient contribution assigned to interval I, and the denominator is the pair proportion in the same interval. This quantity is the empirical counterpart of the surrogate-weight calculations in Appendix E: a value above 1 means that the interval receives more gradient weight than its frequency alone would imply, and a value below 1 means that it receives less. These logs support the two state-level checks used in Section 5: favorable surrogate weights should place more relative gradient weight on severe wrong pairs, and CE warm-up should reduce the misranking ratio at the point where pairwise AUC optimization begins.

For checks involving the severe-misranking mass and correction-direction criterion in Appendix D, the direction is fixed mechanically before the projected quantities are measured. On a selection batch, define the severe-pair set $\mathcal{S}_M(\theta) = \{(i,j) : m_{ij} \le -M\}$. The empirical severe-pair objective on this batch has first-order directional change proportional to

$$\frac{1}{|\mathcal{S}_M(\theta)|} \sum_{(i,j)\in\mathcal{S}_M(\theta)} w_\ell(m_{ij}) \left\langle \boldsymbol{g}_j^- - \boldsymbol{g}_i^+, \boldsymbol{u} \right\rangle .$$

Thus the vector below is the finite-batch severe-error normal: its negative gives the steepest local direction for reducing the surrogate-weighted severe-pair margin error among unit directions in the trainable parameter space. The sign convention is immaterial for the variance term and for the absolute projected update size reported in the mechanism measurements. When $\mathcal{S}_M(\theta)$ is nonempty, compute

$$\widehat{\boldsymbol{u}}_{\mathrm{sev}} = \frac{1}{|\mathcal{S}_M(\theta)|} \sum_{(i,j)\in\mathcal{S}_M(\theta)} w_\ell(m_{ij}) \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+).$$

If $\mathcal{S}_M(\theta)$ is empty or $\|\widehat{\boldsymbol{u}}_{\mathrm{sev}}\|_2$ is negligible, the checkpoint is recorded as having no stable severe-pair correction direction. Otherwise, the empirical severe-error direction is

$$\widehat{\boldsymbol{v}}_{\mathrm{sev}} = -\frac{\widehat{\boldsymbol{u}}_{\mathrm{sev}}}{\|\widehat{\boldsymbol{u}}_{\mathrm{sev}}\|_2}. \tag{1}$$

This matches the steepest-decrease orientation in Appendix D. The deterministic rule is applied to every method, task, seed, and checkpoint before the projected update and Hoeffding terms are computed. On an independent logging batch, write

$$\widehat{\psi}_{ij} = \left\langle w_\ell(m_{ij}) \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+), \widehat{\boldsymbol{v}}_{\mathrm{sev}} \right\rangle, \qquad \overline{\psi} = \frac{1}{B_+ B_-} \sum_{i=1}^{B_+} \sum_{j=1}^{B_-} \widehat{\psi}_{ij}.$$

The projected update size is $|\overline{\psi}|$. The first-order Hoeffding variance estimate uses the row and column means

$$\widehat{r}_i = \frac{1}{B_-} \sum_{j=1}^{B_-} \widehat{\psi}_{ij}, \qquad \widehat{c}_j = \frac{1}{B_+} \sum_{i=1}^{B_+} \widehat{\psi}_{ij},$$

and records

$$\widehat{V}_1 = \widehat{\mathrm{Var}}_i(\widehat{r}_i - \overline{\psi}) + \widehat{\mathrm{Var}}_j(\widehat{c}_j - \overline{\psi}).$$

The first-order variance term reported in Figure 3(a,b) is the scaled quantity

$$\widehat{V}_{1,\mathrm{scaled}} = \frac{\widehat{\mathrm{Var}}_i(\widehat{r}_i - \overline{\psi})}{B_+} + \frac{\widehat{\mathrm{Var}}_j(\widehat{c}_j - \overline{\psi})}{B_-}.$$

This is the finite-batch counterpart of the order-B^{-1} term in Proposition 1. The residual pair-interaction estimate is obtained from

$$\widehat{V}_2 = \widehat{\mathrm{Var}}_{i,j}(\widehat{\psi}_{ij} - \widehat{r}_i - \widehat{c}_j + \overline{\psi}),$$

which enters the same Hoeffding split at scale $(B_+B_-)^{-1}$, reducing to B^{-2} in the equal-size case. Proposition 14 gives the finite-pair check for severe-misranking mass, and Proposition 9 states when severe-pair compression makes $\widehat{V}_1$ small.

Practical diagnose–intervene–recheck protocol. First, use the logging-batch severe-misranking ratio as an error gate: if it decreases, continue direct pairwise training; if it remains high, proceed to the motion gate. Second, construct $\widehat{\boldsymbol{v}}_{\mathrm{sev}}$ on a selection batch and estimate the projected drift $|\overline{\psi}|$ and first-order variance $\widehat{V}_{1,\mathrm{scaled}}$ on an independent logging batch. If either source of corrective motion remains substantial, continue direct pairwise training; if both are weak, apply CE warm-up and repeat both checks before switching to pairwise training. Among the five losses studied here, squared-hinge provides the intended severity-increasing and easy-pair-suppressing weight profile; other surrogates with the same profile can be used analogously. These diagnostics are interpreted along the observed training trajectory rather than through a universal numerical threshold.

Fixed-checkpoint batch-size variance scaling. We measure fixed-checkpoint batch-size scaling on APTOS, SNLI, and ANLI-R1. At each checkpoint, the empirical severe-error direction $\widehat{\boldsymbol{v}}_{\mathrm{sev}}$ is constructed by Eq. (1) from a fixed reference pool of 64 positive and 64 negative examples and is then held fixed across all batch-size measurements at that checkpoint. A disjoint logging pool contains 128 examples per class. Matched checkpoints use the same sample indices but compute their own checkpoint-specific directions.

Severe pairs satisfy $f_\theta(x^+) - f_\theta(x^-) \le -0.05$ on the sigmoid-score difference scale used by the mechanism logs. For each per-class batch size $B \in \{4, 8, 16, 32, 64\}$, we use $5{,}000$ with-replacement draws from the logging pool and float64 aggregation to estimate the variance of the severe-truncated pairwise mini-batch update projected onto $\widehat{\boldsymbol{v}}_{\mathrm{sev}}$. We fit the log empirical variance against $\log B$. Each setting is repeated over three independent reference/logging-pool splits, and Table 9 reports the ranges of severe projected drift and fitted slope across the three splits. Slope stability is checked with $1{,}000$ bootstrap repetitions. Good and bad regions are identified by severe projected drift before the slopes are fitted.

For equal positive and negative batch sizes, the measured projected variance uses the decomposition in Eq. (12), written as

$$\mathrm{Var}(\widehat{g}_B) = \frac{V_1}{B} + \frac{V_2}{B^2},$$

where V_1 is the sum of the first-order row and column Hoeffding projection variances and V_2 is the residual pair-interaction variance.

Table 9: Fixed-checkpoint batch-size scaling of the severe-truncated projected variance. Good and bad regions are determined by severe projected drift. Entries report ranges over three independent reference/logging-pool splits.

Task / surrogate	Region	Severe projected drift	Log-variance–log-B slope
APTOS / direct square	Good	$+0.273$ to $+1.043$	-1.102 to -1.070
APTOS / direct square	Bad	-0.187 to $+0.105$	-1.250 to -1.185
APTOS / direct exponential	Good	$+0.574$ to $+1.304$	-1.099 to -1.081
APTOS / direct exponential	Bad	-0.104 to $+0.164$	-1.470 to -1.351
APTOS / direct logistic	Good	$+0.552$ to $+0.687$	-1.091 to -1.060
APTOS / direct logistic	Bad	-0.003 to $+0.016$	-1.562 to -1.412
SNLI / direct squared-hinge	Good	$+0.213$ to $+0.384$	-1.091 to -1.032
SNLI / direct squared-hinge	Bad	-0.0004 to $+0.0011$	-1.595 to -1.492
ANLI-R1 / direct squared-hinge	Good	$+0.107$ to $+0.151$	-1.062 to -1.038
ANLI-R1 / direct squared-hinge	Bad	$+0.011$ to $+0.037$	-1.201 to -1.141

All variance calculations record the corresponding float64 numerical floor. The smallest measured severe projected variance remains 1.39×10^9 times its corresponding numerical floor. As a separate numerical control, applying the same scaling procedure to an exact-degeneracy construction satisfying Corollary 7 gives a log-variance–log-B slope of approximately -2.000; this controlled result is not included in Table 9.

Figure 5 extends the trajectory check in Figure 3(c,d). On GoEmotions Annoyance and SNLI E/N, direct training remains visibly separated from warm-started training for square, hinge, exponential, and logistic losses, while squared-hinge shows much smaller separation. This supports the use of the main-text trajectory panels as representative examples rather than isolated cases.

Figure 6 complements Figure 5 by plotting the corresponding first-order Hoeffding variance term rather than the misranking ratio. The panels use the same correction-direction construction as Figure 3(a,b), but show all five pairwise losses on two representative NLP tasks.

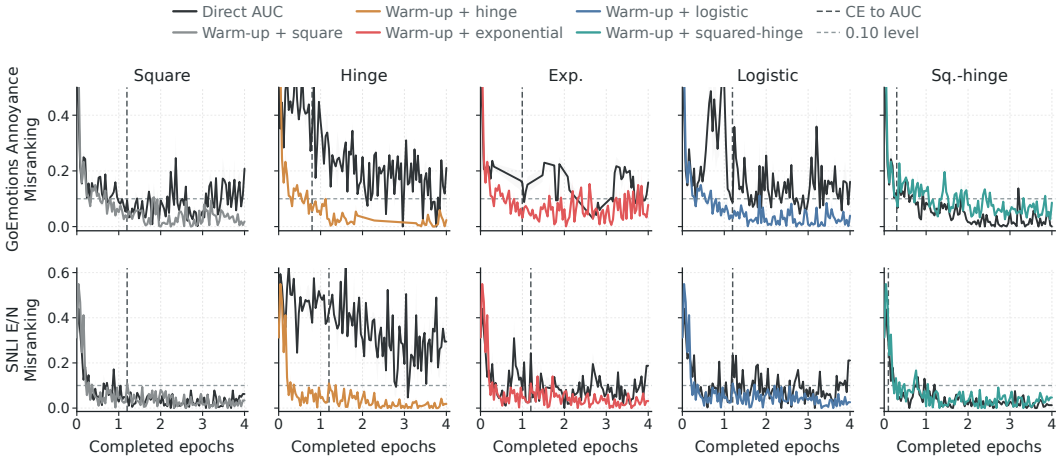


Figure 5: Misranking trajectories for all five pairwise losses on DistilBERT GoEmotions Annoyance and SNLI E/N. Black curves are direct pairwise training; colored curves are CE warm-up followed by AUC training, with loss colors shown in the legend. Vertical dashed lines mark the CE-to-AUC switch.

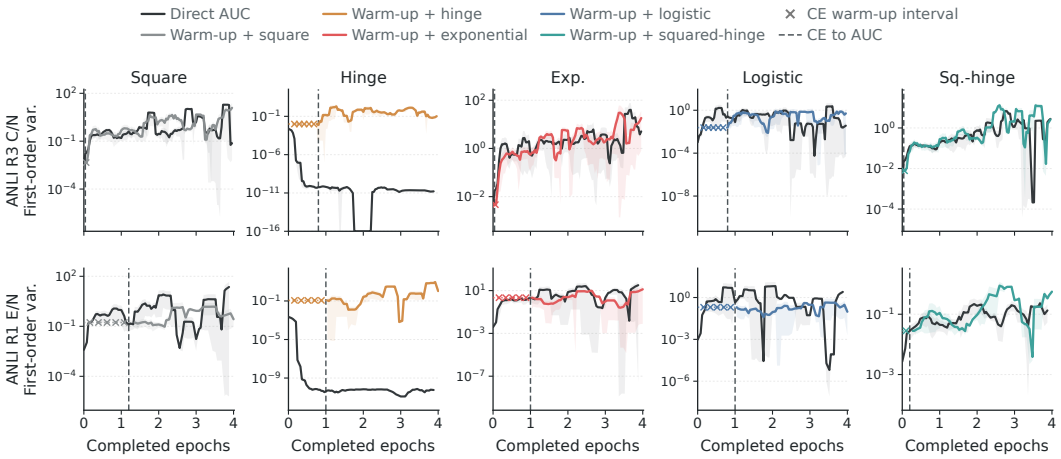


Figure 6: Scaled first-order Hoeffding variance trajectories for all five pairwise losses on DistilBERT ANLI R3 C/N and ANLI R1 E/N. Black curves are direct pairwise training; colored curves are CE warm-up followed by AUC training, with loss colors shown in the legend. Shaded bands show one standard error across matching runs after the same trajectory smoothing used for the displayed mean. x marks denote the CE warm-up interval before the pairwise variance term is recorded, and vertical dashed lines mark the CE-to-AUC switch.

Matched directional intervention. We run matched interventions from fixed direct-hinge checkpoints on ImageNet-LT dog-tail versus siblings and APTOS severe-plus versus moderate. At each checkpoint, the empirical severe-pair normal is computed from a fixed reference pool of 64 positive and 64 negative training examples with $M = 0$, and $\widehat{\boldsymbol{v}}_{\mathrm{sev}}$ is obtained by Eq. (1) and then held fixed throughout the intervention. A disjoint pool of 128 examples per class is retained for mechanism logging. All arms use training/data seed 123 and pool seed 20260725, and share the same checkpoint, data order, augmentations, optimizer, and learning-rate schedule.

For ImageNet-LT dog-tail versus siblings, the intervention starts from update 690, uses learning rate 0.025 and per-update parameter L_2 magnitude $\sigma = 0.05$, and continues for three epochs (207 optimizer updates) over 16 noise seeds 20262000–20262015. For APTOS severe-plus versus moderate, it starts from update 110, uses learning rate 5×10^{-5} and $\sigma = 0.01$, and continues for three epochs (66 optimizer updates) over 24 noise seeds 20263000–20263023. All 40 realizations are retained.

After each optimizer step, the directional arm adds $\epsilon_t \sigma \widehat{\boldsymbol{v}}_{\mathrm{sev}}$, where the $\epsilon_t \in \{-1, +1\}$ are independent Rademacher signs. The matched isotropic arm adds an independent normalized Gaussian direction with the same per-update L_2 magnitude, refreshed at every update. One deterministic no-noise run provides the shared baseline for each task. Endpoint effects are measured at the end of the third intervention epoch on the test set. Let $\Delta\mathrm{AUC}$ and ΔS_{sev} denote the changes in test AUC and test severe-misranking mass $S_{\mathrm{sev}} = \mathbb{P}\{m_\theta(X^+, X^-) \le 0\}$ relative to this baseline. A realization is classified as rescue when $\Delta\mathrm{AUC} \ge 0.5$ and $\Delta S_{\mathrm{sev}} \le -0.5$ percentage points, as harm when $\Delta\mathrm{AUC} \le -0.5$ and $\Delta S_{\mathrm{sev}} \ge 0.5$ percentage points, and as no clear effect otherwise. The same rule is applied to the directional and isotropic arms.

Table 10: Matched directional-intervention data at fixed direct-hinge checkpoints. R/H/N denotes rescue/harm/no clear effect. Endpoint changes are measured at the final intervention epoch relative to the shared no-noise baseline and are reported in percentage points; D/I denotes directional/isotropic.

Task	n	Directional R/H/N	Isotropic R/H/N	Mean ΔAUC D/I	Mean ΔS_{sev} D/I
ImageNet-LT dog-tail	16	2/6/8	0/0/16	$-2.541/-0.001$	$+6.036/-0.046$
APTOS	24	6/13/5	1/1/22	$-1.316/+0.035$	$+1.317/-0.033$
Total	40	8/19/13	1/1/38	—	—

B.6 Multiclass, Multilabel, and Deep-AUC Extensions

Small-scale multiclass experiments. We evaluate CE, direct hinge, warm-up hinge, direct squared-hinge, and warm-up squared-hinge on three long-tailed multiclass tasks. Aircraft Airbus-6 LT contains the A318, A319, A320, A321, A330-200, and A330-300 classes, with $93/200/200$ training/validation/test examples. CIFAR-100 Medium-Mammals-5 LT contains camel, cattle, chimpanzee, elephant, and kangaroo, with $431/250/250$ examples. CIFAR-100 Vehicles-10 LT contains bicycle, bus, motorcycle, pickup truck, train, lawn mower, rocket, streetcar, tank, and tractor, with $818/500/500$ examples.

The experiments use a randomly initialized ResNet18 trained for 10 epochs, with 24 batches per epoch, 64×64 inputs, learning rate 0.002, and weight decay 10^{-4}. Pairwise training uses a macro one-vs-rest objective over class logits, and checkpoints are selected by validation Macro-OVO-AUC.

Table 11: Results on three long-tailed multiclass tasks (AUC%).

Multiclass task	**CE**	**Direct hinge**	**Warm-up hinge**	**Direct squared-hinge**	**Warm-up squared-hinge**
Aircraft Airbus-6 LT	60.92	61.43	61.61	**61.68**	60.91
CIFAR-100 Medium-Mammals-5 LT	81.08	81.13	82.91	83.62	**84.44**
CIFAR-100 Vehicles-10 LT	86.47	88.30	88.64	87.97	**89.08**

Multilabel experiments. We also evaluate multilabel AUC optimization on CheXpert. Competition-5 uses Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion; Cardiac/Fluid-4

uses Cardiomegaly, Edema, Atelectasis, and Pleural Effusion; and Consolidation/Cardiac-2 uses Consolidation and Cardiomegaly. Each image may have multiple positive labels.

The experiments use an ImageNet-pretrained DenseNet121 with a randomly initialized multilabel head. Under the U-Zero rule, uncertain labels (-1) and missing labels are mapped to zero. Patient-disjoint training and internal-validation splits are constructed from `train.csv`, and the official `valid.csv` is used only as the test set. The training and internal-validation sets contain at most $12{,}000$ and $3{,}000$ examples, respectively. Models are trained for 8 epochs with batch size 64, 224×224 inputs, learning rate 10^{-4}, and weight decay 10^{-4}.

The pairwise loss is computed separately between positive and negative examples for each label and macro-averaged over labels that contain both classes in the current batch. Checkpoints are selected by internal-validation macro ROC-AUC, and the final metric is the macro-average of label-wise test ROC-AUC.

Table 12: Results on three CheXpert multilabel tasks (AUC%).

Multilabel task	BCE	Direct hinge	Warm-up hinge	Direct squared-hinge	Warm-up squared-hinge
Competition-5	80.86	82.75	**83.40**	81.51	83.09
Cardiac/Fluid-4	80.76	81.14	**82.13**	79.96	81.44
Consolidation/Cardiac-2	51.52	57.28	56.09	57.08	**59.69**

Large-scale multiclass experiments. We further evaluate large-scale multiclass learning on iNaturalist 2018 and ImageNet-LT, retaining their natural long-tailed training distributions without class resampling. iNaturalist 2018 contains $437{,}513$ training examples and $24{,}426$ official validation examples, for $461{,}939$ samples over $8{,}142$ classes. ImageNet-LT contains $115{,}845$ valid training examples, $20{,}000$ validation examples, and $50{,}000$ test examples, for $185{,}845$ samples over $1{,}000$ classes.

The iNaturalist 2018 experiment uses an ImageNet-supervised pretrained ResNet50 with a randomly initialized $8{,}142$-class head. It is trained for 3 epochs with training and evaluation batch size 128 and 128×128 inputs; the official validation set is evaluated once after the fixed training budget. The ImageNet-LT experiment uses a Places365-supervised pretrained ResNet50 [90] with a randomly initialized classification head. It is trained for 4 epochs with training batch size 128, evaluation batch size 256, and 128×128 inputs; checkpoints are selected by validation Macro-OVO-AUC, and the test set is evaluated only after the selected checkpoint is restored.

We compare CE, direct hinge, warm-up hinge, direct squared-hinge, and warm-up squared-hinge, and report Macro-OVO-AUC and Macro-OVR-AUC.

Table 13: Large-scale multiclass results on iNaturalist 2018 and ImageNet-LT (AUC%).

Dataset	Metric	CE	Hinge direct	Hinge warm-up	Sq.-hinge direct	Sq.-hinge warm-up
iNaturalist	Macro-OVO-AUC	96.64	92.45	99.70	89.18	**99.74**
ImageNet-LT	Macro-OVO-AUC	98.50	97.87	98.45	97.69	**98.91**
iNaturalist	Macro-OVR-AUC	97.70	88.72	97.31	79.94	**98.13**
ImageNet-LT	Macro-OVR-AUC	94.80	93.21	94.81	92.95	**96.08**

Deep-AUC comparison. We compare the results in this paper with Deep-AUC as implemented by LibAUC [91] using `AUCMLoss(v2)` and the PESG optimizer. `AUCMLoss(v2)` uses the squared AUC-margin objective with margin 1.0; PESG uses SGD mode, momentum 0.9, epoch decay 0.002, weight decay 10^{-5}, and positive fraction 0.25.

The squared AUC-margin objective can be reconstructed with auxiliary variables in a min–max form, avoiding explicit construction of all positive–negative pairs at each update. This reconstruction uses class-dependent first- and second-order statistics. In contrast, hinge, logistic, exponential, and squared-hinge assign weights that depend on each pair's margin and generally cannot be reconstructed from fixed class means and second moments alone.

More explicitly, for a target margin c, the quadratic objective

$$\mathbb{E}\big[(c - f_\theta(X^+) + f_\theta(X^-))^2\big]$$

expands exactly into the positive- and negative-class score means, second moments, and their interaction. Auxiliary primal and dual variables represent these quantities as an instance-wise min–max problem, thereby removing explicit pair construction and the corresponding two-sample pairwise estimator for this quadratic objective. By contrast, the margin-dependent weights of hinge, logistic, exponential, and squared-hinge are not generally determined by fixed class means and second moments; reproducing them without explicit pairs would require a different loss-specific reconstruction.

Table 14: Deep-AUC and our results on 17 task–model settings (AUC%).

Task–model setting	**Deep-AUC**	**Ours**
Apple Disease–DenseNet121	55.56	**62.69**
Rice Disease–DenseNet121	**77.76**	75.71
APTOS Sev./Mod.–DenseNet121	82.12	**84.20**
HAM10000 M/N–DenseNet121	97.16	**97.70**
ImageNet-LT Dogs–ResNet50	94.56	**95.83**
Dog Breeds–ResNet50	85.74	**87.00**
CRJ700/CRJ900–ResNet50	53.52	**56.62**
APTOS Sev./Mod.–ResNet50	82.40	**83.92**
DeepWeeds P/S–ResNet50	98.43	**98.72**
GoEmotions Annoyance–DistilBERT	**84.51**	84.38
GoEmotions Fear–DistilBERT	96.11	**96.48**
ANLI R1 E/N–DistilBERT	**71.67**	70.21
ANLI R3 C/N–DistilBERT	**62.63**	62.47
ANLI R3 E/N–DistilBERT	63.65	**63.87**
SNLI E/N–DistilBERT	92.94	**92.98**
WANLI C/N–DistilBERT	82.38	**82.41**
WANLI E/N–DistilBERT	75.40	**75.44**

Deep-AUC is stronger in 4 of the 17 task–model settings, while our validation-selected direct configuration is stronger in 13. The comparison identifies Deep-AUC as a strong scalable solution when its quadratic reconstruction is appropriate, while showing that direct pairwise optimization remains competitive and supports a broader range of surrogate dynamics.

B.7 Assets, Licenses, and Terms

The experiments use existing public datasets, pretrained models, and standard software assets. Dataset sources and task constructions are listed in Appendix B.4 and Appendix B.6. The CV experiments use ImageNet-LT split files derived from ImageNet, Stanford Dogs, FGVC-Aircraft, CIFAR-100, iNaturalist 2018, CheXpert, APTOS 2019, DeepWeeds, D-KAP, Rice Leaf Diseases, and HAM10000. The NLP experiments use the Hugging Face releases or prepared splits of GoEmotions, ANLI, SNLI, and WANLI. The model assets are randomly initialized ResNet18, ImageNet- or Places365-supervised pretrained ResNet50, ImageNet-pretrained DenseNet121, DistilBERT-base, and Qwen2.5-1.5B-Instruct; the DistilBERT and Qwen NLP runs both use LoRA as the parameter-efficient fine-tuning method. The Deep-AUC comparison uses LibAUC `AUCMLoss(v2)` and PESG.

The asset licenses and terms are as follows. ImageNet/ImageNet-LT and Stanford Dogs are used under ImageNet-style non-commercial research and educational terms; FGVC-Aircraft is non-commercial research only, with original photographer copyright retained. CIFAR-100 is distributed through its official project page and used subject to its source terms; iNaturalist 2018 is used subject to the competition data terms and the licenses attached to its images; and CheXpert is obtained through Stanford's official release channel and used under its data-use terms. APTOS 2019 224×224 is CC0 1.0; DeepWeeds is CC BY 4.0; D-KAP is listed as "Database: Open Database, Contents: Database Contents"; the Rice Leaf Diseases release used here does not list a license; and HAM10000 is CC BY-NC-SA 4.0 in the Kaggle release used here. GoEmotions is Apache 2.0, ANLI is CC BY-NC 4.0, SNLI is CC BY-SA 4.0, and WANLI is CC BY 4.0. DistilBERT-base-uncased and Qwen2.5-1.5B-Instruct are Apache 2.0; TorchVision ImageNet-pretrained weights are used subject to

TorchVision notices and ImageNet-derived terms, and Places365-pretrained weights are used subject to their release terms. The software stack uses standard open-source licenses, including Apache 2.0 for Hugging Face Transformers/Datasets/PEFT and BSD-style licenses for PyTorch/TorchVision, NumPy, pandas, and scikit-learn; LibAUC is used under its project license.

Source and use details for these datasets and models appear in Appendix B.4, Appendix B.5, and Appendix B.6. We used the assets for research experiments and did not redistribute the original datasets, pretrained checkpoints, or derived datasets as part of this manuscript. Users who reproduce the experiments should obtain each dataset and pretrained model from the official source or public release and follow the corresponding license, data-use terms, and model-use terms. Any anonymized code or reproduction package should include its own license file and instructions for separately downloading third-party assets from their official sources.

B.8 Ethics, Broader Impacts, and Safeguards

This work studies optimization dynamics for AUC training and does not introduce a deployed decision system. The potential positive impact is improved understanding of when pairwise AUC optimization is reliable under class imbalance and ranking-based evaluation. Potential negative impact could arise if AUC-optimized models are deployed in high-stakes domains without suitable validation, calibration, uncertainty assessment, subgroup analysis, or human oversight. The experiments include public medical-image and natural-language datasets, so any downstream use of the trained models should follow the dataset terms and the safeguards required by the target application.

We do not release a new high-risk model, scraped dataset, or human-subject dataset. The manuscript reports experiments on existing public datasets and pretrained models, and it does not involve new data collection, crowdsourcing, interaction with human subjects, or interventions on human participants. Accordingly, no new compensation protocol, participant instruction text, or institution-specific IRB information is part of this work. No additional release safeguard is introduced beyond respecting the original asset terms.

C Direct Pairwise SGD as a Stochastic Dynamics

This section fixes the direct pairwise SGD process used in the paper and states its local stochastic-dynamics interface. When each step uses a fresh positive-negative mini-batch, the parameter update is Markov in the current parameter: its drift is the negative population pairwise gradient, and its covariance is the mini-batch covariance later decomposed by Hoeffding projections.

Pairwise setup. We use the positive-margin convention

$$m_\theta(x^+, x^-) = f_\theta(x^+) - f_\theta(x^-).$$

For a pairwise surrogate ℓ, define

$$R_\ell(\theta) = \mathbb{E}[\ell(m_\theta(X^+, X^-))], \qquad w_\ell(x) = -\ell'(x),$$

where $X^+ \sim P_+$, $X^- \sim P_-$, and the two draws are independent. Given B_+ positives and B_- negatives, let

$$m_{ij} = f_\theta(x_i^+) - f_\theta(x_j^-), \qquad \boldsymbol{g}_i^+ = \nabla_\theta f_\theta(x_i^+), \qquad \boldsymbol{g}_j^- = \nabla_\theta f_\theta(x_j^-).$$

The direct pairwise mini-batch gradient is

$$\widehat{\boldsymbol{g}}^\ell_{B_+,B_-}(\theta) = \frac{1}{B_+ B_-} \sum_{i=1}^{B_+} \sum_{j=1}^{B_-} w_\ell(m_{ij}) \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+). \tag{2}$$

When $B_+ = B_- = B$, we write $\widehat{\boldsymbol{g}}^\ell_B(\theta) = \widehat{\boldsymbol{g}}^\ell_{B,B}(\theta)$, so the equal-size specialization is

$$\widehat{\boldsymbol{g}}^\ell_B(\theta) = \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} w_\ell(m_{ij}) \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+). \tag{3}$$

For local uses, we follow the localization convention in Li et al. [70, Remark 11]: when the smoothness and Lipschitz assumptions hold only on a neighborhood, the analysis is applied to the stopped process

until it exits that neighborhood, or to a globally regularized objective that agrees with the original objective on the neighborhood. The drift and covariance identities below are unchanged before the exit time.

Theorem 2 (Markov-state finite-time weak SGD diffusion bridge)**.** *Let $(\Gamma_k)_{k\ge 0}$ be i.i.d. random variables, let $F(\cdot,\Gamma_k)$ be random objectives on $\mathbb{R}^q$, and let the Markov state $\boldsymbol{y}_k \in \mathbb{R}^q$, initialized at $\boldsymbol{y}_0 = \boldsymbol{y}_{\text{init}}$, evolve by constant-step SGD*

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \eta \cdot \nabla F(\boldsymbol{y}_k, \Gamma_k), \qquad k = 0, \dots, N-1, \quad N = \lfloor T/\eta \rfloor.$$

Assume the random objective family satisfies the hypotheses of Li et al. [70, Corollary 10] with $f_\gamma = F(\cdot,\Gamma_0)$ and $f = \bar{F}$, where

$$\bar{F}(\boldsymbol{y}) = \mathbb{E}[F(\boldsymbol{y},\Gamma_0)].$$

Define

$$\boldsymbol{A}(\boldsymbol{y}) = \mathrm{Cov}(\nabla F(\boldsymbol{y},\Gamma_0)).$$

Then the Itô SDE

$$d\widetilde{\boldsymbol{y}}_t = -\nabla \bar{F}(\widetilde{\boldsymbol{y}}_t)\, dt + (\eta \cdot \boldsymbol{A}(\widetilde{\boldsymbol{y}}_t))^{1/2}\, d\boldsymbol{W}_t, \qquad \widetilde{\boldsymbol{y}}_0 = \boldsymbol{y}_{\text{init}},$$

where $\boldsymbol{W}_t$ is a standard Brownian motion in $\mathbb{R}^q$, is an order-1 finite-time weak approximation to the discrete dynamics: for every test function g in the G_2 class of Li et al. [70, Corollary 10], there is a constant $C_{T,g}$ independent of η such that

$$\max_{0\le k\le N} |\mathbb{E}[g(\boldsymbol{y}_k)] - \mathbb{E}[g(\widetilde{\boldsymbol{y}}_{k\eta})]| \le C_{T,g} \cdot \eta.$$

Proof. The state process is Markov by assumption: the conditional law of $\boldsymbol{y}_{k+1}$ given the past is determined by the current state $\boldsymbol{y}_k$ and a fresh noise variable Γ_k. Thus the update has the constant-step stochastic-gradient form required by Li et al. [70, Corollary 10]:

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \eta \cdot \nabla F(\boldsymbol{y}_k, \Gamma_k).$$

The mean objective is

$$\bar{F}(\boldsymbol{y}) = \mathbb{E}[F(\boldsymbol{y},\Gamma_0)].$$

Under the exchange-of-derivative and regularity conditions included in Li et al. [70, Corollary 10],

$$\mathbb{E}[\nabla F(\boldsymbol{y},\Gamma_0)] = \nabla \bar{F}(\boldsymbol{y}).$$

Hence the conditional first moment of one step is

$$\mathbb{E}[\boldsymbol{y}_{k+1} - \boldsymbol{y}_k \mid \boldsymbol{y}_k = \boldsymbol{y}] = -\eta \cdot \nabla \bar{F}(\boldsymbol{y}),$$

so the continuous-time drift is $-\nabla \bar{F}(\boldsymbol{y})$. The stochastic-gradient covariance is

$$\mathrm{Cov}(\nabla F(\boldsymbol{y},\Gamma_k) \mid \boldsymbol{y}_k = \boldsymbol{y}) = \mathrm{Cov}(\nabla F(\boldsymbol{y},\Gamma_0)) = \boldsymbol{A}(\boldsymbol{y}),$$

where the second equality uses the identical distribution of the noise variables. Therefore the finite-time weak modified equation has diffusion coefficient

$$(\eta \cdot \boldsymbol{A}(\boldsymbol{y}))^{1/2}.$$

Applying Li et al. [70, Corollary 10] with $f_\gamma = F(\cdot,\Gamma_0)$, mean objective $\bar{F}$, and covariance field $\boldsymbol{A}$ gives the displayed SDE and the weak-error bound for every G_2 test function g. □

Proposition 4 (Plain pairwise mini-batch gradients are unbiased)**.** *Assume $\nabla_\theta \ell(m_\theta(X^+, X^-))$ is integrable and differentiation can be exchanged with expectation. Then the direct pairwise mini-batch gradient in Eq. (2) satisfies*

$$\mathbb{E}[\widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)] = \nabla R_\ell(\theta).$$

The equal-size shorthand $\widehat{\boldsymbol{g}}^{\ell}_{B}$ satisfies the same identity.

Proof. For one independent positive-negative pair,

$$\nabla_\theta \ell(m_\theta(X^+, X^-)) = \ell'(m_\theta(X^+, X^-)) \cdot \big(\nabla f_\theta(X^+) - \nabla f_\theta(X^-)\big).$$

Since $w_\ell = -\ell'$, the same vector can be written as

$$w_\ell(m_\theta(X^+, X^-)) \cdot \big(\nabla f_\theta(X^-) - \nabla f_\theta(X^+)\big).$$

The exchange assumption gives

$$\nabla R_\ell(\theta) = \mathbb{E}[\nabla_\theta \ell(m_\theta(X^+, X^-))].$$

Every pair (X_i^+, X_j^-) inside the mini-batch has the same product distribution as (X^+, X^-), although the B_+B_- pair terms are not mutually independent. Therefore linearity of expectation yields

$$\mathbb{E}[\widehat{\boldsymbol{g}}^\ell_{B_+,B_-}(\theta)] = \frac{1}{B_+B_-}\sum_{i=1}^{B_+}\sum_{j=1}^{B_-}\mathbb{E}[w_\ell(m_\theta(X_i^+, X_j^-)) \cdot (\nabla f_\theta(X_j^-) - \nabla f_\theta(X_i^+))] = \nabla R_\ell(\theta).$$

□

Proposition 5 (One-step conditional moments for pairwise dynamics)**.** *Consider the parameter dynamics*

$$\theta_{k+1} = \theta_k - \eta_k \cdot \widehat{\boldsymbol{g}}^\ell_{B_+,B_-,k}(\theta_k),$$

where, conditionally on θ_k, $\widehat{\boldsymbol{g}}^\ell_{B_+,B_-,k}(\theta_k)$ is computed from a fresh positive-negative mini-batch. Let

$$\Delta_k = \theta_{k+1} - \theta_k.$$

Under Proposition 4,

$$\mathbb{E}[\Delta_k \mid \theta_k] = -\eta_k \cdot \nabla R_\ell(\theta_k).$$

If the conditional second moment is finite, then

$$\mathbb{E}[\Delta_k \cdot \Delta_k^\top \mid \theta_k] = \eta_k^2 \cdot \left\{\mathrm{Cov}(\widehat{\boldsymbol{g}}^\ell_{B_+,B_-,k}(\theta_k) \mid \theta_k) + \nabla R_\ell(\theta_k) \cdot \nabla R_\ell(\theta_k)^\top\right\}.$$

Proof. The increment is

$$\Delta_k = -\eta_k \cdot \widehat{\boldsymbol{g}}^\ell_{B_+,B_-,k}(\theta_k).$$

Because the fresh mini-batch used at step k is independent of the past, the conditional law of $\widehat{\boldsymbol{g}}^\ell_{B_+,B_-,k}(\theta_k)$ given the full history depends on the history only through θ_k. Proposition 4 therefore applies conditionally:

$$\mathbb{E}[\widehat{\boldsymbol{g}}^\ell_{B_+,B_-,k}(\theta_k) \mid \theta_k] = \nabla R_\ell(\theta_k).$$

Multiplying by $-\eta_k$ gives the displayed drift identity.

For the second moment,

$$\mathbb{E}[\Delta_k \cdot \Delta_k^\top \mid \theta_k] = \eta_k^2 \cdot \mathbb{E}[\widehat{\boldsymbol{g}}^\ell_{B_+,B_-,k}(\theta_k) \cdot \widehat{\boldsymbol{g}}^\ell_{B_+,B_-,k}(\theta_k)^\top \mid \theta_k].$$

For any square-integrable vector Y,

$$\mathbb{E}[Y \cdot Y^\top] = \mathrm{Cov}(Y) + \mathbb{E}[Y] \cdot \mathbb{E}[Y]^\top.$$

Applying this identity conditionally with

$$Y = \widehat{\boldsymbol{g}}^\ell_{B_+,B_-,k}(\theta_k)$$

and substituting the conditional unbiasedness identity proves the formula. □

Corollary 2 (Mean-ODE generator for decreasing steps)**.** *In Proposition 5, suppose $\eta_k \downarrow 0$. For the one-step kernel*

$$(K_k\varphi)(\theta) = \mathbb{E}[\varphi(\theta - \eta_k \cdot \widehat{\boldsymbol{g}}^\ell_{B_+,B_-}(\theta))],$$

assume that for every compact C in a local open set U,

$$\sup_{\theta\in C}\mathbb{E}[\|\widehat{\boldsymbol{g}}^\ell_{B_+,B_-}(\theta)\|_2^3] < \infty.$$

Then for every $\varphi \in C_b^3(\mathbb{R}^d)$ and every compact $C \subset U$,

$$\sup_{\theta\in C}\left|\frac{(K_k\varphi)(\theta) - \varphi(\theta)}{\eta_k} + \langle\nabla\varphi(\theta), \nabla R_\ell(\theta)\rangle\right| \to 0.$$

Proof. Taylor expand around θ. For $G = \widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)$,

$$\varphi(\theta - \eta_k \cdot G) = \varphi(\theta) - \eta_k \cdot \langle \nabla\varphi(\theta), G\rangle + \frac{\eta_k^2}{2} \cdot G^\top \cdot \nabla^2\varphi(\theta) \cdot G + \mathcal{R}_k(\theta, G),$$

where, uniformly over $\theta \in C$,

$$|\mathcal{R}_k(\theta, G)| \le C_\varphi \cdot \eta_k^3 \cdot \|G\|_2^3$$

because $\varphi \in C_b^3(\mathbb{R}^d)$. Taking expectation and dividing by η_k gives

$$\begin{aligned}\frac{(K_k\varphi)(\theta) - \varphi(\theta)}{\eta_k} &= -\langle \nabla\varphi(\theta), \mathbb{E}[\widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)]\rangle \\ &\quad + \frac{\eta_k}{2} \cdot \mathbb{E}[\widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)^\top \cdot \nabla^2\varphi(\theta) \cdot \widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)] \\ &\quad + O\Big(\eta_k^2 \cdot \mathbb{E}[\|\widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)\|_2^3]\Big).\end{aligned}$$

The first term equals

$$-\langle \nabla\varphi(\theta), \nabla R_\ell(\theta)\rangle$$

by Proposition 4. On the compact set C,

$$\left|\frac{\eta_k}{2} \cdot \mathbb{E}[\widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)^\top \cdot \nabla^2\varphi(\theta) \cdot \widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)]\right| \le \frac{\eta_k}{2} \cdot \|\nabla^2\varphi\|_{\infty,C} \cdot \mathbb{E}[\|\widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)\|_2^2].$$

Since $\|u\|_2^2 \le 1 + \|u\|_2^3$, the uniform third-moment assumption makes this term converge to 0 uniformly on C. The remainder term also converges to 0 uniformly on C. This proves the displayed convergence. □

Corollary 3 (Pairwise AUC SGD Langevin interface)**.** *Consider constant-step direct pairwise AUC SGD*

$$\theta_{k+1} = \theta_k - \eta \cdot \widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-,k}(\theta_k),$$

with fresh pairwise mini-batches at each step. Define

$$\boldsymbol{\Sigma}_{B_+,B_-,\ell}(\theta) = \mathrm{Cov}(\widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta)).$$

For a unit direction $\boldsymbol{v}$, define the scalar kernel

$$\psi_{\theta,\boldsymbol{v}}(x^+, x^-) = \left\langle \ell'(m_\theta(x^+, x^-)) \cdot \big(\nabla f_\theta(x^+) - \nabla f_\theta(x^-)\big), \boldsymbol{v}\right\rangle.$$

Let $\psi^{(1,0)}_{\theta,\boldsymbol{v}}$, $\psi^{(0,1)}_{\theta,\boldsymbol{v}}$, and $\psi^{(1,1)}_{\theta,\boldsymbol{v}}$ denote the positive-class first-order projection, the negative-class first-order projection, and the residual pair-interaction component of this kernel, as defined explicitly in Theorem 3. Under the usual local smoothness and moment assumptions for stochastic-approximation diffusion approximations [68–71], the local Langevin interface for the pairwise dynamics is

$$d\boldsymbol{\Theta}_t = -\nabla R_\ell(\boldsymbol{\Theta}_t)\, dt + (\eta \cdot \boldsymbol{\Sigma}_{B_+,B_-,\ell}(\boldsymbol{\Theta}_t))^{1/2}\, d\boldsymbol{W}_t.$$

Moreover, along every unit direction $\boldsymbol{v}$,

$$\boldsymbol{v}^\top \cdot \boldsymbol{\Sigma}_{B_+,B_-,\ell}(\theta) \cdot \boldsymbol{v} = \frac{\mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+))}{B_+} + \frac{\mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-))}{B_-} + \frac{\mathrm{Var}(\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-))}{B_+B_-}.$$

In the equal-size case $B_+ = B_- = B$, this reduces to the B^{-1} versus B^{-2} split used below.

Proof. Let

$$\Gamma = (X_1^+, \dots, X_{B_+}^+, X_1^-, \dots, X_{B_-}^-)$$

denote one full positive-negative mini-batch, and define the random batch objective

$$F^{\ell}_{B_+,B_-}(\theta, \Gamma) = \frac{1}{B_+B_-} \sum_{i=1}^{B_+} \sum_{j=1}^{B_-} \ell(m_\theta(X_i^+, X_j^-)).$$

Its gradient is exactly the direct pairwise mini-batch gradient because

$$\nabla_\theta \ell(m_\theta(X_i^+, X_j^-)) = \ell'(m_{ij}) \cdot (\boldsymbol{g}_i^+ - \boldsymbol{g}_j^-) = w_\ell(m_{ij}) \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+).$$

Thus

$$\nabla F^{\ell}_{B_+,B_-}(\theta,\Gamma) = \widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta).$$

The population objective associated with $F^{\ell}_{B_+,B_-}$ is

$$\mathbb{E}[F^{\ell}_{B_+,B_-}(\theta,\Gamma)] = R_{\ell}(\theta),$$

because each pair (X_i^+, X_j^-) has the same product distribution as (X^+, X^-). Theorem 2, applied with Markov state $\boldsymbol{y}_k = \theta_k$, gives the local Langevin interface. Equivalently, the first two conditional moments in Proposition 5 identify the local drift as

$$-\nabla R_{\ell}(\theta)$$

and the local diffusion covariance as

$$\eta \cdot \boldsymbol{\Sigma}_{B_+,B_-,\ell}(\theta),$$

under the stated local regularity and moment assumptions.

For the directional coefficient, use

$$\boldsymbol{v}^{\top} \cdot \boldsymbol{\Sigma}_{B_+,B_-,\ell}(\theta) \cdot \boldsymbol{v} = \mathrm{Var}(\langle \widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta), \boldsymbol{v}\rangle).$$

The exact variance split in Corollary 4 gives

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}^{\ell}_{B_+,B_-}(\theta), \boldsymbol{v}\rangle) = \frac{\mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+))}{B_+} + \frac{\mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-))}{B_-} + \frac{\mathrm{Var}(\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+,X^-))}{B_+B_-}.$$

Thus the first-order Hoeffding projections are the directional variance terms with scales B_+^{-1} and B_-^{-1} in the diffusion covariance of the local Langevin interface. When those projections vanish, the directional variance contribution is governed by the canonical term with scale $(B_+B_-)^{-1}$, which is B^{-2} in the equal-size case. □

C.1 Extensions Beyond Binary AUC

Multiclass one-vs-one AUC is a weighted combination of binary AUC comparisons between class pairs, while one-vs-rest AUC combines binary comparisons between each class and the remaining classes. Multilabel macro-AUC is an average of label-wise binary AUC objectives. The corresponding pairwise gradients can therefore be decomposed through the binary comparisons that constitute each metric.

In these settings, the same example may participate in several class-pair or label-wise objectives. All first-order contributions produced by that example must be aggregated before the variance of the first-order Hoeffding term V_1 is computed. The disappearance of a first-order projection for one class pair or one label is not sufficient for the complete multiclass or multilabel first-order term to vanish. Exact B^{-2} scaling requires the aggregated first-order projection to vanish as a whole.

Partial AUC and fixed top-k metrics can likewise be written as restricted or weighted pairwise objectives, with weights selecting pairs in a specified false-positive-rate region or ranking position. If a partial-AUC threshold or top-k boundary is estimated from the current mini-batch, the sampling fluctuation of that threshold or boundary contributes an additional first-order random term. The corresponding degeneracy condition must therefore apply to the complete aggregated gradient, including the effect of boundary estimation. These decompositions provide a unified form for applying our first-order projection and variance-scaling analysis to multiclass, multilabel, and restricted ranking objectives, while making the corresponding aggregate degeneracy conditions explicit.

D Bad Regions and Hoeffding Projection Collapse

This section collects the technical statements used by Section 2. The statements isolate three ingredients: severe misrankings, loss of deterministic corrective motion, and collapse of the first-order stochastic term in the two-sample mini-batch gradient.

The section begins by defining the severe-error functional and its steepest local correction direction; the corrective-cone formulation is recorded only as an appendix-level extension. Proposition 6

through Corollary 5 then prove the U-statistic variance decomposition, and Proposition 8 records the corresponding appendix-level cone extension. Proposition 9 through Corollary 7 give conditional sufficient conditions under which severe-misranking regions make the first-order stochastic term small, increasingly small as severe-misranking mass grows, or exactly degenerate. Proposition 10 through Corollary 8 give elementary constructions and initialization-distribution lower bounds for severe-misranking mass. Proposition 14 through Corollary 9 give finite-sample verification statements. Proposition 17 and Corollary 10 combine high-error mass, weak drift, and projection collapse into the correction-direction bad-region sufficient condition used in the main text.

Severe-misranking mass and bad-region criterion. For $M > 0$, define

$$\pi_M(\theta) = \mathbb{P}(m_\theta(X^+, X^-) \le -M).$$

A region $\mathcal{U}$ is high-error at level (M, ρ) if

$$\inf_{\theta \in \mathcal{U}} \pi_M(\theta) \ge \rho.$$

It is bad for SGD in a specified correction direction $\boldsymbol{v}$ when high-error mass persists while both local correction components are weak: the projected drift $|\langle \nabla R_\ell(\theta), \boldsymbol{v}\rangle|$ is small, and the first-order Hoeffding variance defined below is small relative to the residual pair-interaction variance. The next two paragraphs give the canonical steepest construction of this direction from the severe-error statistic and the corresponding cone version for comparable first-order corrections.

Severe-error functional and steepest correction direction. The directional conditions above are local statements about motion that can reduce the severe-misranking statistic. To make this explicit, fix a nonnegative severe-pair weight $a_M(x)$. For the differentiable statements below, take a_M to be a smooth approximation of $\mathbf{1}\{x \le -M\}$; the hard-mask version is used only in the finite-batch logging rule in Appendix B.5. Let $A_{M,\ell}$ be a differentiable margin-error potential satisfying

$$A'_{M,\ell}(x) = -a_M(x) \cdot w_\ell(x), \qquad w_\ell(x) = -\ell'(x). \tag{4}$$

Define the local severe-error functional

$$\Phi_{M,\ell}(\theta) = \mathbb{E}\big[A_{M,\ell}(m_\theta(X^+, X^-))\big]. \tag{5}$$

When differentiation under the expectation is justified, for example under the same local smoothness and integrability assumptions used for the pairwise gradient in Appendix C, its gradient is

$$\nabla\Phi_{M,\ell}(\theta) = \mathbb{E}\big[a_M(m_\theta(X^+, X^-)) \cdot w_\ell(m_\theta(X^+, X^-)) \cdot \big(\nabla f_\theta(X^-) - \nabla f_\theta(X^+)\big)\big]. \tag{6}$$

When $\nabla\Phi_{M,\ell}(\theta) \ne \mathbf{0}$, the unit direction

$$\boldsymbol{v}_{\mathrm{sev}}(\theta) = -\frac{\nabla\Phi_{M,\ell}(\theta)}{\|\nabla\Phi_{M,\ell}(\theta)\|_2} \tag{7}$$

is the steepest local direction for reducing $\Phi_{M,\ell}$ under the Euclidean parameter norm. Indeed, by Cauchy–Schwarz, for any unit $\boldsymbol{u}$,

$$-\langle \nabla\Phi_{M,\ell}(\theta), \boldsymbol{u}\rangle \le \|\nabla\Phi_{M,\ell}(\theta)\|_2, \tag{8}$$

with equality at $\boldsymbol{u} = \boldsymbol{v}_{\mathrm{sev}}(\theta)$. Thus a mini-batch update $\Delta\theta$ gives first-order improvement of the severe-error functional according to

$$\Phi_{M,\ell}(\theta + \Delta\theta) - \Phi_{M,\ell}(\theta) = \langle \nabla\Phi_{M,\ell}(\theta), \Delta\theta\rangle + o(\|\Delta\theta\|_2). \tag{9}$$

For this mechanism, stochastic motion is measured after projection onto the steepest direction that decreases $\Phi_{M,\ell}$, with $\boldsymbol{v}_{\mathrm{sev}}$ being the canonical direction.

Corrective cone. The projected statements used in the main text can be stated for a fixed correction direction or for a cone of comparable first-order corrections. For $\alpha > 0$, define

$$\mathcal{C}_\alpha(\theta) = \{\boldsymbol{u} : \|\boldsymbol{u}\|_2 = 1,\ \langle \nabla\Phi_{M,\ell}(\theta), \boldsymbol{u}\rangle \le -\alpha\}. \tag{10}$$

Directions in $\mathcal{C}_\alpha(\theta)$ all reduce the severe-error functional at first order by at least α. The single direction $\boldsymbol{v}_{\mathrm{sev}}$ in Eq. (7) gives the steepest correction direction, while the cone formulation records the appendix-level comparison class. When $\|\nabla\Phi_{M,\ell}(\theta)\|_2 \ge \alpha$, this steepest direction belongs to $\mathcal{C}_\alpha(\theta)$.

Proposition 6 (Directional pairwise gradients are two-sample U-statistics)**.** *Fix θ and a unit vector $\boldsymbol{v}$. Define*

$$\psi_{\theta,\boldsymbol{v}}(x^+, x^-) = \left\langle \ell'(m_\theta(x^+, x^-)) \cdot \big(\nabla f_\theta(x^+) - \nabla f_\theta(x^-)\big), \boldsymbol{v} \right\rangle.$$

Then

$$\langle \widehat{\boldsymbol{g}}^{\ell}_B(\theta), \boldsymbol{v}\rangle = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} \psi_{\theta,\boldsymbol{v}}(X_i^+, X_j^-).$$

Proof. Taking the inner product of Eq. (3) with $\boldsymbol{v}$ and using $w_\ell = -\ell'$ gives

$$\langle \widehat{\boldsymbol{g}}^{\ell}_B(\theta), \boldsymbol{v}\rangle = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} \left\langle -\ell'(m_{ij}) \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+), \boldsymbol{v} \right\rangle.$$

This equals

$$\frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} \left\langle \ell'(m_{ij}) \cdot (\boldsymbol{g}_i^+ - \boldsymbol{g}_j^-), \boldsymbol{v} \right\rangle.$$

Substituting $m_{ij} = m_\theta(X_i^+, X_j^-)$, $\boldsymbol{g}_i^+ = \nabla f_\theta(X_i^+)$, and $\boldsymbol{g}_j^- = \nabla f_\theta(X_j^-)$ yields the displayed two-sample U-statistic representation. □

Theorem 3 (Two-sample Hoeffding decomposition and variance split)**.** *Assume the positive batch is i.i.d., the negative batch is i.i.d., the two classes are independent, and*

$$\mathbb{E}[\psi_{\theta,\boldsymbol{v}}(X^+, X^-)^2] < \infty.$$

Let

$$\mu_{\theta,\boldsymbol{v}} = \mathbb{E}[\psi_{\theta,\boldsymbol{v}}(X^+, X^-)],$$

$$\psi^{(1,0)}_{\theta,\boldsymbol{v}}(x^+) = \mathbb{E}[\psi_{\theta,\boldsymbol{v}}(x^+, X^-)] - \mu_{\theta,\boldsymbol{v}}, \qquad \psi^{(0,1)}_{\theta,\boldsymbol{v}}(x^-) = \mathbb{E}[\psi_{\theta,\boldsymbol{v}}(X^+, x^-)] - \mu_{\theta,\boldsymbol{v}},$$

and

$$\psi^{(1,1)}_{\theta,\boldsymbol{v}}(x^+, x^-) = \psi_{\theta,\boldsymbol{v}}(x^+, x^-) - \mathbb{E}[\psi_{\theta,\boldsymbol{v}}(x^+, X^-)] - \mathbb{E}[\psi_{\theta,\boldsymbol{v}}(X^+, x^-)] + \mu_{\theta,\boldsymbol{v}}.$$

Then

$$\begin{aligned}\langle \widehat{\boldsymbol{g}}^{\ell}_B(\theta), \boldsymbol{v}\rangle - \mu_{\theta,\boldsymbol{v}} &= \frac{1}{B}\sum_{i=1}^{B} \psi^{(1,0)}_{\theta,\boldsymbol{v}}(X_i^+) + \frac{1}{B}\sum_{j=1}^{B} \psi^{(0,1)}_{\theta,\boldsymbol{v}}(X_j^-) \\ &\quad + \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} \psi^{(1,1)}_{\theta,\boldsymbol{v}}(X_i^+, X_j^-), \end{aligned} \tag{11}$$

and

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}^{\ell}_B(\theta), \boldsymbol{v}\rangle) = \frac{\mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-))}{B} + \frac{\mathrm{Var}(\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-))}{B^2}. \tag{12}$$

Proof. For every (x^+, x^-),

$$\psi_{\theta,\boldsymbol{v}}(x^+, x^-) - \mu_{\theta,\boldsymbol{v}} = \psi^{(1,0)}_{\theta,\boldsymbol{v}}(x^+) + \psi^{(0,1)}_{\theta,\boldsymbol{v}}(x^-) + \psi^{(1,1)}_{\theta,\boldsymbol{v}}(x^+, x^-),$$

which follows by expanding the definition of $\psi^{(1,1)}_{\theta,\boldsymbol{v}}$. Averaging this identity over the B^2 pairs proves Eq. (11).

The two first-order projections are centered:

$$\mathbb{E}[\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+)] = 0, \qquad \mathbb{E}[\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-)] = 0.$$

The canonical term satisfies

$$\mathbb{E}[\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-) \mid X^+] = 0, \qquad \mathbb{E}[\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-) \mid X^-] = 0,$$

because the positive and negative classes are independent and the conditional expectations subtracted in its definition match the corresponding conditional means.

Let

$$P_B^+ = \frac{1}{B}\sum_{i=1}^{B} \psi_{\theta,\boldsymbol{v}}^{(1,0)}(X_i^+), \quad P_B^- = \frac{1}{B}\sum_{j=1}^{B} \psi_{\theta,\boldsymbol{v}}^{(0,1)}(X_j^-),$$

and

$$H_B = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} \psi_{\theta,\boldsymbol{v}}^{(1,1)}(X_i^+, X_j^-).$$

The decomposition gives

$$\langle \widehat{\boldsymbol{g}}_B^{\ell}(\theta), \boldsymbol{v}\rangle - \mu_{\theta,\boldsymbol{v}} = P_B^+ + P_B^- + H_B.$$

The positive and negative batches are independent, so

$$\mathrm{Cov}(P_B^+, P_B^-) = 0.$$

Conditioning on the whole positive batch gives

$$\mathbb{E}[H_B \mid X_1^+, \ldots, X_B^+] = 0,$$

and therefore $\mathrm{Cov}(P_B^+, H_B) = 0$. Conditioning on the whole negative batch similarly gives $\mathrm{Cov}(P_B^-, H_B) = 0$. Thus

$$\mathrm{Var}(P_B^+ + P_B^- + H_B) = \mathrm{Var}(P_B^+) + \mathrm{Var}(P_B^-) + \mathrm{Var}(H_B).$$

Since the positive samples are i.i.d. and the first-order projection is centered,

$$\mathrm{Var}(P_B^+) = \frac{1}{B}\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,0)}(X^+)), \qquad \mathrm{Var}(P_B^-) = \frac{1}{B}\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(0,1)}(X^-)).$$

For H_B, every off-diagonal covariance is zero. If one index is shared, conditioning on the shared sample gives zero because the canonical term has zero conditional mean. If both indices differ, the two terms are independent and centered. Therefore only the B^2 diagonal terms remain:

$$\mathrm{Var}(H_B) = \frac{B^2}{B^4}\cdot \mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,1)}(X^+, X^-)) = \frac{\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,1)}(X^+, X^-))}{B^2}.$$

Combining the three variance identities proves Eq. (12). The same decomposition is the standard two-sample Hoeffding decomposition [19, 20]. □

Corollary 4 (Unequal positive and negative batch sizes)**.** *Let the positive and negative mini-batch sizes be B_+ and B_-, and define*

$$\widehat{\boldsymbol{g}}_{B_+,B_-}^{\ell}(\theta) = \frac{1}{B_+B_-}\sum_{i=1}^{B_+}\sum_{j=1}^{B_-} w_\ell(m_{ij})\cdot(\boldsymbol{g}_j^- - \boldsymbol{g}_i^+).$$

Under the assumptions of Theorem 3,

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}_{B_+,B_-}^{\ell}(\theta), \boldsymbol{v}\rangle) = \frac{\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,0)}(X^+))}{B_+} + \frac{\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(0,1)}(X^-))}{B_-} + \frac{\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,1)}(X^+, X^-))}{B_+B_-}.$$

Thus unequal class counts change only the constants attached to the two first-order class projections; when B_+ and B_- grow proportionally, the first-order and residual terms retain the same B^{-1} versus B^{-2} scaling used in the main text.

Proof. The Hoeffding decomposition in Eq. (11) becomes

$$\langle \widehat{\boldsymbol{g}}_{B_+,B_-}^{\ell}(\theta), \boldsymbol{v}\rangle - \mu_{\theta,\boldsymbol{v}} = \frac{1}{B_+}\sum_{i=1}^{B_+}\psi_{\theta,\boldsymbol{v}}^{(1,0)}(X_i^+) + \frac{1}{B_-}\sum_{j=1}^{B_-}\psi_{\theta,\boldsymbol{v}}^{(0,1)}(X_j^-) + \frac{1}{B_+B_-}\sum_{i=1}^{B_+}\sum_{j=1}^{B_-}\psi_{\theta,\boldsymbol{v}}^{(1,1)}(X_i^+, X_j^-).$$

The orthogonality arguments in the proof of Theorem 3 are unchanged. The two first-order averages have variances $B_+^{-1}\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,0)}(X^+))$ and $B_-^{-1}\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(0,1)}(X^-))$. For the residual term, only the B_+B_- diagonal pair terms contribute, giving the factor $(B_+B_-)^{-1}$. □

Proposition 7 (Pointwise projected-variance baseline)**.** *Let $Z_1, \dots, Z_B$ be i.i.d. and define*

$$\bar{H}_B(\theta) = \frac{1}{B} \sum_{i=1}^{B} h_\theta(Z_i).$$

For any unit vector $\boldsymbol{v}$ with

$$\mathrm{Var}(\langle h_\theta(Z), \boldsymbol{v} \rangle) > 0,$$

one has

$$\mathrm{Var}(\langle \bar{H}_B(\theta), \boldsymbol{v} \rangle) = \frac{1}{B} \mathrm{Var}(\langle h_\theta(Z), \boldsymbol{v} \rangle).$$

Proof. Set $Y_i = \langle h_\theta(Z_i), \boldsymbol{v} \rangle$. Then $Y_1, \dots, Y_B$ are i.i.d., and

$$\langle \bar{H}_B(\theta), \boldsymbol{v} \rangle = \frac{1}{B} \sum_{i=1}^{B} Y_i.$$

Expanding the variance gives

$$\mathrm{Var}\left(\frac{1}{B} \sum_{i=1}^{B} Y_i \right) = \frac{1}{B^2} \sum_{i=1}^{B} \mathrm{Var}(Y_i) + \frac{1}{B^2} \sum_{i \neq j} \mathrm{Cov}(Y_i, Y_j).$$

The cross-covariances vanish by independence, which gives the stated B^{-1} variance. □

Corollary 5 (First-order versus second-order stochasticity)**.** *If*

$$\mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-)) > 0,$$

then

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}^{\ell}_B(\theta), \boldsymbol{v} \rangle) = \Omega(B^{-1}).$$

If

$$\mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+)) = \mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-)) = 0$$

and

$$\mathrm{Var}(\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-)) > 0,$$

then

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}^{\ell}_B(\theta), \boldsymbol{v} \rangle) = \Theta(B^{-2}).$$

Proof. The conclusion follows directly from Eq. (12). A nonzero coefficient on B^{-1} gives the first claim. If both first-order variances vanish and the residual pair-interaction variance is positive, the variance equals

$$\frac{\mathrm{Var}(\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-))}{B^2}.$$

□

Proposition 8 (Cone-level sufficient bad-region criterion)**.** *Let $\mathcal{U} \subset \mathbb{R}^d$ be a neighborhood, and let $\mathcal{C}_\alpha(\theta)$ be defined by Eq. (10). Suppose that $\mathcal{U}$ is high-error at level (M, ρ) and that, for every $\theta \in \mathcal{U}$ and every $\boldsymbol{u} \in \mathcal{C}_\alpha(\theta)$,*

$$|\langle \nabla R_\ell(\theta), \boldsymbol{u} \rangle| \leq \varepsilon_{\mathrm{drift}},$$

and

$$\mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{u}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{u}}(X^-)) \leq \varepsilon_1.$$

Then every first-order corrective direction in $\mathcal{C}_\alpha(\theta)$ has projected deterministic correction bounded by $\varepsilon_{\mathrm{drift}}$, and its order-$B^{-1}$ Hoeffding contribution is at most ε_1/B. In particular, if $\boldsymbol{v}_{\mathrm{sev}}(\theta) \in \mathcal{C}_\alpha(\theta)$, the same bound applies to the steepest severe-error correction direction.

Proof. Fix $\theta \in \mathcal{U}$ and $\boldsymbol{u} \in \mathcal{C}_\alpha(\theta)$. Since $\boldsymbol{u}$ is fixed after θ is fixed, Proposition 6 gives the scalar two-sample U-statistic representation of $\langle \widehat{\boldsymbol{g}}_B^\ell(\theta), \boldsymbol{u}\rangle$, and Theorem 3 gives

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}_B^\ell(\theta), \boldsymbol{u}\rangle) = \frac{\mathrm{Var}(\psi_{\theta,\boldsymbol{u}}^{(1,0)}(X^+)) + \mathrm{Var}(\psi_{\theta,\boldsymbol{u}}^{(0,1)}(X^-))}{B} + \frac{\mathrm{Var}(\psi_{\theta,\boldsymbol{u}}^{(1,1)}(X^+, X^-))}{B^2}.$$

The order-B^{-1} coefficient is exactly

$$\mathrm{Var}(\psi_{\theta,\boldsymbol{u}}^{(1,0)}(X^+)) + \mathrm{Var}(\psi_{\theta,\boldsymbol{u}}^{(0,1)}(X^-)).$$

By assumption, this coefficient is at most ε_1, so the first-order Hoeffding contribution is at most ε_1/B. The drift assumption gives

$$|\langle \nabla R_\ell(\theta), \boldsymbol{u}\rangle| \le \varepsilon_{\mathrm{drift}}.$$

Because the argument holds for an arbitrary $\theta \in \mathcal{U}$ and an arbitrary $\boldsymbol{u} \in \mathcal{C}_\alpha(\theta)$, the bounds are uniform over the stated cone. Finally, $\boldsymbol{v}_{\mathrm{sev}}(\theta)$ is a unit direction by Eq. (7); if it also satisfies the cone membership condition, substituting $\boldsymbol{u} = \boldsymbol{v}_{\mathrm{sev}}(\theta)$ gives the last claim. □

Proposition 9 (Severe-pair compression makes first-order projections small)**.** *Fix θ, a unit direction $\boldsymbol{v}$, and $M > 0$. Write*

$$S_M^\theta = \{(x^+, x^-) : m_\theta(x^+, x^-) \le -M\}, \qquad \delta_M(\theta) = \mathbb{P}((X^+, X^-) \notin S_M^\theta),$$

and abbreviate $\psi = \psi_{\theta,\boldsymbol{v}}$. Assume

$$|\psi(X^+, X^-)| \le K \qquad \text{a.s.}$$

for some $K < \infty$. Define the severe-truncated row and column means

$$m_{+,M}(x^+) = \mathbb{E}\Big[\psi(x^+, X^-) \cdot \mathbf{1}_{\{(x^+,X^-)\in S_M^\theta\}}\Big],$$

$$m_{-,M}(x^-) = \mathbb{E}\Big[\psi(X^+, x^-) \cdot \mathbf{1}_{\{(X^+,x^-)\in S_M^\theta\}}\Big].$$

If there exist constants a_+ and a_- such that

$$\mathbb{E}[(m_{+,M}(X^+) - a_+)^2] \le \varepsilon_+^2, \qquad \mathbb{E}[(m_{-,M}(X^-) - a_-)^2] \le \varepsilon_-^2,$$

then the coefficient of the order-B^{-1} term in the Hoeffding variance split satisfies

$$\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,0)}(X^+)) + \mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(0,1)}(X^-)) \le 2 \cdot \varepsilon_+^2 + 2 \cdot \varepsilon_-^2 + 4 \cdot K^2 \cdot \delta_M(\theta).$$

Proof. Let

$$m_+(x^+) = \mathbb{E}[\psi(x^+, X^-)], \qquad m_-(x^-) = \mathbb{E}[\psi(X^+, x^-)].$$

By definition,

$$\psi_{\theta,\boldsymbol{v}}^{(1,0)}(X^+) = m_+(X^+) - \mu_{\theta,\boldsymbol{v}}, \qquad \psi_{\theta,\boldsymbol{v}}^{(0,1)}(X^-) = m_-(X^-) - \mu_{\theta,\boldsymbol{v}}.$$

Therefore, for any constant a_+,

$$\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,0)}(X^+)) = \mathrm{Var}(m_+(X^+)) \le \mathbb{E}[(m_+(X^+) - a_+)^2].$$

Decompose

$$m_+(x^+) = m_{+,M}(x^+) + r_+(x^+),$$

where

$$r_+(x^+) = \mathbb{E}\Big[\psi(x^+, X^-) \cdot \mathbf{1}_{\{(x^+,X^-)\notin S_M^\theta\}}\Big].$$

The inequality $(u+v)^2 \le 2 \cdot u^2 + 2 \cdot v^2$ gives

$$\mathbb{E}[(m_+(X^+) - a_+)^2] \le 2 \cdot \mathbb{E}[(m_{+,M}(X^+) - a_+)^2] + 2 \cdot \mathbb{E}[r_+(X^+)^2].$$

Let

$$q_+(x^+) = \mathbb{P}((x^+, X^-) \notin S_M^\theta).$$

Since $|\psi| \le K$,

$$|r_+(x^+)| \le K \cdot q_+(x^+),$$

and hence

$$\mathbb{E}[r_+(X^+)^2] \le K^2 \cdot \mathbb{E}[q_+(X^+)^2] \le K^2 \cdot \mathbb{E}[q_+(X^+)] = K^2 \cdot \delta_M(\theta).$$

Thus

$$\mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+)) \le 2 \cdot \varepsilon_+^2 + 2 \cdot K^2 \cdot \delta_M(\theta).$$

The column projection is identical. Namely, with

$$r_-(x^-) = \mathbb{E}\Big[\psi(X^+, x^-) \cdot \mathbf{1}_{\{(X^+,x^-)\notin S^\theta_M\}}\Big], \qquad q_-(x^-) = \mathbb{P}((X^+, x^-) \notin S^\theta_M),$$

we have

$$\mathbb{E}[q_-(X^-)] = \delta_M(\theta)$$

and therefore

$$\mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-)) \le 2 \cdot \varepsilon_-^2 + 2 \cdot K^2 \cdot \delta_M(\theta).$$

Adding the two bounds proves the claim. □

Corollary 6 (Conditional monotonicity in severe-misranking mass)**.** *Fix $M > 0$, a unit direction $\boldsymbol{v}$, and $K < \infty$. Consider any collection of parameters for which the assumptions of Proposition 9 hold with the same envelope K. For each such θ, let $\varepsilon_+(\theta)$ and $\varepsilon_-(\theta)$ be admissible row and column compression errors in Proposition 9. Then*

$$\mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-)) \le 2 \cdot \varepsilon_+(\theta)^2 + 2 \cdot \varepsilon_-(\theta)^2 + 4 \cdot K^2 \cdot (1 - \pi_M(\theta)).$$

Consequently, if two parameters θ_1, θ_2 have the same admissible compression errors and

$$\pi_M(\theta_1) \ge \pi_M(\theta_2),$$

then the projection-collapse upper bound for θ_1 is no larger than that for θ_2. More generally, along any sequence $(\theta_n)_{n\ge1}$ for which

$$\pi_M(\theta_n) \to 1, \qquad \varepsilon_+(\theta_n) \to 0, \qquad \varepsilon_-(\theta_n) \to 0,$$

the order-B^{-1} coefficient in the Hoeffding variance split converges to 0. If, in addition,

$$\liminf_{n\to\infty} \mathrm{Var}(\psi^{(1,1)}_{\theta_n,\boldsymbol{v}}(X^+, X^-)) > 0,$$

then the leading single-class stochastic term disappears along the sequence while the canonical pair-interaction term remains nontrivial.

Proof. The displayed bound is Proposition 9 with

$$\delta_M(\theta) = \mathbb{P}((X^+, X^-) \notin S^\theta_M) = 1 - \pi_M(\theta).$$

If the admissible compression errors are the same for θ_1 and θ_2, the only term in the bound that depends on the severe-misranking mass is

$$4 \cdot K^2 \cdot (1 - \pi_M(\theta)).$$

This term is nonincreasing in $\pi_M(\theta)$, proving the monotonic comparison. For the sequence statement, the right-hand side satisfies

$$2 \cdot \varepsilon_+(\theta_n)^2 + 2 \cdot \varepsilon_-(\theta_n)^2 + 4 \cdot K^2 \cdot (1 - \pi_M(\theta_n)) \to 0.$$

Therefore

$$\mathrm{Var}(\psi^{(1,0)}_{\theta_n,\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\theta_n,\boldsymbol{v}}(X^-)) \to 0.$$

The final claim follows by combining this convergence with the assumed positive lower limit of the canonical variance in the variance split of Theorem 3. □

Corollary 7 (Exact collapse in a fully severe compressed region)**.** *In the setting of Proposition 9, suppose*

$$\mathbb{P}((X^+, X^-) \in S_M^\theta) = 1$$

and the row and column conditional means are constant:

$$\mathbb{E}[\psi_{\theta,\boldsymbol{v}}(X^+, X^-) \mid X^+] = a_+ \quad a.s., \qquad \mathbb{E}[\psi_{\theta,\boldsymbol{v}}(X^+, X^-) \mid X^-] = a_- \quad a.s.$$

Then $a_+ = a_- = \mu_{\theta,\boldsymbol{v}}$, both first-order Hoeffding projections vanish, and

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}_B^\ell(\theta), \boldsymbol{v}\rangle) = \frac{\mathrm{Var}(\psi_{\theta,\boldsymbol{v}}^{(1,1)}(X^+, X^-))}{B^2}.$$

Proof. Taking expectations in the row-mean identity gives

$$a_+ = \mathbb{E}[\mathbb{E}[\psi_{\theta,\boldsymbol{v}}(X^+, X^-) \mid X^+]] = \mu_{\theta,\boldsymbol{v}}.$$

Taking expectations in the column-mean identity gives $a_- = \mu_{\theta,\boldsymbol{v}}$. Hence

$$\psi_{\theta,\boldsymbol{v}}^{(1,0)}(X^+) = \mathbb{E}[\psi_{\theta,\boldsymbol{v}}(X^+, X^-) \mid X^+] - \mu_{\theta,\boldsymbol{v}} = 0 \quad \text{a.s.},$$

and similarly

$$\psi_{\theta,\boldsymbol{v}}^{(0,1)}(X^-) = 0 \quad \text{a.s.}$$

The variance identity follows from Theorem 3. □

Proposition 9 is deliberately conditional. A large value of $\pi_M(\theta)$ alone does not force projection collapse. It makes collapse relevant when the severe pairs that dominate the local pair distribution have nearly constant row and column effects in the projected update. Corollary 6 states the precise monotonic consequence: under such compression, increasing severe-misranking mass tightens the upper bound on the first-order projections. Corollary 7 records the corresponding exact case.

Proposition 10 (Elementary severe-misranking and projection-collapse constructions)**.** *The two ingredients used in the bad-region mechanism are nonempty.*

1. *Let the sample space contain one positive point x^+ and one negative point x^-, and define*

   $$f_\theta(x^+) = -\theta, \qquad f_\theta(x^-) = \theta.$$

   Then $m_\theta(x^+, x^-) = -2 \cdot \theta$, and for every $M > 0$, the ray $\{\theta \geq M/2\}$ satisfies $\pi_M(\theta) = 1$.

2. *Let A and C be independent Rademacher random variables, let $f_\theta(z) = \theta \cdot z$, and use $\ell(x) = x^2$. For $\boldsymbol{v} = 1$, the scalar pairwise-gradient kernel is*

   $$G_\theta(A, C) = 2 \cdot \theta \cdot (A - C)^2 = 4 \cdot \theta - 4 \cdot \theta \cdot A \cdot C.$$

   Both first-order Hoeffding projections vanish, and

   $$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}_B^\ell(\theta), \boldsymbol{v}\rangle) = \frac{16 \cdot \theta^2}{B^2}.$$

Proof. The first claim follows from

$$m_\theta(x^+, x^-) = (-\theta) - \theta = -2 \cdot \theta.$$

If $\theta \geq M/2$, then $m_\theta(x^+, x^-) \leq -M$, so the severe-misranking event has probability 1.

For the second claim,

$$m_\theta(A, C) = \theta \cdot (A - C).$$

Since $w(x) = -\ell'(x) = -2 \cdot x$, the scalar gradient kernel in direction 1 is

$$G_\theta(A, C) = w(m_\theta(A, C)) \cdot (C - A) = 2 \cdot \theta \cdot (A - C)^2.$$

Using $A^2 = C^2 = 1$ gives

$$(A - C)^2 = 2 - 2 \cdot A \cdot C,$$

and hence

$$G_\theta(A, C) = 4 \cdot \theta - 4 \cdot \theta \cdot A \cdot C.$$

Its mean is $4 \cdot \theta$. Conditioning on $A = a$ gives

$$\mathbb{E}[G_\theta(a, C)] - 4 \cdot \theta = -4 \cdot \theta \cdot a \cdot \mathbb{E}[C] = 0,$$

and conditioning on $C = c$ gives the same result. Therefore both first-order projections vanish. The centered canonical term is

$$-4 \cdot \theta \cdot A \cdot C,$$

whose variance is $16 \cdot \theta^2$. Applying Eq. (12) gives the displayed B^{-2} variance. □

Proposition 11 (High-error starts under random or label-blind initialization)**.** *High-error starts can be phrased as probability statements.*

1. *In the severe-misranking ray of Proposition 10, if $\Theta_0 \sim N(0, \sigma^2)$, then*

$$\mathbb{P}(\Theta_0 \geq M/2) = 1 - \Phi\left(\frac{M}{2 \cdot \sigma}\right).$$

2. *Let Θ_0 be a random initialization. Conditional on Θ_0, define*

$$Y^+ = f_{\Theta_0}(X^+), \qquad Y^- = f_{\Theta_0}(X^-).$$

If Y^+ and Y^- are conditionally independent and identically distributed, then

$$\pi_0(\Theta_0) \geq \frac{1}{2} \qquad \text{a.s.}$$

If additionally

$$Y^+, Y^- \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \quad \text{conditional on } \Theta_0,$$

then for every $M \geq 0$,

$$\pi_M(\Theta_0) = 1 - \Phi\left(\frac{M}{\sqrt{2} \cdot \sigma}\right).$$

Proof. The Gaussian ray formula follows by standardization. For the label-blind statement, conditional on Θ_0, the distribution of $Y^+ - Y^-$ is symmetric about 0 because (Y^+, Y^-) and (Y^-, Y^+) have the same conditional law. Hence

$$\mathbb{P}(Y^+ - Y^- \leq 0 \mid \Theta_0) = \frac{1 + \mathbb{P}(Y^+ - Y^- = 0 \mid \Theta_0)}{2} \geq \frac{1}{2}.$$

Under the conditional Gaussian assumption,

$$Y^+ - Y^- \mid \Theta_0 \sim N(0, 2 \cdot \sigma^2),$$

which gives the displayed formula for $\pi_M(\Theta_0)$. □

Proposition 12 (Random isotropic linear heads give severe-misranking lower bounds)**.** *Let $h : \mathcal{X} \to \mathbb{R}^d$ be measurable and consider*

$$f_{\boldsymbol{w}}(x) = \boldsymbol{w}^\top h(x).$$

Let

$$\boldsymbol{W}_0 \sim N(0, \alpha^2 \cdot \boldsymbol{I}_d), \qquad \alpha > 0,$$

independently of (X^+, X^-). Define

$$\boldsymbol{D} = h(X^+) - h(X^-), \qquad \pi_M(\boldsymbol{w}) = \mathbb{P}(\boldsymbol{w}^\top \boldsymbol{D} \leq -M).$$

Then

$$\mathbb{E}[\pi_0(\boldsymbol{W}_0)] = \frac{1}{2} + \frac{1}{2} \cdot \mathbb{P}(\boldsymbol{D} = \boldsymbol{0}) \geq \frac{1}{2}.$$

For $M > 0$,

$$\mathbb{E}[\pi_M(\boldsymbol{W}_0)] = \mathbb{E}[H_M(\|\boldsymbol{D}\|_2)],$$

where

$$H_M(r) = \begin{cases} 1 - \Phi\left(\dfrac{M}{\alpha \cdot r}\right), & r > 0, \\ 0, & r = 0. \end{cases}$$

If $R > 0$ and $\mathbb{P}(\|\boldsymbol{D}\|_2 \geq R) \geq q$, then

$$\mathbb{E}[\pi_M(\boldsymbol{W}_0)] \geq q \cdot \left\{1 - \Phi\left(\frac{M}{\alpha \cdot R}\right)\right\}.$$

Proof. Condition on $\boldsymbol{D} = \boldsymbol{d}$. If $\boldsymbol{d} \neq \boldsymbol{0}$, then

$$\boldsymbol{W}_0^\top \boldsymbol{d} \sim N(0, \alpha^2 \cdot \|\boldsymbol{d}\|_2^2).$$

If $\boldsymbol{d} = \boldsymbol{0}$, then $\boldsymbol{W}_0^\top \boldsymbol{d} = 0$ almost surely. This proves the $M = 0$ identity and the exact H_M formula. If $R > 0$ and $\|\boldsymbol{D}\|_2 \geq R$, then

$$\frac{M}{\alpha \cdot \|\boldsymbol{D}\|_2} \leq \frac{M}{\alpha \cdot R},$$

and $1 - \Phi(\cdot)$ is nonincreasing. Taking expectation over this event gives the lower bound. □

Lemma 1 (One-sided second-moment lower bound)**.** *Let $X \geq 0$, $0 < \mathbb{E}[X] < \infty$, and $\mathbb{E}[X^2] < \infty$. For every $\lambda \in [0, 1)$,*

$$\mathbb{P}(X \geq \lambda \cdot \mathbb{E}[X]) \geq \frac{(1-\lambda)^2 \cdot \mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

Proof. Let $A_\lambda = \{X \geq \lambda \cdot \mathbb{E}[X]\}$. On A_λ^c, $X < \lambda \cdot \mathbb{E}[X]$, so

$$\mathbb{E}[X \cdot \mathbf{1}_{A_\lambda}] = \mathbb{E}[X] - \mathbb{E}[X \cdot \mathbf{1}_{A_\lambda^c}] \geq (1-\lambda) \cdot \mathbb{E}[X].$$

By Cauchy-Schwarz,

$$\mathbb{E}[X \cdot \mathbf{1}_{A_\lambda}]^2 \leq \mathbb{E}[X^2] \cdot \mathbb{P}(A_\lambda).$$

Combining the two inequalities proves the claim. □

Proposition 13 (High-dimensional scaled random heads retain nontrivial bad mass)**.** *For each $d \geq 1$, let $h_d : \mathcal{X} \to \mathbb{R}^d$ and*

$$\boldsymbol{d}_d = h_d(X^+) - h_d(X^-).$$

Let

$$\boldsymbol{w}_{0,d} \sim N\left(0, \frac{\sigma^2}{d} \cdot \boldsymbol{I}_d\right), \qquad \sigma > 0,$$

independently of (X^+, X^-). Assume there exist constants $a, b > 0$, independent of d, such that

$$\mathbb{E}[\|\boldsymbol{d}_d\|_2^2] \geq a^2 \cdot d, \qquad \mathbb{E}[\|\boldsymbol{d}_d\|_2^4] \leq b^4 \cdot d^2.$$

Then for every $M > 0$,

$$\mathbb{E}[\pi_{M,d}(\boldsymbol{w}_{0,d})] \geq \frac{a^4}{4 \cdot b^4} \left\{1 - \Phi\left(\frac{\sqrt{2} \cdot M}{\sigma \cdot a}\right)\right\},$$

where $\pi_{M,d}(\boldsymbol{w}) = \mathbb{P}(\boldsymbol{w}^\top \boldsymbol{d}_d \leq -M)$.

Proof. Let $X_d = \|\boldsymbol{d}_d\|_2^2$. Lemma 1 with $\lambda = 1/2$ gives

$$\mathbb{P}\left(X_d \geq \frac{1}{2} \cdot \mathbb{E}[X_d]\right) \geq \frac{\mathbb{E}[X_d]^2}{4 \cdot \mathbb{E}[X_d^2]} \geq \frac{a^4}{4 \cdot b^4}.$$

Since $\mathbb{E}[X_d] \geq a^2 \cdot d$, this event implies

$$\|\boldsymbol{d}_d\|_2 \geq a \cdot \sqrt{d/2}.$$

Apply Proposition 12 with

$$\alpha = \frac{\sigma}{\sqrt{d}}, \qquad R = a \cdot \sqrt{d/2}, \qquad q = \frac{a^4}{4 \cdot b^4}.$$

Because $\alpha \cdot R = \sigma \cdot a/\sqrt{2}$, the result follows. □

Corollary 8 (Average severe-misranking mass gives a nontrivial high-error initialization probability)**.** *Let W_0 be any random initialization and $Y_M = \pi_M(W_0) \in [0, 1]$. If $\mathbb{E}[Y_M] \geq \mu_M$ for $\mu_M \in (0, 1]$, then for every $\tau \in [0, \mu_M)$,*

$$\mathbb{P}(Y_M \geq \tau) \geq \frac{\mu_M - \tau}{1 - \tau}.$$

In particular,

$$\mathbb{P}\left(Y_M \geq \frac{\mu_M}{2}\right) \geq \frac{\mu_M}{2 - \mu_M} \geq \frac{\mu_M}{2}.$$

Proof. Since $0 \le Y_M \le 1$,

$$Y_M \le \tau + (1-\tau)\cdot \mathbf{1}_{\{Y_M \ge \tau\}}.$$

Taking expectations and using $\mathbb{E}[Y_M] \ge \mu_M$ gives the first inequality. The final display follows by taking $\tau = \mu_M/2$. □

Proposition 14 (Finite-pair high-error check)**.** *Fix θ, $M \ge 0$, and*

$$\pi_M(\theta) = \mathbb{P}(m_\theta(X^+, X^-) \le -M).$$

Let $(\widetilde{X}_k^+, \widetilde{X}_k^-)_{k=1}^K$ be independent positive-negative pairs and define

$$\widehat{\pi}_{M,K}(\theta) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{\{m_\theta(\widetilde{X}_k^+, \widetilde{X}_k^-) \le -M\}}.$$

Then

$$\mathbb{E}[\widehat{\pi}_{M,K}(\theta)] = \pi_M(\theta), \qquad \mathrm{Var}(\widehat{\pi}_{M,K}(\theta)) \le \frac{1}{4\cdot K}.$$

Consequently, for every $\varepsilon > 0$,

$$\mathbb{P}(|\widehat{\pi}_{M,K}(\theta) - \pi_M(\theta)| \ge \varepsilon) \le \frac{1}{4\cdot K\cdot \varepsilon^2}.$$

Equivalently,

$$\mathbb{P}(\pi_M(\theta) \ge \widehat{\pi}_{M,K}(\theta) - \varepsilon) \ge 1 - \frac{1}{4\cdot K\cdot \varepsilon^2},$$

and

$$\mathbb{P}(\pi_M(\theta) \le \widehat{\pi}_{M,K}(\theta) + \varepsilon) \ge 1 - \frac{1}{4\cdot K\cdot \varepsilon^2}.$$

Proof. The indicators in the empirical average are i.i.d. Bernoulli random variables with mean $\pi_M(\theta)$. Their average has expectation $\pi_M(\theta)$ and variance

$$\frac{\pi_M(\theta)\cdot(1-\pi_M(\theta))}{K} \le \frac{1}{4\cdot K}.$$

Chebyshev's inequality gives the deviation bound. The one-sided lower and upper statements follow because the events

$$\{\widehat{\pi}_{M,K}(\theta) - \pi_M(\theta) \ge \varepsilon\}, \qquad \{\pi_M(\theta) - \widehat{\pi}_{M,K}(\theta) \ge \varepsilon\}$$

are both contained in

$$\{|\widehat{\pi}_{M,K}(\theta) - \pi_M(\theta)| \ge \varepsilon\}.$$

□

Proposition 15 (Finite-initialization lower bound for high-error starts)**.** *Fix $M \ge 0$ and a seed-level threshold $\tau \in [0,1]$. Let*

$$\theta^{(1)}, \dots, \theta^{(S)}$$

be S parameter points. For each seed s, construct an independent pairwise estimate

$$\widehat{\pi}^{(s)}_{M,K_s}(\theta^{(s)})$$

as in Proposition 14, choose $\varepsilon_s > 0$, and define

$$C_s = \mathbf{1}_{\{\widehat{\pi}^{(s)}_{M,K_s}(\theta^{(s)}) - \varepsilon_s \ge \tau\}}.$$

Then

$$\mathbb{P}\left(\frac{1}{S}\sum_{s=1}^{S} \mathbf{1}_{\{\pi_M(\theta^{(s)}) \ge \tau\}} \ge \frac{1}{S}\sum_{s=1}^{S} C_s\right) \ge 1 - \sum_{s=1}^{S} \frac{1}{4\cdot K_s\cdot \varepsilon_s^2}.$$

Proof. For each s, define

$$E_s = \left\{ \pi_M(\theta^{(s)}) \geq \widehat{\pi}^{(s)}_{M,K_s}(\theta^{(s)}) - \varepsilon_s \right\}.$$

By Proposition 14,

$$\mathbb{P}(E_s) \geq 1 - \frac{1}{4 \cdot K_s \cdot \varepsilon_s^2}.$$

The union bound gives

$$\mathbb{P}\left(\bigcap_{s=1}^{S} E_s \right) \geq 1 - \sum_{s=1}^{S} \frac{1}{4 \cdot K_s \cdot \varepsilon_s^2}.$$

On $\bigcap_{s=1}^{S} E_s$, if $C_s = 1$, then

$$\widehat{\pi}^{(s)}_{M,K_s}(\theta^{(s)}) - \varepsilon_s \geq \tau$$

and therefore

$$\pi_M(\theta^{(s)}) \geq \tau.$$

Thus

$$C_s \leq \mathbf{1}_{\{\pi_M(\theta^{(s)}) \geq \tau\}}$$

for every s. Averaging over $s = 1, \ldots, S$ proves the claim. □

Corollary 9 (Initialization-distribution lower bound for high-error starts)**.** *Let W_0 be a random initialization and define*

$$p_{M,\tau} = \mathbb{P}(\pi_M(W_0) \geq \tau).$$

Let $W_0^{(1)}, \ldots, W_0^{(S)}$ be independent draws from the same initialization distribution. For each s, build C_s as in Proposition 15, using pairwise samples independent of the initializations. Then for every $\rho > 0$,

$$\mathbb{P}\left(p_{M,\tau} \geq \frac{1}{S} \sum_{s=1}^{S} C_s - \rho \right) \geq 1 - \sum_{s=1}^{S} \frac{1}{4 \cdot K_s \cdot \varepsilon_s^2} - \frac{1}{4 \cdot S \cdot \rho^2}.$$

Proof. Define the true seed indicators

$$B_s = \mathbf{1}_{\{\pi_M(W_0^{(s)}) \geq \tau\}}, \qquad s = 1, \ldots, S.$$

Then $B_1, \ldots, B_S$ are i.i.d. Bernoulli random variables with mean $p_{M,\tau}$, and

$$\mathrm{Var}\left(\frac{1}{S} \sum_{s=1}^{S} B_s \right) = \frac{p_{M,\tau} \cdot (1 - p_{M,\tau})}{S} \leq \frac{1}{4 \cdot S}.$$

Chebyshev's inequality gives

$$\mathbb{P}\left(p_{M,\tau} \geq \frac{1}{S} \sum_{s=1}^{S} B_s - \rho \right) \geq 1 - \frac{1}{4 \cdot S \cdot \rho^2}.$$

By Proposition 15,

$$\mathbb{P}\left(\frac{1}{S} \sum_{s=1}^{S} B_s \geq \frac{1}{S} \sum_{s=1}^{S} C_s \right) \geq 1 - \sum_{s=1}^{S} \frac{1}{4 \cdot K_s \cdot \varepsilon_s^2}.$$

On the intersection of these two events,

$$p_{M,\tau} \geq \frac{1}{S} \sum_{s=1}^{S} B_s - \rho \geq \frac{1}{S} \sum_{s=1}^{S} C_s - \rho.$$

Another union bound proves the result. □

Proposition 16 (Severe-misranking neighborhoods persist under local score stability)**.** *Fix $M > 0$ and $\delta \in (0, 1)$. Suppose there exist $\bar{\theta}$, $\rho > 0$, and $L_{\mathrm{sc}} > 0$ such that*

$$\mathbb{P}(m_{\bar{\theta}}(X^+, X^-) \le -(M + \rho)) \ge 1 - \delta$$

and

$$|m_\theta(x^+, x^-) - m_{\bar{\theta}}(x^+, x^-)| \le L_{\mathrm{sc}} \cdot \|\theta - \bar{\theta}\|_2$$

for all relevant pairs. Then every θ with

$$\|\theta - \bar{\theta}\|_2 \le \rho / L_{\mathrm{sc}}$$

satisfies

$$\pi_M(\theta) \ge 1 - \delta.$$

Proof. On the event $m_{\bar{\theta}}(X^+, X^-) \le -(M + \rho)$, the score-stability bound gives

$$m_\theta(X^+, X^-) \le -(M + \rho) + L_{\mathrm{sc}} \cdot \|\theta - \bar{\theta}\|_2 \le -M.$$

Thus the severe event at $\bar{\theta}$ is contained in the severe event at θ. Taking probabilities proves the claim. □

Proposition 17 (Single-coordinate degenerate high-error near-stationary regions are bad)**.** *Let $\mathcal{U} \subset \mathbb{R}^d$ be a neighborhood and fix $M, \rho > 0$, a unit direction $\boldsymbol{v}$, and constants $\varepsilon_{\mathrm{drift}}, \varepsilon_1, c_2 \ge 0$ with $c_2 > 0$. Assume*

$$\inf_{\theta \in \mathcal{U}} \pi_M(\theta) \ge \rho, \qquad \sup_{\theta \in \mathcal{U}} |\langle \nabla R_\ell(\theta), \boldsymbol{v} \rangle| \le \varepsilon_{\mathrm{drift}},$$

$$\sup_{\theta \in \mathcal{U}} \left\{ \mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-)) \right\} \le \varepsilon_1, \qquad \inf_{\theta \in \mathcal{U}} \mathrm{Var}(\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-)) \ge c_2.$$

Then $\mathcal{U}$ is bad for SGD in the specified correction direction $\boldsymbol{v}$: severe misrankings persist, deterministic correction along $\boldsymbol{v}$ is bounded by $\varepsilon_{\mathrm{drift}}$, and the first-order Hoeffding contribution to the projected variance is at most ε_1 / B. This fixed-direction statement is the form used in the projected escape-time argument. When the steepest direction $\boldsymbol{v}_{\mathrm{sev}}(\theta)$ varies with θ, Proposition 8 gives the corresponding cone-level uniform version. If $\varepsilon_1 = 0$, then

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}^\ell_B(\theta), \boldsymbol{v} \rangle) = \frac{\mathrm{Var}(\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-))}{B^2} \qquad \text{for all } \theta \in \mathcal{U}.$$

Proof. The first assumption says exactly that $\mathcal{U}$ is high-error at level (M, ρ). The second assumption gives the stated projected-drift bound. The variance split in Theorem 3 gives, for each $\theta \in \mathcal{U}$,

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}^\ell_B(\theta), \boldsymbol{v} \rangle) = \frac{\mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-))}{B} + \frac{\mathrm{Var}(\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-))}{B^2}.$$

The first-order contribution is therefore at most ε_1 / B, while the residual pair-interaction component remains nontrivial because its variance is at least c_2. If $\varepsilon_1 = 0$, the B^{-1} term vanishes identically and the displayed B^{-2} identity follows. □

Corollary 10 (Degenerate stationary neighborhoods are bad in a correction direction)**.** *Let $\theta_\star$ be a stationary point of R_ℓ, let $\mathcal{U} = B(\theta_\star, r)$, and fix $M, \rho > 0$, a unit direction $\boldsymbol{v}$, and constants $\varepsilon_1, c_2 \ge 0$ with $c_2 > 0$. Assume ∇R_ℓ is L_R-Lipschitz on $\mathcal{U}$. If*

$$\inf_{\theta \in \mathcal{U}} \pi_M(\theta) \ge \rho,$$

$$\sup_{\theta \in \mathcal{U}} \left\{ \mathrm{Var}(\psi^{(1,0)}_{\theta,\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\theta,\boldsymbol{v}}(X^-)) \right\} \le \varepsilon_1, \qquad \inf_{\theta \in \mathcal{U}} \mathrm{Var}(\psi^{(1,1)}_{\theta,\boldsymbol{v}}(X^+, X^-)) \ge c_2 > 0,$$

then $\mathcal{U}$ is bad for SGD in the sense of Proposition 17, with

$$|\langle \nabla R_\ell(\theta), \boldsymbol{v} \rangle| \le L_R \cdot r \qquad \text{for every } \theta \in \mathcal{U}.$$

Proof. Since $\theta_\star$ is stationary, $\nabla R_\ell(\theta_\star) = \mathbf{0}$. For every $\theta \in B(\theta_\star, r)$,

$$\|\nabla R_\ell(\theta)\|_2 = \|\nabla R_\ell(\theta) - \nabla R_\ell(\theta_\star)\|_2 \le L_R \cdot \|\theta - \theta_\star\|_2 \le L_R \cdot r.$$

Hence

$$|\langle \nabla R_\ell(\theta), \boldsymbol{v} \rangle| \le L_R \cdot r.$$

Apply Proposition 17 with $\varepsilon_{\mathrm{drift}} = L_R \cdot r$. □

E Surrogate Weight Calculations

This section records the surrogate-weight calculations used by Section 3. The comparison is local and weight-based. It asks how the pairwise weight

$$w_\ell(x) = -\ell'(x)$$

changes as a positive-negative margin x moves from severe violation to easy separation. The five canonical profiles studied in the main text are logistic, exponential, square, hinge, and squared-hinge.

E.1 Canonical Weights

Proposition 18 (Canonical pairwise weights)**.** *For*

$$\ell_{\log}(x) = \log(1 + e^{-x}), \qquad \ell_{\exp}(x) = e^{-x}, \qquad \ell_{\mathrm{sq},m}(x) = (m - x)^2,$$

$$\ell_{\mathrm{h},m}(x) = (m - x)_+, \qquad \ell_{\mathrm{sqh},m}(x) = (m - x)_+^2,$$

the corresponding weights are

$$w_{\log}(x) = \frac{1}{1 + e^x}, \qquad w_{\exp}(x) = e^{-x}, \qquad w_{\mathrm{sq},m}(x) = 2 \cdot (m - x),$$

$$w_{\mathrm{h},m}(x) = \mathbf{1}_{\{x<m\}}, \qquad w_{\mathrm{sqh},m}(x) = 2 \cdot (m - x)_+.$$

For hinge, the boundary subgradient at $x = m$ is chosen to be zero.

Proof. For logistic,

$$\ell'_{\log}(x) = \frac{-e^{-x}}{1 + e^{-x}} = -\frac{1}{1 + e^x},$$

and hence $w_{\log}(x) = 1/(1 + e^x)$. For exponential,

$$\ell'_{\exp}(x) = -e^{-x},$$

so $w_{\exp}(x) = e^{-x}$. For square,

$$\ell'_{\mathrm{sq},m}(x) = 2 \cdot (m - x) \cdot (-1) = -2 \cdot (m - x),$$

so $w_{\mathrm{sq},m}(x) = 2 \cdot (m - x)$.

For hinge, $(m - x)_+$ has derivative -1 on $x < m$ and derivative 0 on $x > m$. With the stated boundary convention, $w_{\mathrm{h},m}(x) = \mathbf{1}_{\{x<m\}}$. For squared-hinge, the derivative is $-2 \cdot (m - x)$ on $x < m$ and 0 on $x \geq m$, which gives $w_{\mathrm{sqh},m}(x) = 2 \cdot (m - x)_+$. □

E.2 Active-Side Correction Versus Class-Mean Collapse

The next statements describe what happens on a batch where all pairs are active, $m_{ij} < m$. Define

$$\bar{s}_+ = \frac{1}{B}\sum_{i=1}^{B} f_\theta(x_i^+), \qquad \bar{s}_- = \frac{1}{B}\sum_{j=1}^{B} f_\theta(x_j^-),$$

$$\bar{\boldsymbol{g}}_+ = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{g}_i^+, \qquad \bar{\boldsymbol{g}}_- = \frac{1}{B}\sum_{j=1}^{B} \boldsymbol{g}_j^-, \qquad \boldsymbol{d}_B = \bar{\boldsymbol{g}}_- - \bar{\boldsymbol{g}}_+,$$

and the centered score-gradient terms

$$\boldsymbol{c}_+ = \frac{1}{B}\sum_{i=1}^{B} \big(f_\theta(x_i^+) - \bar{s}_+\big) \cdot \boldsymbol{g}_i^+, \qquad \boldsymbol{c}_- = \frac{1}{B}\sum_{j=1}^{B} \big(f_\theta(x_j^-) - \bar{s}_-\big) \cdot \boldsymbol{g}_j^-,$$

$$\boldsymbol{c}_B = \boldsymbol{c}_+ + \boldsymbol{c}_-.$$

The vector $\boldsymbol{d}_B$ is the class-mean gradient direction, while $\boldsymbol{c}_B$ records within-class score-gradient variation.

Proposition 19 (Hinge exact decoupling on the active region)**.** *If all current pairs satisfy $m_{ij} < m$, then the hinge mini-batch gradient satisfies*

$$\widehat{\boldsymbol{g}}_B^{\mathrm{h},m}(\theta) = \bar{\boldsymbol{g}}_- - \bar{\boldsymbol{g}}_+, \qquad \bar{\boldsymbol{g}}_+ = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{g}_i^+, \quad \bar{\boldsymbol{g}}_- = \frac{1}{B}\sum_{j=1}^{B} \boldsymbol{g}_j^-.$$

Proof. If $m_{ij} < m$ for every pair, then $w_{\mathrm{h},m}(m_{ij}) = 1$. Substituting this into Eq. (3) gives

$$\widehat{\boldsymbol{g}}_B^{\mathrm{h},m}(\theta) = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}(\boldsymbol{g}_j^- - \boldsymbol{g}_i^+) = \bar{\boldsymbol{g}}_- - \bar{\boldsymbol{g}}_+.$$

□

Lemma 2 (Affine active-side decomposition)**.** *Suppose $w_\ell(x) = a \cdot x + b$ on the current batch. Then*

$$\widehat{\boldsymbol{g}}_B^{\ell}(\theta) = \big(a \cdot (\bar{s}_+ - \bar{s}_-) + b\big) \cdot \boldsymbol{d}_B - a \cdot \boldsymbol{c}_B.$$

Proof. Split the weight into constant and affine parts. The constant part gives

$$\frac{b}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}(\boldsymbol{g}_j^- - \boldsymbol{g}_i^+) = b \cdot \boldsymbol{d}_B.$$

For the affine part, using $m_{ij} = f_\theta(x_i^+) - f_\theta(x_j^-)$,

$$\begin{aligned}
\frac{1}{B^2}\sum_{i,j} m_{ij} \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+) &= \frac{1}{B^2}\sum_{i,j} f_\theta(x_i^+) \cdot \boldsymbol{g}_j^- - \frac{1}{B^2}\sum_{i,j} f_\theta(x_i^+) \cdot \boldsymbol{g}_i^+ \\
&\quad - \frac{1}{B^2}\sum_{i,j} f_\theta(x_j^-) \cdot \boldsymbol{g}_j^- + \frac{1}{B^2}\sum_{i,j} f_\theta(x_j^-) \cdot \boldsymbol{g}_i^+ \\
&= \bar{s}_+ \cdot \bar{\boldsymbol{g}}_- - \frac{1}{B}\sum_{i} f_\theta(x_i^+) \cdot \boldsymbol{g}_i^+ - \frac{1}{B}\sum_{j} f_\theta(x_j^-) \cdot \boldsymbol{g}_j^- \\
&\quad + \bar{s}_- \cdot \bar{\boldsymbol{g}}_+.
\end{aligned}$$

By the definitions of $\boldsymbol{c}_+$ and $\boldsymbol{c}_-$,

$$\frac{1}{B}\sum_{i} f_\theta(x_i^+) \cdot \boldsymbol{g}_i^+ = \boldsymbol{c}_+ + \bar{s}_+ \cdot \bar{\boldsymbol{g}}_+, \qquad \frac{1}{B}\sum_{j} f_\theta(x_j^-) \cdot \boldsymbol{g}_j^- = \boldsymbol{c}_- + \bar{s}_- \cdot \bar{\boldsymbol{g}}_-.$$

Substitution gives

$$\frac{1}{B^2}\sum_{i,j} m_{ij} \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+) = (\bar{s}_+ - \bar{s}_-) \cdot (\bar{\boldsymbol{g}}_- - \bar{\boldsymbol{g}}_+) - (\boldsymbol{c}_+ + \boldsymbol{c}_-),$$

or equivalently

$$\frac{1}{B^2}\sum_{i,j} m_{ij} \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+) = (\bar{s}_+ - \bar{s}_-) \cdot \boldsymbol{d}_B - \boldsymbol{c}_B.$$

Multiplying this identity by a and adding the constant part proves the claim. □

Corollary 11 (Squared-hinge retains within-class correction)**.** *If all current pairs satisfy $m_{ij} < m$, then*

$$\widehat{\boldsymbol{g}}_B^{\mathrm{sqh},m}(\theta) = 2 \cdot (m - \bar{s}_+ + \bar{s}_-) \cdot \boldsymbol{d}_B + 2 \cdot \boldsymbol{c}_B.$$

Consequently,

$$\operatorname{dist}\Big(\widehat{\boldsymbol{g}}_B^{\mathrm{sqh},m}(\theta), \operatorname{span}\{\boldsymbol{d}_B\}\Big) = 2 \cdot \operatorname{dist}(\boldsymbol{c}_B, \operatorname{span}\{\boldsymbol{d}_B\})\,.$$

Proof. On the active region, squared-hinge has

$$w_{\mathrm{sqh},m}(x) = 2 \cdot (m - x) = -2 \cdot x + 2 \cdot m.$$

Lemma 2 applies with $a = -2$ and $b = 2 \cdot m$, giving the displayed decomposition.

Let $L_\theta = \mathrm{span}\{\boldsymbol{d}_B\}$. The first term in the decomposition lies in L_θ, so the distance from $\widehat{\boldsymbol{g}}_B^{\mathrm{sqh},m}(\theta)$ to L_θ is exactly the distance from $2 \cdot \boldsymbol{c}_B$ to L_θ. This proves the distance identity, including the case $\boldsymbol{d}_B = \mathbf{0}$, where $L_\theta = \{\mathbf{0}\}$. □

Proposition 20 (Logistic near-decoupling on severe tails)**.** *Assume*

$$\|\boldsymbol{g}_i^+\|_2 \le L_B(\theta), \qquad \|\boldsymbol{g}_j^-\|_2 \le L_B(\theta)$$

for all i, j, and define

$$q_{B,M}(\theta) = \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \mathbf{1}_{\{m_{ij}(\theta) \le -M\}}.$$

Then

$$\left\| \widehat{\boldsymbol{g}}_B^{\log}(\theta) - (\bar{\boldsymbol{g}}_- - \bar{\boldsymbol{g}}_+) \right\|_2 \le 2 \cdot L_B(\theta) \cdot \{1 - q_{B,M}(\theta) \cdot (1 - e^{-M})\}.$$

Proof. Let $\delta_{ij} = 1 - w_{\log}(m_{ij})$. Since

$$w_{\log}(x) = \frac{1}{1 + e^x},$$

we have

$$\widehat{\boldsymbol{g}}_B^{\log}(\theta) - (\bar{\boldsymbol{g}}_- - \bar{\boldsymbol{g}}_+) = -\frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \delta_{ij} \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+).$$

Taking norms and using

$$\|\boldsymbol{g}_j^- - \boldsymbol{g}_i^+\|_2 \le 2 \cdot L_B(\theta)$$

gives

$$\left\| \widehat{\boldsymbol{g}}_B^{\log}(\theta) - (\bar{\boldsymbol{g}}_- - \bar{\boldsymbol{g}}_+) \right\|_2 \le \frac{2 \cdot L_B(\theta)}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \delta_{ij}.$$

If $m_{ij} \le -M$, then

$$\delta_{ij} = \frac{e^{m_{ij}}}{1 + e^{m_{ij}}} \le e^{-M}.$$

If $m_{ij} > -M$, then $\delta_{ij} \le 1$. Hence

$$\frac{1}{B^2} \sum_{i,j} \delta_{ij} \le q_{B,M}(\theta) \cdot e^{-M} + (1 - q_{B,M}(\theta)) = 1 - q_{B,M}(\theta) \cdot (1 - e^{-M}).$$

Substituting this bound proves the result. □

E.3 Easy-Side Silence Versus Overshoot

Proposition 21 (Active-linear easy-side silent profile of squared-hinge within the studied family)**.** *Among logistic, exponential, square, hinge, and squared-hinge, squared-hinge has a weight that is affine with nonzero slope on the whole active region $x < m$ and identically zero on the whole easy region $x \ge m$.*

Proof. Use the explicit weights in Proposition 18. Logistic has

$$w_{\log}(x) = \frac{1}{1 + e^x},$$

which is not affine on $x < m$ and is strictly positive for every x. Exponential has

$$w_{\exp}(x) = e^{-x},$$

which is also non-affine on $x < m$ and strictly positive everywhere. Hinge has

$$w_{\mathrm{h},m}(x) = \mathbf{1}_{\{x<m\}},$$

so it is silent on $x \geq m$, but it is constant on the active side and has zero active-side slope. Square has

$$w_{\mathrm{sq},m}(x) = 2 \cdot (m - x),$$

which is affine with nonzero slope and remains nonzero on the easy side. Squared-hinge has

$$w_{\mathrm{sqh},m}(x) = 2 \cdot (m - x)_+,$$

which equals $2 \cdot (m - x)$ on $x < m$ and 0 on $x \geq m$. It is therefore the only member of the five with both properties. □

Proposition 22 (Exact square-versus-squared-hinge separation)**.** *Square and squared-hinge satisfy*

$$\widehat{\boldsymbol{g}}_B^{\mathrm{sq},m}(\theta) = \widehat{\boldsymbol{g}}_B^{\mathrm{sqh},m}(\theta) - \frac{2}{B^2} \sum_{i,j} (m_{ij} - m)_+ \cdot (\boldsymbol{g}_j^- - \boldsymbol{g}_i^+).$$

Consequently, if

$$\|\boldsymbol{g}_i^+\|_2 \leq L_B(\theta), \qquad \|\boldsymbol{g}_j^-\|_2 \leq L_B(\theta)$$

for all i, j, then

$$\left\| \widehat{\boldsymbol{g}}_B^{\mathrm{sq},m}(\theta) - \widehat{\boldsymbol{g}}_B^{\mathrm{sqh},m}(\theta) \right\|_2 \leq 4 \cdot L_B(\theta) \cdot \frac{1}{B^2} \sum_{i,j} (m_{ij} - m)_+.$$

Proof. For every scalar x,

$$m - x = (m - x)_+ - (x - m)_+.$$

Multiplying by 2 gives

$$w_{\mathrm{sq},m}(x) = 2 \cdot (m - x) = 2 \cdot (m - x)_+ - 2 \cdot (x - m)_+ = w_{\mathrm{sqh},m}(x) - 2 \cdot (x - m)_+.$$

Substituting this identity into Eq. (3) proves the exact separation formula.

For the norm bound, apply the triangle inequality:

$$\left\| \widehat{\boldsymbol{g}}_B^{\mathrm{sq},m}(\theta) - \widehat{\boldsymbol{g}}_B^{\mathrm{sqh},m}(\theta) \right\|_2 \leq \frac{2}{B^2} \sum_{i,j} (m_{ij} - m)_+ \cdot \|\boldsymbol{g}_j^- - \boldsymbol{g}_i^+\|_2.$$

The assumed bounds imply

$$\|\boldsymbol{g}_j^- - \boldsymbol{g}_i^+\|_2 \leq \|\boldsymbol{g}_j^-\|_2 + \|\boldsymbol{g}_i^+\|_2 \leq 2 \cdot L_B(\theta),$$

which gives the displayed inequality. □

E.4 Hard-Tail Alignment

Proposition 23 (Hard-pair reweighting increments)**.** *Fix $t > 0$. Then*

$$\lim_{x \to -\infty} \big(w_{\log}(x - t) - w_{\log}(x) \big) = 0,$$

$$w_{\mathrm{sqh},m}(x - t) - w_{\mathrm{sqh},m}(x) = 2 \cdot t \qquad \textit{whenever } x < m - t,$$

and

$$w_{\exp}(x - t) - w_{\exp}(x) = (e^t - 1) \cdot e^{-x} \to \infty \qquad \textit{as } x \to -\infty.$$

Proof. For logistic, $w_{\log}(x) = 1/(1 + e^x)$, and both $w_{\log}(x)$ and $w_{\log}(x - t)$ converge to 1 as $x \to -\infty$. This proves the first limit.

For squared-hinge, if $x < m - t$, then both x and $x - t$ are active. Hence

$$w_{\mathrm{sqh},m}(x - t) - w_{\mathrm{sqh},m}(x) = 2 \cdot (m - x + t) - 2 \cdot (m - x) = 2 \cdot t.$$

For exponential,

$$w_{\exp}(x - t) - w_{\exp}(x) = e^{-(x-t)} - e^{-x} = (e^t - 1) \cdot e^{-x},$$

which diverges because $t > 0$ and $e^{-x} \to \infty$. □

Proposition 24 (Loss-gradient coupling in the hard tail)**.** *The hard-tail ratios satisfy*

$$\frac{w_{\log}(x)^2}{\ell_{\log}(x)} \to 0 \qquad \text{as } x \to -\infty,$$

$$\frac{w_{\mathrm{sqh},m}(x)^2}{\ell_{\mathrm{sqh},m}(x)} = 4 \qquad \text{for every } x < m,$$

and

$$\frac{w_{\exp}(x)^2}{\ell_{\exp}(x)} = e^{-x} \to \infty \qquad \text{as } x \to -\infty.$$

Proof. For logistic, $w_{\log}(x) \to 1$ as $x \to -\infty$, while

$$\ell_{\log}(x) = \log(1 + e^{-x}) \sim -x \to \infty.$$

Thus $w_{\log}(x)^2/\ell_{\log}(x) \to 0$. For squared-hinge and $x < m$,

$$w_{\mathrm{sqh},m}(x) = 2 \cdot (m - x), \qquad \ell_{\mathrm{sqh},m}(x) = (m - x)^2,$$

so the ratio is 4. For exponential,

$$\frac{w_{\exp}(x)^2}{\ell_{\exp}(x)} = \frac{e^{-2x}}{e^{-x}} = e^{-x},$$

which diverges as $x \to -\infty$. □

E.5 Projected Stochastic Separation

Proposition 25 (Variance separation example: hinge versus squared-hinge)**.** *Fix θ and a unit vector $\boldsymbol{v}$. Assume all sampled pairs satisfy $m_{ij} < m$ almost surely, and there exist constants c_+ and c_- such that*

$$\langle \nabla_\theta f_\theta(X^+), \boldsymbol{v} \rangle = c_+ \quad \text{a.s.}, \qquad \langle \nabla_\theta f_\theta(X^-), \boldsymbol{v} \rangle = c_- \quad \text{a.s.}$$

Then

$$\mathrm{Var}\Big(\langle \widehat{\boldsymbol{g}}_B^{\mathrm{h},m}(\theta), \boldsymbol{v} \rangle \Big) = 0,$$

whereas

$$\mathrm{Var}\Big(\langle \widehat{\boldsymbol{g}}_B^{\mathrm{sqh},m}(\theta), \boldsymbol{v} \rangle \Big) = \frac{4 \cdot (c_- - c_+)^2}{B} \cdot \big\{ \mathrm{Var}(f_\theta(X^+)) + \mathrm{Var}(f_\theta(X^-)) \big\}.$$

Proof. By Proposition 19,

$$\widehat{\boldsymbol{g}}_B^{\mathrm{h},m}(\theta) = \bar{\boldsymbol{g}}_- - \bar{\boldsymbol{g}}_+.$$

Projecting onto $\boldsymbol{v}$ gives

$$\langle \widehat{\boldsymbol{g}}_B^{\mathrm{h},m}(\theta), \boldsymbol{v} \rangle = c_- - c_+,$$

which is deterministic, so the hinge variance is zero.

For squared-hinge, Corollary 11 gives

$$\widehat{\boldsymbol{g}}_B^{\mathrm{sqh},m}(\theta) = 2 \cdot (m - \bar{s}_+ + \bar{s}_-) \cdot \boldsymbol{d}_B + 2 \cdot \boldsymbol{c}_B.$$

The constant projected-gradient assumptions imply

$$\langle \boldsymbol{d}_B, \boldsymbol{v} \rangle = c_- - c_+.$$

They also imply

$$\langle \boldsymbol{c}_B, \boldsymbol{v} \rangle = \frac{1}{B} \sum_i \big(f_\theta(X_i^+) - \bar{s}_+ \big) \cdot c_+ + \frac{1}{B} \sum_j \big(f_\theta(X_j^-) - \bar{s}_- \big) \cdot c_- = 0,$$

because both centered sums vanish. Hence

$$\langle \widehat{\boldsymbol{g}}_B^{\mathrm{sqh},m}(\theta), \boldsymbol{v} \rangle = 2 \cdot (c_- - c_+) \cdot (m - \bar{s}_+ + \bar{s}_-).$$

The constant m does not affect variance, and the positive and negative batches are independent. Therefore

$$\mathrm{Var}(m - \bar{s}_+ + \bar{s}_-) = \mathrm{Var}(\bar{s}_+) + \mathrm{Var}(\bar{s}_-) = \frac{\mathrm{Var}(f_\theta(X^+)) + \mathrm{Var}(f_\theta(X^-))}{B}.$$

Multiplying by $4 \cdot (c_- - c_+)^2$ proves the displayed formula. □

F Changing Where Pairwise AUC Optimization Begins

This section gives the formal statements behind Section 4. The results have a common structure. CE warm-up can change the starting point of the AUC phase by reducing severe positive-negative misrankings. Parameter-efficient fine-tuning can reduce the need for a separate warm-up stage when the trainable subspace starts in, and remains inside, a favorable neighborhood where severe misranking is controlled and first-order projected stochastic motion persists over the finite training horizon.

F.1 CE Warm-Up and Severe-Misranking Mass

Proposition 26 (Pointwise CE risk controls severe-misranking mass)**.** *Define the pointwise logistic risk*

$$R_{\mathrm{CE}}(\theta) = \mathbb{E}[\log(1 + e^{-f_\theta(X^+)})] + \mathbb{E}[\log(1 + e^{f_\theta(X^-)})].$$

Then, for every $M > 0$,

$$\pi_M(\theta) \le \frac{R_{\mathrm{CE}}(\theta)}{\log(1 + e^{M/2})}.$$

Proof. The severe-misranking event is

$$A = \{m_\theta(X^+, X^-) \le -M\} = \{f_\theta(X^-) - f_\theta(X^+) \ge M\}.$$

Let

$$B_+ = \{f_\theta(X^+) \le -M/2\}, \qquad B_- = \{f_\theta(X^-) \ge M/2\}.$$

If A occurs and neither B_+ nor B_- occurs, then

$$f_\theta(X^+) > -M/2, \qquad f_\theta(X^-) < M/2,$$

so $f_\theta(X^-) - f_\theta(X^+) < M$, contradicting A. Therefore $A \subseteq B_+ \cup B_-$, and

$$\pi_M(\theta) \le \mathbb{P}(B_+) + \mathbb{P}(B_-).$$

On B_+,

$$\log(1 + e^{-f_\theta(X^+)}) \ge \log(1 + e^{M/2}),$$

and Markov's inequality gives

$$\mathbb{P}(B_+) \le \frac{\mathbb{E}[\log(1 + e^{-f_\theta(X^+)})]}{\log(1 + e^{M/2})}.$$

The same argument gives

$$\mathbb{P}(B_-) \le \frac{\mathbb{E}[\log(1 + e^{f_\theta(X^-)})]}{\log(1 + e^{M/2})}.$$

Adding the two inequalities proves the claim. □

Corollary 12 (Small CE risk implies low severe-misranking mass)**.** *If a warm-up endpoint θ_{wu} satisfies*

$$R_{\mathrm{CE}}(\theta_{\mathrm{wu}}) \le \delta \cdot \log(1 + e^{M/2}),$$

then

$$\pi_M(\theta_{\mathrm{wu}}) \le \delta.$$

Proof. Apply Proposition 26 at θ_{wu} and substitute the assumed upper bound on $R_{\mathrm{CE}}(\theta_{\mathrm{wu}})$. □

Proposition 27 (Finite-step CE warm-up under local PL and smoothness)**.** *Assume $L_{\mathrm{CE}} > 0$, $\mu_{\mathrm{CE}} > 0$, R_{CE} has L_{CE}-Lipschitz gradient on a region $\mathcal{U}$, and the Polyak-Lojasiewicz condition*

$$\frac{1}{2} \cdot \|\nabla R_{\mathrm{CE}}(\theta)\|^2 \ge \mu_{\mathrm{CE}} \cdot \big(R_{\mathrm{CE}}(\theta) - R^\star_{\mathrm{CE}}\big) \qquad (\theta \in \mathcal{U}),$$

where $R^\star_{\mathrm{CE}} = \inf_{\theta \in \mathcal{U}} R_{\mathrm{CE}}(\theta)$. Consider the fixed-step CE dynamics

$$\theta_{t+1} = \theta_t - \alpha \cdot \nabla R_{\mathrm{CE}}(\theta_t), \qquad 0 < \alpha \le \min\{1/L_{\mathrm{CE}}, 1/\mu_{\mathrm{CE}}\},$$

and suppose $\theta_0, \dots, \theta_T \in \mathcal{U}$. *Then*

$$R_{\mathrm{CE}}(\theta_t) - R_{\mathrm{CE}}^\star \le (1 - \alpha \cdot \mu_{\mathrm{CE}})^t \cdot \big(R_{\mathrm{CE}}(\theta_0) - R_{\mathrm{CE}}^\star\big) \qquad (0 \le t \le T).$$

Consequently, if

$$R_{\mathrm{CE}}^\star + (1 - \alpha \cdot \mu_{\mathrm{CE}})^T \cdot \big(R_{\mathrm{CE}}(\theta_0) - R_{\mathrm{CE}}^\star\big) \le \delta \cdot \log(1 + e^{M/2}),$$

then $\pi_M(\theta_T) \le \delta$.

Proof. For $t < T$, smoothness gives

$$R_{\mathrm{CE}}(\theta_{t+1}) \le R_{\mathrm{CE}}(\theta_t) + \langle \nabla R_{\mathrm{CE}}(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{L_{\mathrm{CE}}}{2} \cdot \|\theta_{t+1} - \theta_t\|^2.$$

Substituting the CE gradient step,

$$\theta_{t+1} - \theta_t = -\alpha \cdot \nabla R_{\mathrm{CE}}(\theta_t),$$

gives

$$R_{\mathrm{CE}}(\theta_{t+1}) \le R_{\mathrm{CE}}(\theta_t) - \left(\alpha - \frac{L_{\mathrm{CE}} \cdot \alpha^2}{2}\right) \cdot \|\nabla R_{\mathrm{CE}}(\theta_t)\|^2.$$

Because $\alpha \le 1/L_{\mathrm{CE}}$,

$$\alpha - \frac{L_{\mathrm{CE}} \cdot \alpha^2}{2} \ge \frac{\alpha}{2}.$$

The PL inequality therefore gives

$$R_{\mathrm{CE}}(\theta_{t+1}) - R_{\mathrm{CE}}^\star \le (1 - \alpha \cdot \mu_{\mathrm{CE}}) \cdot \big(R_{\mathrm{CE}}(\theta_t) - R_{\mathrm{CE}}^\star\big).$$

Iterating the one-step contraction proves the displayed risk bound. The final severe-misranking bound follows by evaluating the risk bound at T and applying Corollary 12. □

Proposition 28 (Warm-up transfer to a low-error endpoint)**.** *Fix* $M' > 0$ *and* $r \ge 0$. *Assume pairwise margins are locally* L_{sc}*-Lipschitz on a neighborhood containing* $\bar{\theta}$ *and* θ_{wu}*:*

$$|m_\theta(x^+, x^-) - m_{\theta'}(x^+, x^-)| \le L_{\mathrm{sc}} \cdot \|\theta - \theta'\|$$

for all relevant θ, θ' *and all pairs* (x^+, x^-). *If*

$$\|\theta_{\mathrm{wu}} - \bar{\theta}\| \le r \qquad \textit{and} \qquad \pi_{M'}(\bar{\theta}) \le \delta,$$

then

$$\pi_{M' + L_{\mathrm{sc}} \cdot r}(\theta_{\mathrm{wu}}) \le \delta.$$

Proof. On the event

$$\{m_{\theta_{\mathrm{wu}}}(X^+, X^-) \le -(M' + L_{\mathrm{sc}} \cdot r)\},$$

the Lipschitz condition and $\|\theta_{\mathrm{wu}} - \bar{\theta}\| \le r$ imply

$$m_{\bar{\theta}}(X^+, X^-) \le m_{\theta_{\mathrm{wu}}}(X^+, X^-) + L_{\mathrm{sc}} \cdot \|\theta_{\mathrm{wu}} - \bar{\theta}\| \le -M'.$$

Thus

$$\{m_{\theta_{\mathrm{wu}}}(X^+, X^-) \le -(M' + L_{\mathrm{sc}} \cdot r)\} \subseteq \{m_{\bar{\theta}}(X^+, X^-) \le -M'\}.$$

Taking probabilities proves the claim. □

Proposition 29 (Post-warm-up low severe-misranking neighborhood)**.** *Fix* $M > 0$, $r \ge 0$, *and* $L_{\mathrm{sc}} \ge 0$ *such that* $M - L_{\mathrm{sc}} \cdot r > 0$. *Assume that the pairwise margin map is locally* L_{sc}*-Lipschitz around a warm-up endpoint* θ_{wu}*:*

$$|m_\theta(x^+, x^-) - m_{\theta_{\mathrm{wu}}}(x^+, x^-)| \le L_{\mathrm{sc}} \cdot \|\theta - \theta_{\mathrm{wu}}\|$$

for all θ *with* $\|\theta - \theta_{\mathrm{wu}}\| \le r$ *and all sample pairs. If*

$$\pi_{M - L_{\mathrm{sc}} \cdot r}(\theta_{\mathrm{wu}}) \le \delta,$$

then every θ *satisfying* $\|\theta - \theta_{\mathrm{wu}}\| \le r$ *also satisfies*

$$\pi_M(\theta) \le \delta.$$

Proof. If $m_\theta(X^+, X^-) \le -M$ and $\|\theta - \theta_{\mathrm{wu}}\| \le r$, then

$$m_{\theta_{\mathrm{wu}}}(X^+, X^-) \le m_\theta(X^+, X^-) + L_{\mathrm{sc}} \cdot r \le -(M - L_{\mathrm{sc}} \cdot r).$$

Therefore

$$\{m_\theta(X^+, X^-) \le -M\} \subseteq \{m_{\theta_{\mathrm{wu}}}(X^+, X^-) \le -(M - L_{\mathrm{sc}} \cdot r)\}.$$

Taking probabilities gives $\pi_M(\theta) \le \delta$. □

Corollary 13 (Finite-horizon post-warm-up neighborhood check)**.** *Under the assumptions of Proposition 29, let*

$$\widehat{\pi}_{M - L_{\mathrm{sc}} \cdot r, K}(\theta_{\mathrm{wu}})$$

be an independent finite-pair estimator as in Proposition 14. Consider a second-stage trajectory

$$\theta_0, \theta_1, \dots, \theta_T, \qquad \theta_0 = \theta_{\mathrm{wu}},$$

such that $\|\theta_t - \theta_{\mathrm{wu}}\| \le r$ *for all* $0 \le t \le T$*. Then, for every* $\varepsilon > 0$*, with probability at least*

$$1 - \frac{1}{4 \cdot K \cdot \varepsilon^2},$$

one has the implication

$$\widehat{\pi}_{M - L_{\mathrm{sc}} \cdot r, K}(\theta_{\mathrm{wu}}) + \varepsilon \le \delta \quad \Longrightarrow \quad \pi_M(\theta_t) \le \delta \quad \textit{for all } 0 \le t \le T.$$

Proof. By Proposition 14, with probability at least $1 - 1/(4 \cdot K \cdot \varepsilon^2)$,

$$\pi_{M - L_{\mathrm{sc}} \cdot r}(\theta_{\mathrm{wu}}) \le \widehat{\pi}_{M - L_{\mathrm{sc}} \cdot r, K}(\theta_{\mathrm{wu}}) + \varepsilon.$$

On this event, the premise of the implication gives

$$\pi_{M - L_{\mathrm{sc}} \cdot r}(\theta_{\mathrm{wu}}) \le \delta.$$

Since every θ_t lies within radius r of θ_{wu}, Proposition 29 yields $\pi_M(\theta_t) \le \delta$ for all $0 \le t \le T$. □

F.2 Parameter-Efficient Fine-Tuning in a Stable Trainable Subspace

We now consider a pretrained model with a frozen base and a low-dimensional trainable subspace, such as a head, adapter, or LoRA module. Let $\boldsymbol{\xi}$ denote the trainable coordinates and absorb the frozen base into the score map $f_{\boldsymbol{\xi}}$. The results below state sufficient conditions under which direct pairwise AUC-surrogate training starts in a favorable neighborhood and remains there for the finite horizon of interest.

Proposition 30 (Trainable-subspace transfer of pairwise identities)**.** *Let the trainable parameter be* $\boldsymbol{\xi}$*, and let the frozen pretrained base be absorbed into* $f_{\boldsymbol{\xi}}$*. If* $f_{\boldsymbol{\xi}}$ *and* $\nabla_{\boldsymbol{\xi}} f_{\boldsymbol{\xi}}$ *are locally well defined, then the pairwise gradient identity, Hoeffding variance split, and surrogate-weight identities from Appendices D and E remain valid after replacing* θ *by* $\boldsymbol{\xi}$*.*

Proof. Each identity used in those appendices depends only on the score map, its gradient with respect to the trainable variable, the positive margin

$$m_{\boldsymbol{\xi}}(x^+, x^-) = f_{\boldsymbol{\xi}}(x^+) - f_{\boldsymbol{\xi}}(x^-),$$

and algebraic operations such as pair averaging, conditioning, variance decomposition, and applying $w_\ell(x) = -\ell'(x)$. If the pretrained base is frozen, the same objects exist with $\boldsymbol{\xi}$ as the trainable variable:

$$f_\theta \mapsto f_{\boldsymbol{\xi}}, \qquad \nabla_\theta f_\theta \mapsto \nabla_{\boldsymbol{\xi}} f_{\boldsymbol{\xi}}.$$

No derivation requires the frozen coordinates to be trainable. Therefore the same identities hold verbatim in the trainable subspace. □

Proposition 31 (Score perturbation in a trainable subspace)**.** *Fix a trainable-subspace initialization* $\boldsymbol{\xi}_0$ *and radius* $\rho > 0$*. Suppose*

$$\sup_{\boldsymbol{\zeta} \in B(\boldsymbol{\xi}_0, \rho)} \sup_x \|\nabla_{\boldsymbol{\xi}} f_{\boldsymbol{\zeta}}(x)\| \le L_f.$$

Then, for any $\boldsymbol{\xi}, \boldsymbol{\xi}' \in B(\boldsymbol{\xi}_0, \rho)$ *and every input* x*,*

$$|f_{\boldsymbol{\xi}}(x) - f_{\boldsymbol{\xi}'}(x)| \le L_f \cdot \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|,$$

and for every pair,

$$|m_{\boldsymbol{\xi}}(x^+, x^-) - m_{\boldsymbol{\xi}'}(x^+, x^-)| \le 2 \cdot L_f \cdot \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|.$$

Proof. For a fixed x, define

$$\phi(t) = f_{\boldsymbol{\xi}'+t\cdot(\boldsymbol{\xi}-\boldsymbol{\xi}')}(x), \qquad t \in [0,1].$$

The line segment between $\boldsymbol{\xi}'$ and $\boldsymbol{\xi}$ lies in $B(\boldsymbol{\xi}_0, \rho)$. By the fundamental theorem of calculus and Cauchy's inequality,

$$\begin{aligned} |f_{\boldsymbol{\xi}}(x) - f_{\boldsymbol{\xi}'}(x)| &= \left| \int_0^1 \left\langle \nabla_{\boldsymbol{\xi}} f_{\boldsymbol{\xi}'+t\cdot(\boldsymbol{\xi}-\boldsymbol{\xi}')}(x), \boldsymbol{\xi} - \boldsymbol{\xi}' \right\rangle dt \right| \\ &\le \int_0^1 L_f \cdot \|\boldsymbol{\xi} - \boldsymbol{\xi}'\| \, dt = L_f \cdot \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|. \end{aligned}$$

The margin bound follows from

$$m_{\boldsymbol{\xi}}(x^+, x^-) - m_{\boldsymbol{\xi}'}(x^+, x^-) = \big(f_{\boldsymbol{\xi}}(x^+) - f_{\boldsymbol{\xi}'}(x^+)\big) - \big(f_{\boldsymbol{\xi}}(x^-) - f_{\boldsymbol{\xi}'}(x^-)\big)$$

and the triangle inequality. □

Proposition 32 (Uniform severe-misranking control in a stable trainable subspace)**.** *Under the assumptions of Proposition 31, suppose $M > 4 \cdot L_f \cdot \rho$ and a reference point $\bar{\boldsymbol{\xi}} \in B(\boldsymbol{\xi}_0, \rho)$ satisfies*

$$\pi_{M-4\cdot L_f\cdot\rho}(\bar{\boldsymbol{\xi}}) \le \delta_M.$$

Then every $\boldsymbol{\xi} \in B(\boldsymbol{\xi}_0, \rho)$ satisfies

$$\pi_M(\boldsymbol{\xi}) \le \delta_M.$$

Proof. Fix $\boldsymbol{\xi} \in B(\boldsymbol{\xi}_0, \rho)$. On the event $m_{\boldsymbol{\xi}}(X^+, X^-) \le -M$, Proposition 31 gives

$$m_{\bar{\boldsymbol{\xi}}}(X^+, X^-) \le -M + 2 \cdot L_f \cdot \|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}\|.$$

Since both $\boldsymbol{\xi}$ and $\bar{\boldsymbol{\xi}}$ lie in $B(\boldsymbol{\xi}_0, \rho)$,

$$\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}\| \le \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\| + \|\bar{\boldsymbol{\xi}} - \boldsymbol{\xi}_0\| \le 2 \cdot \rho.$$

Therefore

$$m_{\bar{\boldsymbol{\xi}}}(X^+, X^-) \le -(M - 4 \cdot L_f \cdot \rho).$$

Thus

$$\{m_{\boldsymbol{\xi}}(X^+, X^-) \le -M\} \subseteq \{m_{\bar{\boldsymbol{\xi}}}(X^+, X^-) \le -(M - 4 \cdot L_f \cdot \rho)\}.$$

Taking probabilities proves the claim. □

Proposition 33 (Persistence of first-order projected variance)**.** *Fix a unit vector $\boldsymbol{v}$. Suppose that, on a neighborhood $\mathcal{U}$ of a trainable-subspace initialization, the maps*

$$\boldsymbol{\xi} \mapsto \psi^{(1,0)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^+), \qquad \boldsymbol{\xi} \mapsto \psi^{(0,1)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^-)$$

are continuous in L_2. If, for some reference point $\bar{\boldsymbol{\xi}} \in \mathcal{U}$,

$$\mathrm{Var}(\psi^{(1,0)}_{\bar{\boldsymbol{\xi}},\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\bar{\boldsymbol{\xi}},\boldsymbol{v}}(X^-)) > 0,$$

then there is a smaller neighborhood $\mathcal{U}' \subseteq \mathcal{U}$ and a constant $\sigma^2_{1,\boldsymbol{v}} > 0$ such that

$$\mathrm{Var}(\psi^{(1,0)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^-)) \ge \sigma^2_{1,\boldsymbol{v}} \qquad (\boldsymbol{\xi} \in \mathcal{U}').$$

Proof. Let

$$V(\boldsymbol{\xi}) = \mathrm{Var}(\psi^{(1,0)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^-)).$$

L_2-continuity of each projection implies continuity of its second moment and its mean, hence continuity of its variance. Therefore V is continuous. Since $V(\bar{\boldsymbol{\xi}}) > 0$, there exists a neighborhood $\mathcal{U}' \subseteq \mathcal{U}$ such that

$$|V(\boldsymbol{\xi}) - V(\bar{\boldsymbol{\xi}})| \le \frac{V(\bar{\boldsymbol{\xi}})}{2} \qquad (\boldsymbol{\xi} \in \mathcal{U}').$$

Setting $\sigma^2_{1,\boldsymbol{v}} = V(\bar{\boldsymbol{\xi}})/2$ proves the result. □

Proposition 34 (Finite-horizon confinement from a bounded update budget)**.** *Consider discrete dynamics*

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \cdot \widehat{\boldsymbol{g}}_t, \qquad t = 0, \dots, T-1.$$

Assume there are $\rho > 0$ and $g_* < \infty$ such that $B(\boldsymbol{\xi}_0, \rho) \subseteq \mathcal{U}$,

$$\|\widehat{\boldsymbol{g}}_t\| \le g_* \qquad \textit{whenever } \boldsymbol{\xi}_t \in \mathcal{U},$$

and

$$\sum_{t=0}^{T-1} \eta_t \cdot g_* \le \rho.$$

Then $\boldsymbol{\xi}_t \in B(\boldsymbol{\xi}_0, \rho) \subseteq \mathcal{U}$ for every $0 \le t \le T$.

Proof. We prove by induction that $\boldsymbol{\xi}_t \in B(\boldsymbol{\xi}_0, \rho)$. The claim is trivial at $t = 0$. If it holds up to time t, then $\boldsymbol{\xi}_s \in \mathcal{U}$ for $0 \le s \le t$, and hence

$$\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_0 = -\sum_{s=0}^{t} \eta_s \cdot \widehat{\boldsymbol{g}}_s.$$

Taking norms gives

$$\|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_0\| \le \sum_{s=0}^{t} \eta_s \cdot \|\widehat{\boldsymbol{g}}_s\| \le \sum_{s=0}^{T-1} \eta_s \cdot g_* \le \rho.$$

Thus $\boldsymbol{\xi}_{t+1} \in B(\boldsymbol{\xi}_0, \rho)$, closing the induction. □

Proposition 35 (Finite-horizon reduced need for warm-up)**.** *Suppose a neighborhood $\mathcal{U}$ of the trainable-subspace initialization satisfies*

$$\sup_{\boldsymbol{\xi} \in \mathcal{U}} \pi_M(\boldsymbol{\xi}) \le \delta_M$$

and, for some unit vector $\boldsymbol{v}$,

$$\mathrm{Var}(\psi^{(1,0)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^-)) \ge \sigma^2_{1,\boldsymbol{v}} > 0 \qquad (\boldsymbol{\xi} \in \mathcal{U}).$$

If a finite-horizon trajectory remains in $\mathcal{U}$, then every iterate on that horizon satisfies

$$\pi_M(\boldsymbol{\xi}_t) \le \delta_M, \qquad \mathrm{Var}(\langle \widehat{\boldsymbol{g}}^{\ell}_B(\boldsymbol{\xi}_t), \boldsymbol{v} \rangle) \ge \frac{\sigma^2_{1,\boldsymbol{v}}}{B}.$$

Proof. The severe-misranking bound follows directly from the uniform hypothesis. For the variance bound, Proposition 30 transfers the variance split in Eq. (12) to the trainable coordinate $\boldsymbol{\xi}_t$:

$$\mathrm{Var}(\langle \widehat{\boldsymbol{g}}^{\ell}_B(\boldsymbol{\xi}_t), \boldsymbol{v} \rangle) = \frac{\mathrm{Var}(\psi^{(1,0)}_{\boldsymbol{\xi}_t,\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\boldsymbol{\xi}_t,\boldsymbol{v}}(X^-))}{B} + \frac{\mathrm{Var}(\psi^{(1,1)}_{\boldsymbol{\xi}_t,\boldsymbol{v}}(X^+, X^-))}{B^2}.$$

The residual variance is nonnegative, and the first-order term is bounded below by $\sigma^2_{1,\boldsymbol{v}}$ throughout $\mathcal{U}$. The displayed lower bound follows. □

Proposition 36 (Three checks for finite-horizon direct pairwise AUC-surrogate training from a pretrained start)**.** *A finite-horizon conclusion of the form in Proposition 35 follows from three checks:*

1. ***Favorable initialization.*** *Verify a neighborhood $\mathcal{U}$ of the pretrained trainable-subspace initialization where severe-misranking mass is uniformly controlled and the first-order projected variance has a uniform positive lower bound.*
2. ***Finite-horizon confinement.*** *Verify that the direct pairwise AUC-surrogate trajectory remains in $\mathcal{U}$, for example by Proposition 34.*
3. ***Pointwise transfer.*** *Once each iterate lies in $\mathcal{U}$, the severe-misranking and first-order variance bounds transfer to every iterate by Proposition 35.*

Proof. The first check supplies the two uniform hypotheses of Proposition 35. The second check supplies its trajectory containment assumption. Substituting each iterate into the uniform neighborhood bounds gives the pointwise severe-misranking control and first-order variance lower bound on the whole finite horizon. □

F.3 Local Pairwise AUC-Surrogate Descent in a Favorable Subspace

Proposition 37 (Local pairwise AUC-surrogate descent in a favorable trainable subspace)**.** *Let*

$$R_\ell(\boldsymbol{\xi}) = \mathbb{E}[\ell(m_{\boldsymbol{\xi}}(X^+, X^-))]$$

be the pairwise AUC surrogate risk in the trainable subspace. Suppose $L_\ell > 0$ and R_ℓ has L_ℓ-Lipschitz gradient on a neighborhood $\mathcal{U}$. Consider dynamics

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t \cdot \widehat{\boldsymbol{g}}_t, \qquad t = 0, \dots, T-1,$$

adapted to a filtration $(\mathcal{F}_t)_{t\geq 0}$, with $\boldsymbol{\xi}_0, \dots, \boldsymbol{\xi}_T \in \mathcal{U}$ almost surely. Assume

$$\mathbb{E}[\widehat{\boldsymbol{g}}_t \mid \mathcal{F}_t] = \nabla R_\ell(\boldsymbol{\xi}_t), \qquad \mathbb{E}[\|\widehat{\boldsymbol{g}}_t - \nabla R_\ell(\boldsymbol{\xi}_t)\|^2 \mid \mathcal{F}_t] \leq \sigma^2,$$

and $0 < \eta_t \leq 1/L_\ell$. Then

$$\mathbb{E}[R_\ell(\boldsymbol{\xi}_{t+1}) \mid \mathcal{F}_t] \leq R_\ell(\boldsymbol{\xi}_t) - \frac{\eta_t}{2} \cdot \|\nabla R_\ell(\boldsymbol{\xi}_t)\|^2 + \frac{L_\ell \cdot \eta_t^2}{2} \cdot \sigma^2.$$

Consequently, if $R^\star_{\ell,\mathcal{U}} = \inf_{\boldsymbol{\xi}\in\mathcal{U}} R_\ell(\boldsymbol{\xi})$, then

$$\sum_{t=0}^{T-1} \frac{\eta_t}{2} \cdot \mathbb{E}[\|\nabla R_\ell(\boldsymbol{\xi}_t)\|^2] \leq R_\ell(\boldsymbol{\xi}_0) - R^\star_{\ell,\mathcal{U}} + \frac{L_\ell \cdot \sigma^2}{2} \sum_{t=0}^{T-1} \eta_t^2.$$

Proof. Smoothness gives

$$R_\ell(\boldsymbol{\xi}_{t+1}) \leq R_\ell(\boldsymbol{\xi}_t) + \langle \nabla R_\ell(\boldsymbol{\xi}_t), \boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t \rangle + \frac{L_\ell}{2} \cdot \|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\|^2.$$

Substitute $\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t = -\eta_t \cdot \widehat{\boldsymbol{g}}_t$ and condition on $\mathcal{F}_t$:

$$\begin{aligned}\mathbb{E}[R_\ell(\boldsymbol{\xi}_{t+1}) \mid \mathcal{F}_t] \leq R_\ell(\boldsymbol{\xi}_t) &- \eta_t \cdot \langle \nabla R_\ell(\boldsymbol{\xi}_t), \mathbb{E}[\widehat{\boldsymbol{g}}_t \mid \mathcal{F}_t] \rangle \\ &+ \frac{L_\ell \cdot \eta_t^2}{2} \cdot \mathbb{E}[\|\widehat{\boldsymbol{g}}_t\|^2 \mid \mathcal{F}_t].\end{aligned}$$

The conditional unbiasedness assumption gives the inner product $\|\nabla R_\ell(\boldsymbol{\xi}_t)\|^2$. Also,

$$\mathbb{E}[\|\widehat{\boldsymbol{g}}_t\|^2 \mid \mathcal{F}_t] \leq \|\nabla R_\ell(\boldsymbol{\xi}_t)\|^2 + \sigma^2,$$

because the cross term in $\widehat{\boldsymbol{g}}_t = (\widehat{\boldsymbol{g}}_t - \nabla R_\ell(\boldsymbol{\xi}_t)) + \nabla R_\ell(\boldsymbol{\xi}_t)$ has conditional expectation zero. Thus

$$\mathbb{E}[R_\ell(\boldsymbol{\xi}_{t+1}) \mid \mathcal{F}_t] \leq R_\ell(\boldsymbol{\xi}_t) - \eta_t \cdot \left(1 - \frac{L_\ell \cdot \eta_t}{2}\right) \cdot \|\nabla R_\ell(\boldsymbol{\xi}_t)\|^2 + \frac{L_\ell \cdot \eta_t^2}{2} \cdot \sigma^2.$$

Since $\eta_t \leq 1/L_\ell$, $1 - L_\ell \cdot \eta_t/2 \geq 1/2$, proving the one-step bound. Taking expectations, summing over t, and using $\mathbb{E}[R_\ell(\boldsymbol{\xi}_T)] \geq R^\star_{\ell,\mathcal{U}}$ proves the finite-horizon inequality. □

Corollary 14 (Local PL contraction in a favorable trainable subspace)**.** *In the setting of Proposition 37, assume additionally that $\mu > 0$ and R_ℓ satisfies the local PL inequality on $\mathcal{U}$:*

$$\frac{1}{2} \cdot \|\nabla R_\ell(\boldsymbol{\xi})\|^2 \geq \mu \cdot \big(R_\ell(\boldsymbol{\xi}) - R^\star_{\ell,\mathcal{U}}\big) \qquad (\boldsymbol{\xi} \in \mathcal{U}).$$

For a constant step size $0 < \eta \leq \min\{1/L_\ell, 1/\mu\}$, if $\boldsymbol{\xi}_0, \dots, \boldsymbol{\xi}_T \in \mathcal{U}$ almost surely, then

$$\mathbb{E}[R_\ell(\boldsymbol{\xi}_t) - R^\star_{\ell,\mathcal{U}}] \leq (1 - \eta \cdot \mu)^t \cdot \big(R_\ell(\boldsymbol{\xi}_0) - R^\star_{\ell,\mathcal{U}}\big) + \frac{L_\ell \cdot \eta \cdot \sigma^2}{2 \cdot \mu} \cdot \big(1 - (1 - \eta \cdot \mu)^t\big)$$

for every $0 \leq t \leq T$.

Proof. With $\eta_t = \eta$, Proposition 37 gives

$$\mathbb{E}[R_\ell(\boldsymbol{\xi}_{t+1}) \mid \mathcal{F}_t] \leq R_\ell(\boldsymbol{\xi}_t) - \frac{\eta}{2} \cdot \|\nabla R_\ell(\boldsymbol{\xi}_t)\|^2 + \frac{L_\ell \cdot \eta^2}{2} \cdot \sigma^2.$$

The PL inequality gives

$$\frac{1}{2} \cdot \|\nabla R_\ell(\boldsymbol{\xi}_t)\|^2 \geq \mu \cdot \big(R_\ell(\boldsymbol{\xi}_t) - R^\star_{\ell,\mathcal{U}}\big).$$

After subtracting $R^\star_{\ell,\mathcal{U}}$, taking expectations, and iterating the resulting scalar inequality,

$$a_{t+1} \leq (1 - \eta \cdot \mu) \cdot a_t + \frac{L_\ell \cdot \eta^2}{2} \cdot \sigma^2, \qquad a_t = \mathbb{E}[R_\ell(\boldsymbol{\xi}_t) - R^\star_{\ell,\mathcal{U}}],$$

we obtain

$$a_t \leq (1 - \eta \cdot \mu)^t a_0 + \frac{L_\ell \cdot \eta^2}{2} \cdot \sigma^2 \sum_{s=0}^{t-1} (1 - \eta \cdot \mu)^s.$$

The geometric sum equals

$$\frac{1 - (1 - \eta \cdot \mu)^t}{\eta \cdot \mu},$$

which proves the claim. □

Proposition 38 (Direct pairwise AUC-surrogate training in a favorable trainable subspace)**.** *Suppose a trainable-subspace neighborhood $\mathcal{U}$ satisfies*

$$\sup_{\boldsymbol{\xi} \in \mathcal{U}} \pi_M(\boldsymbol{\xi}) \leq \delta_M$$

and, for some unit vector $\boldsymbol{v}$,

$$\mathrm{Var}(\psi^{(1,0)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^+)) + \mathrm{Var}(\psi^{(0,1)}_{\boldsymbol{\xi},\boldsymbol{v}}(X^-)) \geq \sigma^2_{1,\boldsymbol{v}} > 0 \qquad (\boldsymbol{\xi} \in \mathcal{U}).$$

Assume the direct pairwise AUC-surrogate dynamics in the trainable subspace are conditionally unbiased, satisfy the moment bound of Proposition 37, and stay in $\mathcal{U}$ on the finite horizon. Then, on that horizon,

$$\pi_M(\boldsymbol{\xi}_t) \leq \delta_M, \qquad \mathrm{Var}(\langle \widehat{\boldsymbol{g}}^\ell_B(\boldsymbol{\xi}_t), \boldsymbol{v} \rangle) \geq \frac{\sigma^2_{1,\boldsymbol{v}}}{B},$$

and the finite-horizon descent bound in Proposition 37 holds. If the local PL condition also holds, the contraction bound in Corollary 14 holds as well.

Proof. The severe-misranking and first-order variance statements follow from Proposition 35. The conditional unbiasedness and moment assumptions place the same trajectory in the setting of Proposition 37, so the descent bound holds. If the local PL condition is added, Corollary 14 applies. □

G Model-Class Realizations of Projected Bad Components

This section supports the model-class side of Section 2. It constructs Transformer-style scorers in which the projected bad-component conditions from Appendix D hold inside the model parameterization, beyond the abstract linear example. The attention map is genuinely data-dependent: the query, key, value, and output maps are nonzero, and the classification token assigns different attention weights to the data token for the two possible inputs. The results are existence statements. They produce neighborhoods in which a specified trainable direction has weak drift, collapsed first-order stochastic variation, and nonzero residual pair-interaction variance. Degeneracy is established for the constructed projected direction.

G.1 Data-Dependent One-Layer Construction

Proposition 39 (Data-dependent one-layer Transformer region with high error and first-order variance collapse)**.** *Fix $M > 0$, a local radius $r > 0$, a deterministic-drift tolerance $\delta > 0$, and any $\lambda > 0$. There is a one-head, one-layer residual self-attention scorer with nonzero query, key, value, and output maps, a single trainable scalar θ, and a scale $\gamma > 0$ such that the classification-token attention weight depends on the input and, on the nonempty interval*

$$\mathcal{U}_{\gamma,M,r} := \left[\frac{M}{\gamma}, \frac{M}{\gamma} + r\right],$$

the following properties hold for the pairwise square loss $\ell(x) = x^2$. For this loss, let

$$G_\theta(A, C) := \frac{d}{d\theta}\ell(m_\theta(A, C))$$

denote the scalar gradient kernel in the trainable direction.

1. *The attention weight placed by the classification token on the data token is*

$$\alpha_z = \frac{\exp(\lambda \cdot z)}{1 + \exp(\lambda \cdot z)}, \qquad z \in \{-1, 1\},$$

   so $\alpha_{+1} \neq \alpha_{-1}$.

2. *The severe-misranking mass at level M is*

$$\pi_M(\theta) := \mathbb{P}(m_\theta(X^+, X^-) \leq -M) = \frac{1}{4}, \qquad \theta \in \mathcal{U}_{\gamma,M,r}.$$

3. *The expected scalar gradient in the trainable direction satisfies*

$$\sup_{\theta \in \mathcal{U}_{\gamma,M,r}} \left|\mathbb{E}[G_\theta(X^+, X^-)]\right| \leq \delta$$

   whenever $\gamma > 0$ is chosen sufficiently small.

4. *Both first-order Hoeffding projections of the scalar pairwise-gradient kernel vanish for every θ:*

$$\mathbb{E}[G_\theta(A, C) \mid A] - \mathbb{E}[G_\theta(A, C)] = 0, \qquad \mathbb{E}[G_\theta(A, C) \mid C] - \mathbb{E}[G_\theta(A, C)] = 0.$$

   Consequently, for a mini-batch with B positives and B negatives,

$$\mathrm{Var}\left(\frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} G_\theta(A_i, C_j)\right) = \frac{\gamma^4 \cdot \theta^2}{B^2}.$$

Proof. Let A and C be independent Rademacher random variables. A positive example is the two-token sequence $X^+ = s_A$, and a negative example is the two-token sequence $X^- = s_C$. For $z \in \{-1, 1\}$, the sequence s_z consists of a classification token and a data token with embeddings

$$\boldsymbol{e}_{\mathrm{cls}} = (1, 0, 0), \qquad \boldsymbol{e}_z = (0, z, 0) \qquad \text{in } \mathbb{R}^3.$$

There are no positional embeddings. Consider a single-head residual self-attention block without layer normalization. For a two-token hidden-state matrix $\boldsymbol{H} \in \mathbb{R}^{2\times 3}$, the attention output is

$$\mathrm{Attn}(\boldsymbol{H}) = \mathrm{softmax}\big((\boldsymbol{H} \cdot \boldsymbol{W}_Q) \cdot (\boldsymbol{H} \cdot \boldsymbol{W}_K)^\top\big) \cdot (\boldsymbol{H} \cdot \boldsymbol{W}_V) \cdot \boldsymbol{W}_O.$$

Choose

$$\boldsymbol{W}_Q = \begin{pmatrix}1\\0\\0\end{pmatrix}, \qquad \boldsymbol{W}_K = \begin{pmatrix}0\\\lambda\\0\end{pmatrix}, \qquad \boldsymbol{W}_V = \begin{pmatrix}0\\1\\0\end{pmatrix}, \qquad \boldsymbol{W}_O = \begin{pmatrix}0 & 1 & 0\end{pmatrix}.$$

All four maps are nonzero. For the classification token, the query is 1, the classification-token key is 0, and the data-token key is $\lambda \cdot z$. Hence the classification-token attention weight on the data token is

$$\alpha_z = \frac{\exp(\lambda \cdot z)}{1 + \exp(\lambda \cdot z)}.$$

Since $\lambda > 0$, $\alpha_{+1} \neq \alpha_{-1}$. The value at the classification token is 0, while the value at the data token is z:

$$\boldsymbol{e}_{\mathrm{cls}} \cdot \boldsymbol{W}_V = 0, \qquad \boldsymbol{e}_z \cdot \boldsymbol{W}_V = z.$$

Thus the attention output added to the classification token is

$$(0, \alpha_z \cdot z, 0).$$

The residual addition leaves the classification-token representation equal to

$$\boldsymbol{r}_z = (1, \alpha_z \cdot z, 0).$$

Let the feed-forward map be identically zero, so the residual feed-forward sublayer leaves $\boldsymbol{r}_z$ unchanged.

Now attach the trainable scalar head

$$\boldsymbol{a}_\theta = (0, \gamma \cdot \theta, 0)^\top, \qquad \gamma > 0,$$

and define the score by

$$f_\theta(s_z) = \boldsymbol{a}_\theta^\top \cdot \boldsymbol{r}_z.$$

Substituting $\boldsymbol{r}_z = (1, \alpha_z \cdot z, 0)$ gives

$$f_\theta(s_z) = (0, \gamma \cdot \theta, 0) \cdot (1, \alpha_z \cdot z, 0) = \gamma \cdot \theta \cdot \alpha_z \cdot z.$$

Thus the one-layer Transformer realizes a one-dimensional scorer through a nonconstant attention weight. Define

$$h_z = \alpha_z \cdot z.$$

Because $\alpha_{+1} + \alpha_{-1} = 1$, we have

$$h_{+1} - h_{-1} = 1.$$

For a positive value A and a negative value C, the positive margin is

$$m_\theta(A, C) = f_\theta(s_A) - f_\theta(s_C) = \gamma \cdot \theta \cdot (h_A - h_C).$$

If $\theta \in \mathcal{U}_{\gamma,M,r}$, then $\theta \geq M/\gamma$. On the event $\{A = -1, C = 1\}$, which has probability $1/4$,

$$m_\theta(A, C) = -\gamma \cdot \theta \leq -M.$$

On the other three Rademacher outcomes, $h_A - h_C \in \{0, 1\}$, so $m_\theta(A, C) \geq 0$. Hence

$$\pi_M(\theta) = \mathbb{P}(m_\theta(A, C) \leq -M) = \frac{1}{4}.$$

For $\ell(x) = x^2$, the scalar gradient kernel in the trainable direction is

$$G_\theta(A, C) = \frac{d}{d\theta}\ell(m_\theta(A, C)) = 2 \cdot m_\theta(A, C) \cdot \frac{d}{d\theta} m_\theta(A, C).$$

Since

$$\frac{d}{d\theta} m_\theta(A, C) = \gamma \cdot (h_A - h_C),$$

we obtain

$$G_\theta(A, C) = 2 \cdot \gamma^2 \cdot \theta \cdot (h_A - h_C)^2.$$

Because $h_A - h_C$ equals 0 when $A = C$, equals 1 when $(A, C) = (1, -1)$, and equals -1 when $(A, C) = (-1, 1)$,

$$(h_A - h_C)^2 = \frac{1 - A \cdot C}{2},$$

and therefore

$$G_\theta(A, C) = \gamma^2 \cdot \theta \cdot (1 - A \cdot C).$$

Its mean is

$$\mathbb{E}[G_\theta(A, C)] = \gamma^2 \cdot \theta,$$

because $\mathbb{E}[A \cdot C] = \mathbb{E}[A] \cdot \mathbb{E}[C] = 0$. For $\theta \in \mathcal{U}_{\gamma,M,r}$,

$$|\mathbb{E}[G_\theta(A, C)]| \leq \gamma^2 \left(\frac{M}{\gamma} + r \right) = \gamma \cdot M + \gamma^2 \cdot r.$$

Choosing $\gamma > 0$ small enough that

$$\gamma \cdot M + \gamma^2 \cdot r \leq \delta$$

gives the stated drift bound.

It remains to verify the Hoeffding projections and the mini-batch variance. Conditioning on $A = a$ gives

$$\mathbb{E}[G_\theta(a, C)] = \gamma^2 \cdot \theta - \gamma^2 \cdot \theta \cdot a \cdot \mathbb{E}[C] = \gamma^2 \cdot \theta = \mathbb{E}[G_\theta(A, C)].$$

The same calculation conditioning on $C = c$ gives

$$\mathbb{E}[G_\theta(A, c)] = \gamma^2 \cdot \theta - \gamma^2 \cdot \theta \cdot \mathbb{E}[A] \cdot c = \mathbb{E}[G_\theta(A, C)].$$

Thus both first-order Hoeffding projections vanish identically. The centered canonical term is

$$G_\theta(A, C) - \mathbb{E}[G_\theta(A, C)] = -\gamma^2 \cdot \theta \cdot A \cdot C.$$

Since $A \cdot C$ is Rademacher,

$$\mathrm{Var}(G_\theta(A, C) - \mathbb{E}[G_\theta(A, C)]) = \gamma^4 \cdot \theta^2.$$

For independent positive samples $A_1, \ldots, A_B$ and negative samples $C_1, \ldots, C_B$, the first-order projections are zero, so only the canonical second-order term remains in the two-sample Hoeffding variance split. Equivalently, the only nonzero covariances among the centered terms occur when both indices match. Hence

$$\mathrm{Var}\left(\frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B} G_\theta(A_i, C_j)\right) = \frac{B^2}{B^4} \cdot \gamma^4 \cdot \theta^2 = \frac{\gamma^4 \cdot \theta^2}{B^2}.$$

□

G.2 Projected Bad Components in Trainable Transformer Neighborhoods

The next two corollaries move from a single trainable scalar to neighborhoods in the trainable Transformer parameter space. In both statements, $\Psi_{\boldsymbol{\omega},\boldsymbol{v}}(A, C)$ denotes the scalar projected pairwise-gradient kernel

$$\Psi_{\boldsymbol{\omega},\boldsymbol{v}}(A, C) = \langle \nabla_{\boldsymbol{\omega}} \ell(m_{\boldsymbol{\omega}}(A, C)), \boldsymbol{v} \rangle .$$

Its components $\Psi^{(1,0)}_{\boldsymbol{\omega},\boldsymbol{v}}$, $\Psi^{(0,1)}_{\boldsymbol{\omega},\boldsymbol{v}}$, and $\Psi^{(1,1)}_{\boldsymbol{\omega},\boldsymbol{v}}$ are defined by the same two-sample Hoeffding projection formulas as in Theorem 3.

Corollary 15 (Full-parameter one-layer Transformer neighborhood with a projected bad component)**.** *Fix $M > 0$, a slack $s > 0$, a line radius $r > 0$, and tolerances*

$$\delta_{\mathrm{drift}} > 0, \qquad \varepsilon_1 > 0.$$

In the one-layer architecture of Proposition 39, let the trainable parameter vector $\boldsymbol{\omega}$ collect all query, key, value, output, and feed-forward parameters, together with the scaled scalar head coordinate θ in the parameterization

$$\boldsymbol{a}_\theta = (0, \gamma \cdot \theta, 0)^\top.$$

Let $\boldsymbol{v}_\theta$ be the unit direction corresponding to this scalar coordinate. Then there are $\gamma > 0$, a compact line segment

$$\mathcal{L}_{\gamma,M,s,r} = \left\{ \boldsymbol{\omega}(\theta) : \theta \in \left[\frac{M+s}{\gamma}, \frac{M+s}{\gamma} + r \right] \right\},$$

and an open neighborhood $\mathcal{U}_{\mathrm{tr}}$ of this segment in the corresponding trainable parameter space such that the following statements hold for the pairwise square loss.

1. *Severe misrankings persist:*

$$\inf_{\boldsymbol{\omega} \in \mathcal{U}_{\mathrm{tr}}} \mathbb{P}(m_{\boldsymbol{\omega}}(X^+, X^-) \le -M) \ge \frac{1}{4}.$$

2. *The projected deterministic drift is small:*

$$\sup_{\boldsymbol{\omega} \in \mathcal{U}_{\mathrm{tr}}} |\langle \nabla R_\ell(\boldsymbol{\omega}), \boldsymbol{v}_\theta \rangle| \le \delta_{\mathrm{drift}}.$$

3. *The leading projected stochastic component is small:*

$$\sup_{\boldsymbol{\omega} \in \mathcal{U}_{\mathrm{tr}}} \left\{ \mathrm{Var}(\Psi^{(1,0)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A)) + \mathrm{Var}(\Psi^{(0,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(C)) \right\} \le \varepsilon_1.$$

4. The residual pair-interaction component remains nonzero:

$$\inf_{\boldsymbol{\omega}\in\mathcal{U}_{\mathrm{tr}}} \mathrm{Var}(\Psi^{(1,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C)) > 0.$$

Thus a one-layer Transformer with all listed parameters trainable contains a neighborhood that is bad in the projected trainable direction $\boldsymbol{v}_\theta$. This is a projected-component statement for the constructed direction.

Proof. Use the base construction in Proposition 39, but replace M there by $M+s$. Along the compact segment $\mathcal{L}_{\gamma,M,s,r}$, the pair $(A,C)=(-1,1)$ has probability $1/4$ and margin

$$m_{\boldsymbol{\omega}(\theta)}(-1,1) = -\gamma\cdot\theta \le -(M+s).$$

All score values on the finite input family $\{s_{-1}, s_{+1}\}$ are smooth functions of the trainable parameter vector $\boldsymbol{\omega}$, because they are obtained by matrix multiplication and softmax. Therefore, for each point $\boldsymbol{\omega}(\theta)$ on the compact segment, there is a trainable-parameter neighborhood on which

$$m_{\boldsymbol{\omega}}(-1,1) \le -M.$$

A finite subcover of the compact segment gives an open neighborhood $\mathcal{U}_0$ of $\mathcal{L}_{\gamma,M,s,r}$ such that this inequality holds throughout $\mathcal{U}_0$. Hence

$$\mathbb{P}(m_{\boldsymbol{\omega}}(X^+,X^-)\le -M) \ge \mathbb{P}(A=-1, C=1) = \frac{1}{4}$$

for every $\boldsymbol{\omega}\in\mathcal{U}_0$.

Next consider the scalar projected pairwise-gradient kernel

$$\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C) = \langle \nabla_{\boldsymbol{\omega}}\ell(m_{\boldsymbol{\omega}}(A,C)), \boldsymbol{v}_\theta\rangle.$$

Define its positive-class, negative-class, and residual pair-interaction components by

$$\mu_{\boldsymbol{\omega}} = \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C)],$$

$$\Psi^{(1,0)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a) = \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a,C)] - \mu_{\boldsymbol{\omega}}, \qquad \Psi^{(0,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(c) = \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,c)] - \mu_{\boldsymbol{\omega}},$$

and

$$\Psi^{(1,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a,c) = \Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a,c) - \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a,C)] - \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,c)] + \mu_{\boldsymbol{\omega}}.$$

On the segment $\mathcal{L}_{\gamma,M,s,r}$, this is exactly the scalar kernel $G_\theta(A,C)$ in Proposition 39, with M replaced by $M+s$. Therefore

$$\mathbb{E}[\Psi_{\boldsymbol{\omega}(\theta),\boldsymbol{v}_\theta}(A,C)] = \gamma^2\cdot\theta,$$

and hence

$$\left|\mathbb{E}[\Psi_{\boldsymbol{\omega}(\theta),\boldsymbol{v}_\theta}(A,C)]\right| \le \gamma^2\left(\frac{M+s}{\gamma}+r\right) = \gamma\cdot(M+s)+\gamma^2\cdot r.$$

Choose $\gamma>0$ small enough that

$$\gamma\cdot(M+s)+\gamma^2\cdot r \le \frac{\delta_{\mathrm{drift}}}{2}.$$

The map

$$\boldsymbol{\omega}\mapsto \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C)]$$

is continuous on the finite input family, so after possibly shrinking $\mathcal{U}_0$ to a smaller open neighborhood $\mathcal{U}_1$ of the same compact segment,

$$\sup_{\boldsymbol{\omega}\in\mathcal{U}_1} |\langle \nabla R_\ell(\boldsymbol{\omega}), \boldsymbol{v}_\theta\rangle| \le \delta_{\mathrm{drift}}.$$

On the segment, Proposition 39 gives exact first-order projection collapse:

$$\mathrm{Var}(\Psi^{(1,0)}_{\boldsymbol{\omega}(\theta),\boldsymbol{v}_\theta}(A)) + \mathrm{Var}(\Psi^{(0,1)}_{\boldsymbol{\omega}(\theta),\boldsymbol{v}_\theta}(C)) = 0.$$

The two first-order projection variances are continuous functions of $\boldsymbol{\omega}$, again because the input family is finite and the score map is smooth. Thus, after shrinking to an open neighborhood $\mathcal{U}_2\subseteq\mathcal{U}_1$,

$$\sup_{\boldsymbol{\omega}\in\mathcal{U}_2}\left\{\mathrm{Var}(\Psi^{(1,0)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A)) + \mathrm{Var}(\Psi^{(0,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(C))\right\} \le \varepsilon_1.$$

Finally, on the segment the canonical variance is

$$\mathrm{Var}(\Psi^{(1,1)}_{\boldsymbol{\omega}(\theta),\boldsymbol{v}_\theta}(A,C)) = \gamma^4 \cdot \theta^2 \geq \gamma^2 \cdot (M+s)^2 > 0.$$

By continuity and compactness, after one last shrinking to an open neighborhood

$$\mathcal{U}_{\mathrm{tr}} \subseteq \mathcal{U}_2$$

of $\mathcal{L}_{\gamma,M,s,r}$, we have

$$\inf_{\boldsymbol{\omega}\in\mathcal{U}_{\mathrm{tr}}} \mathrm{Var}(\Psi^{(1,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C)) \geq \frac{1}{2} \cdot \gamma^2 \cdot (M+s)^2 > 0.$$

All four displayed properties now hold on $\mathcal{U}_{\mathrm{tr}}$, which proves the corollary. □

Corollary 16 (Two-layer and finite-depth Transformer embedding)**.** *There is a two-layer residual self-attention scorer whose pairwise square-loss dynamics satisfy all conclusions of Proposition 39 on the same interval $\mathcal{U}_{\gamma,M,r}$. More generally, for every finite integer $L \geq 1$, the same conclusions hold for some L-layer residual self-attention scorer.*

Proof. Use Proposition 39 as the first layer. For the second layer, choose the value map and feed-forward map to be zero while keeping the residual connections. The self-attention output and feed-forward output of the second layer are then zero, so the second residual block is the identity map on every token representation. In particular, the classification-token representation remains

$$\boldsymbol{r}_z = (1, \alpha_z \cdot z, 0)$$

for every $z \in \{-1, 1\}$. Attach the same trainable scalar head

$$\boldsymbol{a}_\theta = (0, \gamma \cdot \theta, 0)^\top.$$

The resulting two-layer scorer is therefore identical to the scorer in Proposition 39 on the input family $\{s_{-1}, s_{+1}\}$. The pairwise margins, square-loss gradient kernel, conditional expectations, Hoeffding projections, and mini-batch variance are identical as well. Hence all conclusions of Proposition 39 hold for the two-layer scorer.

The finite-depth statement is the same embedding repeated. For any $L > 2$, choose the value map and feed-forward map of every additional layer $\ell = 2, \ldots, L$ to be zero while keeping residual connections. Each added block is again the identity map, so the final scorer remains identical to the one-layer scorer on $\{s_{-1}, s_{+1}\}$. The same pairwise dynamics and variance identities therefore hold for every finite $L \geq 1$. □

Corollary 17 (Full-parameter two-layer Transformer neighborhood with a projected bad component)**.** *Fix $M > 0$, a slack $s > 0$, a line radius $r > 0$, and tolerances*

$$\delta_{\mathrm{drift}} > 0, \qquad \varepsilon_1 > 0.$$

Consider the two-layer residual self-attention architecture obtained by embedding the one-layer construction of Proposition 39 into the first layer and initializing the second layer as an identity residual block, as in Corollary 16. Let the trainable parameter vector $\boldsymbol{\omega}$ collect all query, key, value, output, and feed-forward parameters in both layers, together with the scaled scalar head coordinate θ in

$$\boldsymbol{a}_\theta = (0, \gamma \cdot \theta, 0)^\top.$$

Let $\boldsymbol{v}_\theta$ be the unit direction corresponding to this scalar coordinate. Then there are $\gamma > 0$, a compact line segment

$$\mathcal{L}^{(2)}_{\gamma,M,s,r} = \left\{ \boldsymbol{\omega}^{(2)}(\theta) : \theta \in \left[\frac{M+s}{\gamma}, \frac{M+s}{\gamma} + r \right] \right\},$$

and an open neighborhood $\mathcal{U}^{(2)}_{\mathrm{tr}}$ of this segment in the full two-layer trainable parameter space such that the following statements hold for the pairwise square loss.

1. *Severe misrankings persist:*

$$\inf_{\boldsymbol{\omega}\in\mathcal{U}^{(2)}_{\mathrm{tr}}} \mathbb{P}(m_{\boldsymbol{\omega}}(X^+, X^-) \leq -M) \geq \frac{1}{4}.$$

2. *The projected deterministic drift is small:*

$$\sup_{\boldsymbol{\omega}\in\mathcal{U}_{\mathrm{tr}}^{(2)}} |\langle \nabla R_\ell(\boldsymbol{\omega}), \boldsymbol{v}_\theta\rangle| \le \delta_{\mathrm{drift}}.$$

3. *The leading projected stochastic component is small:*

$$\sup_{\boldsymbol{\omega}\in\mathcal{U}_{\mathrm{tr}}^{(2)}} \left\{ \mathrm{Var}(\Psi^{(1,0)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A)) + \mathrm{Var}(\Psi^{(0,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(C)) \right\} \le \varepsilon_1.$$

4. *The residual pair-interaction component remains nonzero:*

$$\inf_{\boldsymbol{\omega}\in\mathcal{U}_{\mathrm{tr}}^{(2)}} \mathrm{Var}(\Psi^{(1,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C)) > 0.$$

Thus a two-layer Transformer with all listed parameters trainable contains a neighborhood that is bad in the projected trainable direction $\boldsymbol{v}_\theta$. This is a projected-component statement for the constructed direction.

Proof. Start from the one-layer construction in Proposition 39, replacing M there by $M+s$. Embed it as the first layer of a two-layer residual self-attention scorer. Initialize the second layer as in Corollary 16: its value map and feed-forward map are zero, and residual connections are kept. Along the line segment $\mathcal{L}^{(2)}_{\gamma,M,s,r}$, all first-layer and second-layer parameters are fixed at these values, and only the scalar head coordinate θ varies.

At every point of $\mathcal{L}^{(2)}_{\gamma,M,s,r}$, the second residual block is the identity map on the finite input family $\{s_{-1}, s_{+1}\}$. Therefore the two-layer score agrees exactly with the one-layer score:

$$f_{\boldsymbol{\omega}^{(2)}(\theta)}(s_z) = \gamma\cdot\theta\cdot\alpha_z\cdot z, \qquad z\in\{-1,1\}.$$

Consequently, for a positive value A and a negative value C,

$$m_{\boldsymbol{\omega}^{(2)}(\theta)}(A,C) = \gamma\cdot\theta\cdot(h_A - h_C), \qquad h_z = \alpha_z\cdot z.$$

On the event $(A,C)=(-1,1)$, which has probability $1/4$,

$$m_{\boldsymbol{\omega}^{(2)}(\theta)}(-1,1) = -\gamma\cdot\theta \le -(M+s)$$

for every

$$\theta \in \left[\frac{M+s}{\gamma}, \frac{M+s}{\gamma} + r\right].$$

All score values on $\{s_{-1}, s_{+1}\}$ are smooth functions of the full two-layer parameter vector $\boldsymbol{\omega}$: they are obtained by matrix multiplication, softmax, residual addition, and the final linear head. Since $\mathcal{L}^{(2)}_{\gamma,M,s,r}$ is compact, the strict slack $s>0$ implies that there is an open neighborhood $\mathcal{U}_0^{(2)}$ of the segment such that

$$m_{\boldsymbol{\omega}}(-1,1) \le -M \qquad \text{for all } \boldsymbol{\omega}\in\mathcal{U}_0^{(2)}.$$

Hence

$$\mathbb{P}(m_{\boldsymbol{\omega}}(X^+,X^-)\le -M) \ge \mathbb{P}(A=-1, C=1) = \frac{1}{4}$$

throughout $\mathcal{U}_0^{(2)}$.

Next define the scalar projected pairwise-gradient kernel

$$\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C) = \langle \nabla_{\boldsymbol{\omega}}\ell(m_{\boldsymbol{\omega}}(A,C)), \boldsymbol{v}_\theta\rangle.$$

Because $\boldsymbol{v}_\theta$ is the unit direction of the scalar head coordinate, on the segment this projected kernel is exactly the one-dimensional kernel from Proposition 39, with M replaced by $M+s$:

$$\Psi_{\boldsymbol{\omega}^{(2)}(\theta),\boldsymbol{v}_\theta}(A,C) = \gamma^2\cdot\theta\cdot(1-A\cdot C).$$

Therefore

$$\mathbb{E}[\Psi_{\boldsymbol{\omega}^{(2)}(\theta),\boldsymbol{v}_\theta}(A,C)] = \gamma^2\cdot\theta,$$

and for θ on the segment,

$$\left|\mathbb{E}[\Psi_{\boldsymbol{\omega}^{(2)}(\theta),\boldsymbol{v}_\theta}(A,C)]\right| \le \gamma^2\left(\frac{M+s}{\gamma}+r\right) = \gamma\cdot(M+s)+\gamma^2\cdot r.$$

Choose $\gamma>0$ so that

$$\gamma\cdot(M+s)+\gamma^2\cdot r \le \frac{\delta_{\mathrm{drift}}}{2}.$$

The map

$$\boldsymbol{\omega}\mapsto \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C)]$$

is continuous on the finite input family. Shrinking $\mathcal{U}_0^{(2)}$ if necessary gives an open neighborhood $\mathcal{U}_1^{(2)}$ of the segment such that

$$\sup_{\boldsymbol{\omega}\in\mathcal{U}_1^{(2)}} |\langle \nabla R_\ell(\boldsymbol{\omega}), \boldsymbol{v}_\theta\rangle| \le \delta_{\mathrm{drift}}.$$

For the Hoeffding projections, define

$$\mu_{\boldsymbol{\omega}} = \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C)],$$

$$\Psi^{(1,0)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a) = \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a,C)]-\mu_{\boldsymbol{\omega}}, \qquad \Psi^{(0,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(c) = \mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,c)]-\mu_{\boldsymbol{\omega}},$$

and

$$\Psi^{(1,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a,c) = \Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a,c)-\mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(a,C)]-\mathbb{E}[\Psi_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,c)]+\mu_{\boldsymbol{\omega}}.$$

On the line segment,

$$\mathbb{E}[\Psi_{\boldsymbol{\omega}^{(2)}(\theta),\boldsymbol{v}_\theta}(a,C)] = \mathbb{E}[\Psi_{\boldsymbol{\omega}^{(2)}(\theta),\boldsymbol{v}_\theta}(A,C)]$$

for each $a\in\{-1,1\}$, and the analogous identity holds when conditioning on $C=c$. Thus

$$\mathrm{Var}(\Psi^{(1,0)}_{\boldsymbol{\omega}^{(2)}(\theta),\boldsymbol{v}_\theta}(A)) + \mathrm{Var}(\Psi^{(0,1)}_{\boldsymbol{\omega}^{(2)}(\theta),\boldsymbol{v}_\theta}(C)) = 0.$$

The two projection variances are continuous functions of $\boldsymbol{\omega}$, again because the input family is finite and the score map is smooth. Hence there is an open neighborhood $\mathcal{U}_2^{(2)}\subseteq\mathcal{U}_1^{(2)}$ of the same segment such that

$$\sup_{\boldsymbol{\omega}\in\mathcal{U}_2^{(2)}} \left\{\mathrm{Var}(\Psi^{(1,0)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A)) + \mathrm{Var}(\Psi^{(0,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(C))\right\} \le \varepsilon_1.$$

Finally, on the segment the residual pair-interaction term is

$$\Psi^{(1,1)}_{\boldsymbol{\omega}^{(2)}(\theta),\boldsymbol{v}_\theta}(A,C) = -\gamma^2\cdot\theta\cdot A\cdot C,$$

so

$$\mathrm{Var}(\Psi^{(1,1)}_{\boldsymbol{\omega}^{(2)}(\theta),\boldsymbol{v}_\theta}(A,C)) = \gamma^4\cdot\theta^2 \ge \gamma^2\cdot(M+s)^2 > 0.$$

By continuity and compactness, after shrinking once more to an open neighborhood

$$\mathcal{U}_{\mathrm{tr}}^{(2)}\subseteq\mathcal{U}_2^{(2)}$$

of $\mathcal{L}^{(2)}_{\gamma,M,s,r}$, we obtain

$$\inf_{\boldsymbol{\omega}\in\mathcal{U}_{\mathrm{tr}}^{(2)}} \mathrm{Var}(\Psi^{(1,1)}_{\boldsymbol{\omega},\boldsymbol{v}_\theta}(A,C)) \ge \frac{1}{2}\cdot\gamma^2\cdot(M+s)^2 > 0.$$

The severe-misranking, drift, first-order projection, and residual-variance conclusions all hold on $\mathcal{U}_{\mathrm{tr}}^{(2)}$. The proof controls a full two-layer parameter neighborhood, while the bad component is the projected direction $\boldsymbol{v}_\theta$. □

H Localized Escape-Time Bounds for Stopped SGD

This section supplies the general probability tool used by Section 2. Formally, the escape time is the hitting time of the complement of a stopped region, or of a projected interval along a fixed direction. The main-text use is the projected interval along the optimistic steepest correction direction. The result is stated for a stopped SGD trajectory rather than for a specific loss, so that the escape-time argument remains separate from the AUC-specific Hoeffding calculation. Appendix D determines whether the local directional second moment scales as B^{-1} or B^{-2}; this section converts that scale into the corresponding projected escape-time lower bound used in the main text.

We first record two one-dimensional forms. They are useful when the relevant movement is projected onto a single direction $\boldsymbol{v}$. The subsequent proposition is the vector-valued local version used by the main text.

Proposition 40 (One-dimensional stopped local escape)**.** *Consider an adapted process*

$$S_{t+1} = S_t + \Delta_{t+1}, \qquad S_0 = 0,$$

with respect to a filtration $(\mathcal{F}_t)_{t\geq 0}$, and define

$$\tau_r = \inf\{t \geq 0 : \ |S_t| \geq r\}.$$

Assume that, on $\{t < \tau_r\}$,

$$\mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\Delta_{t+1}^2 \mid \mathcal{F}_t] \leq \nu_B^2 \qquad \text{a.s.}$$

Assume also that $\mathbb{E}[(S_{t\wedge\tau_r})^2] < \infty$ for every finite t, and that either $\mathbb{P}(\tau_r < \infty) = 1$ or $\mathbb{E}[\tau_r] = \infty$. Then

$$\nu_B = 0 \quad \Longrightarrow \quad \tau_r = \infty \text{ a.s.},$$

and otherwise

$$\mathbb{E}[\tau_r] \geq \frac{r^2}{\nu_B^2}.$$

Consequently, if $\nu_B^2 \leq C_1/B$, then $\mathbb{E}[\tau_r] \geq (r^2/C_1)\cdot B$; if $\nu_B^2 \leq C_2/B^2$, then $\mathbb{E}[\tau_r] \geq (r^2/C_2)\cdot B^2$.

Proof. If $\nu_B = 0$, then on $\{t < \tau_r\}$,

$$\mathbb{E}[\Delta_{t+1}^2 \mid \mathcal{F}_t] \leq 0,$$

so $\Delta_{t+1} = 0$ almost surely before exit. Since $S_0 = 0$, induction gives $S_t = 0$ for every t, and $\tau_r = \infty$ almost surely.

Assume $\nu_B > 0$. If $\mathbb{E}[\tau_r] = \infty$, the lower bound is immediate. It remains to consider the case $\tau_r < \infty$ almost surely and $\mathbb{E}[\tau_r] < \infty$. Define

$$M_t = S_{t\wedge\tau_r}.$$

Then

$$M_{t+1} = M_t + \Delta_{t+1}\cdot \mathbf{1}_{\{t<\tau_r\}}.$$

Because $\mathbf{1}_{\{t<\tau_r\}}$ is $\mathcal{F}_t$-measurable and $\mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t] = 0$ on $\{t < \tau_r\}$, (M_t) is a square-integrable martingale. Moreover,

$$\mathbb{E}[M_{t+1}^2 - M_t^2 \mid \mathcal{F}_t] = \mathbb{E}[\Delta_{t+1}^2 \cdot \mathbf{1}_{\{t<\tau_r\}} \mid \mathcal{F}_t] \leq \nu_B^2 \cdot \mathbf{1}_{\{t<\tau_r\}}.$$

Taking expectations and summing from 0 to $n-1$ gives

$$\mathbb{E}[M_n^2] \leq \nu_B^2 \cdot \mathbb{E}[n \wedge \tau_r].$$

On $\{\tau_r \leq n\}$, one has $|M_n| = |S_{\tau_r}| \geq r$, hence

$$r^2 \cdot \mathbb{P}(\tau_r \leq n) \leq \mathbb{E}[M_n^2] \leq \nu_B^2 \cdot \mathbb{E}[n \wedge \tau_r].$$

Letting $n \to \infty$, the probability converges to 1, and monotone convergence gives

$$r^2 \leq \nu_B^2 \cdot \mathbb{E}[\tau_r].$$

The two scaling statements follow by substituting the corresponding upper bounds on ν_B^2. □

Proposition 41 (Small-drift one-dimensional stopped local escape)**.** *Under the setup of Proposition 40, replace the zero-drift condition by*

$$|\mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t]| \le m_B, \qquad \mathbb{E}[\Delta_{t+1}^2 \mid \mathcal{F}_t] \le \nu_B^2 \qquad \text{a.s. on } \{t < \tau_r\}.$$

Then for every integer $T \ge 1$,

$$\big(r - m_B \cdot T\big)_+^2 \cdot \mathbb{P}(\tau_r \le T) \le \nu_B^2 \cdot T.$$

Proof. For $s \ge 1$, define

$$\mu_s = \mathbb{E}[\Delta_s \mid \mathcal{F}_{s-1}], \qquad \widetilde{\mu}_s = \mu_s \cdot \mathbf{1}_{\{s-1<\tau_r\}},$$

and

$$\widetilde{\xi}_s = (\Delta_s - \mu_s) \cdot \mathbf{1}_{\{s-1<\tau_r\}}.$$

Let

$$\widetilde{A}_t = \sum_{s=1}^{t} \widetilde{\mu}_s, \qquad \widetilde{M}_t = \sum_{s=1}^{t} \widetilde{\xi}_s.$$

The stopped process decomposes as

$$S_{t\wedge\tau_r} = \widetilde{A}_t + \widetilde{M}_t.$$

Since $\mathbf{1}_{\{s-1<\tau_r\}}$ is $\mathcal{F}_{s-1}$-measurable,

$$\mathbb{E}[\widetilde{\xi}_s \mid \mathcal{F}_{s-1}] = 0,$$

so $(\widetilde{M}_t)$ is a martingale. Its increments satisfy

$$\mathbb{E}[\widetilde{\xi}_s^2 \mid \mathcal{F}_{s-1}] = \mathbf{1}_{\{s-1<\tau_r\}} \mathrm{Var}(\Delta_s \mid \mathcal{F}_{s-1}) \le \nu_B^2.$$

Therefore

$$\mathbb{E}[\widetilde{M}_T^2] \le \nu_B^2 \cdot T.$$

Doob's L^2 maximal inequality gives, for every $a > 0$,

$$\mathbb{P}\left(\max_{0\le t\le T} |\widetilde{M}_t| \ge a\right) \le \frac{\nu_B^2 \cdot T}{a^2}.$$

The stopped drift satisfies

$$|\widetilde{A}_t| \le \sum_{s=1}^{t} m_B \cdot \mathbf{1}_{\{s-1<\tau_r\}} \le m_B \cdot t \le m_B \cdot T \qquad \text{for } t \le T.$$

If $\tau_r \le T$, then

$$r \le |S_{\tau_r}| \le |\widetilde{M}_{\tau_r}| + |\widetilde{A}_{\tau_r}| \le |\widetilde{M}_{\tau_r}| + m_B \cdot T.$$

Hence

$$\{\tau_r \le T\} \subseteq \left\{\max_{0\le t\le T} |\widetilde{M}_t| \ge (r - m_B \cdot T)_+\right\}.$$

If $r - m_B \cdot T \le 0$, the desired inequality is trivial. Otherwise apply Doob's inequality with $a = r - m_B \cdot T$ and multiply both sides by $(r - m_B \cdot T)^2$. □

Let $(\mathcal{F}_t)_{t\ge0}$ be a filtration and let $(\theta_t)_{t\ge0}$ be an adapted parameter trajectory in $\mathbb{R}^d$. Fix a measurable region $\mathcal{U} \subset \mathbb{R}^d$, an initial point $\theta_0 \in \mathcal{U}$, and define the first hitting time of the complement of $\mathcal{U}$ by

$$\tau_{\mathcal{U}} = \inf\{t \ge 0 : \theta_t \notin \mathcal{U}\}.$$

Assume that the Euclidean ball $\mathcal{B}(\theta_0, r) = \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\|_2 < r\}$ is contained in $\mathcal{U}$. Then leaving $\mathcal{U}$ before time T requires a displacement of at least r before time T. The next proposition turns this geometric fact into an escape-time bound.

Proposition 42 (Localized SGD escape-time lower bound)**.** *Let*

$$\Delta_{t+1} := \theta_{t+1} - \theta_t$$

and suppose that, on the event $\{t < \tau_{\mathcal{U}}\}$, the stopped increments satisfy

$$\|\mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t]\|_2 \le m_B, \qquad \mathbb{E}[\|\Delta_{t+1}\|_2^2 \mid \mathcal{F}_t] \le \nu_B^2 \qquad \text{a.s.}$$

for constants $m_B, \nu_B \ge 0$. Then, for every integer $T \ge 1$,

$$\big(r - m_B \cdot T\big)_+^2 \cdot \mathbb{P}(\tau_{\mathcal{U}} \le T) \le \nu_B^2 \cdot T.$$

If $\nu_B = 0$ and $m_B = 0$, then $\tau_{\mathcal{U}} = \infty$ almost surely. If $m_B \le 8 \cdot \nu_B^2/r$ and $0 < \nu_B^2 \le r^2/32$, then

$$\mathbb{E}[\tau_{\mathcal{U}}] \ge \frac{3 \cdot r^2}{128 \cdot \nu_B^2}.$$

In particular, the positive-variance zero-drift case $m_B = 0$ is included. Consequently, under the same small-drift and small-step conditions, if $\nu_B^2 \le C_1/B$ for a B-independent constant $C_1 > 0$, then for all sufficiently large B,

$$\mathbb{E}[\tau_{\mathcal{U}}] \ge \frac{3 \cdot r^2}{128 \cdot C_1} \cdot B,$$

whereas if $\nu_B^2 \le C_2/B^2$ for a B-independent constant $C_2 > 0$, then for all sufficiently large B,

$$\mathbb{E}[\tau_{\mathcal{U}}] \ge \frac{3 \cdot r^2}{128 \cdot C_2} \cdot B^2.$$

Proof. Define the stopped increment

$$\widetilde{\Delta}_{t+1} := \Delta_{t+1} \cdot \mathbf{1}_{\{t < \tau_{\mathcal{U}}\}},$$

and decompose it into its predictable and centered parts:

$$\boldsymbol{\mu}_{t+1} := \mathbb{E}[\widetilde{\Delta}_{t+1} \mid \mathcal{F}_t], \qquad \boldsymbol{\xi}_{t+1} := \widetilde{\Delta}_{t+1} - \boldsymbol{\mu}_{t+1}.$$

Let

$$\boldsymbol{A}_T := \sum_{t=0}^{T-1} \boldsymbol{\mu}_{t+1}, \qquad \boldsymbol{M}_T := \sum_{t=0}^{T-1} \boldsymbol{\xi}_{t+1}.$$

Then $(\boldsymbol{M}_T)_{T \ge 0}$ is a square-integrable martingale with respect to $(\mathcal{F}_T)$, because

$$\mathbb{E}[\boldsymbol{\xi}_{t+1} \mid \mathcal{F}_t] = \mathbf{0}.$$

Moreover, since $\mathbf{1}_{\{t < \tau_{\mathcal{U}}\}}$ is $\mathcal{F}_t$-measurable,

$$\|\boldsymbol{\mu}_{t+1}\|_2 = \big\|\mathbf{1}_{\{t < \tau_{\mathcal{U}}\}} \mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t]\big\|_2 \le m_B \cdot \mathbf{1}_{\{t < \tau_{\mathcal{U}}\}} \le m_B.$$

Therefore

$$\|\boldsymbol{A}_T\|_2 \le m_B \cdot T.$$

For the martingale term, use

$$\|\boldsymbol{\xi}_{t+1}\|_2^2 \le 2 \cdot \|\widetilde{\Delta}_{t+1}\|_2^2 + 2 \cdot \|\boldsymbol{\mu}_{t+1}\|_2^2.$$

This bound is sufficient for a finite-time small-drift inequality, but it introduces an unnecessary factor. A sharper route uses the Hilbert-space variance identity:

$$\mathbb{E}[\|\boldsymbol{\xi}_{t+1}\|_2^2 \mid \mathcal{F}_t] = \mathbb{E}[\|\widetilde{\Delta}_{t+1}\|_2^2 \mid \mathcal{F}_t] - \|\boldsymbol{\mu}_{t+1}\|_2^2 \le \mathbb{E}[\|\widetilde{\Delta}_{t+1}\|_2^2 \mid \mathcal{F}_t].$$

The stopped second moment is

$$\mathbb{E}[\|\widetilde{\Delta}_{t+1}\|_2^2 \mid \mathcal{F}_t] = \mathbf{1}_{\{t < \tau_{\mathcal{U}}\}} \mathbb{E}[\|\Delta_{t+1}\|_2^2 \mid \mathcal{F}_t] \le \nu_B^2 \cdot \mathbf{1}_{\{t < \tau_{\mathcal{U}}\}} \le \nu_B^2.$$

Martingale increments are orthogonal in L^2, so

$$\mathbb{E}[\|\boldsymbol{M}_T\|_2^2] = \sum_{t=0}^{T-1} \mathbb{E}[\|\boldsymbol{\xi}_{t+1}\|_2^2] \le \nu_B^2 \cdot T.$$

Thus Markov's inequality gives, for every $a > 0$,

$$\mathbb{P}(\|\boldsymbol{M}_T\|_2 \geq a) \leq \frac{\nu_B^2 \cdot T}{a^2}.$$

If $\tau_{\mathcal{U}} \leq T$, then because $\mathcal{B}(\theta_0, r) \subset \mathcal{U}$, the stopped trajectory must satisfy

$$\|\theta_{\tau_{\mathcal{U}}} - \theta_0\|_2 \geq r.$$

The stopped displacement equals

$$\theta_{\tau_{\mathcal{U}} \wedge T} - \theta_0 = \sum_{t=0}^{T-1} \widetilde{\Delta}_{t+1} = \boldsymbol{A}_T + \boldsymbol{M}_T.$$

Therefore, on $\{\tau_{\mathcal{U}} \leq T\}$,

$$r \leq \|\boldsymbol{A}_T + \boldsymbol{M}_T\|_2 \leq \|\boldsymbol{A}_T\|_2 + \|\boldsymbol{M}_T\|_2 \leq m_B \cdot T + \|\boldsymbol{M}_T\|_2.$$

Hence

$$\{\tau_{\mathcal{U}} \leq T\} \subseteq \left\{ \|\boldsymbol{M}_T\|_2 \geq \left(r - m_B \cdot T\right)_+ \right\}.$$

If $r - m_B \cdot T \leq 0$, the desired inequality is trivial. Otherwise applying the preceding Markov bound with $a = r - m_B \cdot T$ yields

$$\mathbb{P}(\tau_{\mathcal{U}} \leq T) \leq \frac{\nu_B^2 \cdot T}{(r - m_B \cdot T)^2},$$

which proves the finite-time bound.

If $\nu_B = 0$ and $m_B = 0$, the finite-time bound gives

$$r^2 \cdot \mathbb{P}(\tau_{\mathcal{U}} \leq T) \leq 0$$

for every $T \geq 1$. Hence $\mathbb{P}(\tau_{\mathcal{U}} \leq T) = 0$ for every finite T, and $\tau_{\mathcal{U}} = \infty$ almost surely.

It remains to prove the expected escape-time statement. Let

$$T_\star := \left\lfloor \frac{r^2}{16 \cdot \nu_B^2} \right\rfloor.$$

The condition $\nu_B^2 \leq r^2/32$ implies $r^2/(16 \cdot \nu_B^2) \geq 2$, and hence

$$T_\star \geq \frac{r^2}{32 \cdot \nu_B^2}.$$

If $m_B \leq 8 \cdot \nu_B^2/r$, then

$$m_B \cdot T_\star \leq \frac{8 \cdot \nu_B^2}{r} \cdot \frac{r^2}{16 \cdot \nu_B^2} = \frac{r}{2}.$$

Applying the finite-time bound at the integer time $T_\star$ gives

$$\mathbb{P}(\tau_{\mathcal{U}} \leq T_\star) \leq \frac{\nu_B^2 \cdot T_\star}{(r/2)^2} \leq \frac{\nu_B^2}{(r/2)^2} \cdot \frac{r^2}{16 \cdot \nu_B^2} = \frac{1}{4}.$$

Therefore

$$\mathbb{P}(\tau_{\mathcal{U}} > T_\star) \geq \frac{3}{4}.$$

Since $\tau_{\mathcal{U}}$ is nonnegative,

$$\mathbb{E}[\tau_{\mathcal{U}}] \geq T_\star \cdot \mathbb{P}(\tau_{\mathcal{U}} > T_\star) \geq \frac{3}{4} \cdot \frac{r^2}{32 \cdot \nu_B^2} = \frac{3 \cdot r^2}{128 \cdot \nu_B^2}.$$

The final two bounds follow by substituting $\nu_B^2 \leq C_1/B$ and $\nu_B^2 \leq C_2/B^2$, respectively. □

Corollary 18 (Projected escape-time consequence of first-order variance collapse)**.** *Fix a unit direction $\boldsymbol{v}$, and define the projected process*

$$S_t = \langle \theta_t - \theta_0, \boldsymbol{v} \rangle, \qquad \tau_{\boldsymbol{v},r} = \inf\{t \geq 0 : \ |S_t| \geq r\}.$$

If, before $\tau_{\boldsymbol{v},r}$, the projected stopped SGD increment satisfies the small-drift and small-step conditions in Proposition 42 and has conditional second moment of order B^{-1}, then the proposition gives an order-B lower-bound scale for $\mathbb{E}[\tau_{\boldsymbol{v},r}]$. If the first-order stochastic term vanishes in direction $\boldsymbol{v}$ and the projected conditional second moment is of order B^{-2}, then the lower-bound scale becomes order B^2.

Proof. Apply Proposition 42 to the one-dimensional trajectory S_t with region $(-r, r)$, first with $\nu_B^2 = O(B^{-1})$ and then with $\nu_B^2 = O(B^{-2})$. The AUC-specific reason these two cases arise is the two-sample Hoeffding variance split of the directional pairwise mini-batch gradient: nonzero first-order projections produce a B^{-1} term, while vanishing first-order projections leave only the B^{-2} canonical term. For a fixed step size, the centered variance contribution is multiplied by η^2. The corollary applies to the conditional second moment after the projected drift is separately controlled by the small-drift condition, so the order in B is unchanged. □

I Limitations

The theory studies projected escape along the steepest severe-error correction direction because this is an optimistic direction for escape. Among unit directions, it is the most direct local direction for reducing the severe-error functional, and therefore gives a favorable projected direction to test. Since there are infinitely many possible directions and they cannot all be enumerated, we focus on this steepest and most optimistic escape direction. If escape is slow even along this direction, escape along other directions may be even harder.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's theoretical mechanism, surrogate-weight claims, warm-up/pretraining interpretation, and experimental scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix I explains why the theory studies projected escape along the steepest severe-error correction direction as an optimistic direction among infinitely many possible directions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main text states the principal assumptions and proof sketches, and Appendices C–H provide the formal statements and proofs used by the main results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix B.3–B.5 gives task construction, split sizes, seeds, model settings, optimization hyperparameters, warm-up and margin grids, checkpoint selection, and environment details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the

case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Appendix B.4 lists public data sources and Appendix B.5 gives the experimental protocol, but this submission does not include an anonymized code package with exact reproduction commands.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix B.4 reports dataset construction and split sizes, and Appendix B.5 specifies models, optimizers, learning rates, batch sizes, schedules, warm-up and margin grids, checkpointing, and selection rules.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main and expanded appendix tables report mean held-out AUC with sample standard deviation, and Appendix B.3–B.5 defines the reported quantity, seed set, and checkpoint-selection protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B.5 reports the GPU type, memory, one-GPU-per-run scheduling, and a log-derived total compute estimate for all experiments involved in this paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Appendix B.8 states that the work studies optimization dynamics, uses existing public datasets and pretrained models, and does not involve new human-subject data collection or a deployed decision system.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix B.8 discusses potential positive impacts from improved understanding of AUC optimization and potential negative impacts from deployment in high-stakes domains without validation, calibration, subgroup analysis, or oversight.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Appendix B.8 states that the work does not release a new high-risk model, scraped dataset, or human-subject dataset. The paper reports experiments on existing public datasets and pretrained models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendix B.7 credits the public datasets, pretrained models, and standard software assets, explicitly lists the corresponding licenses or terms, and states that the manuscript does not redistribute third-party assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release a new dataset, model checkpoint, or benchmark asset; it reports experiments on existing public datasets and pretrained models, as stated in Appendix B.7.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Appendix B.8 states that the work does not involve new data collection, crowdsourcing, interaction with human subjects, or interventions on human participants.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Appendix B.8 states that the work does not involve crowdsourcing, human-subject interaction, interventions on human participants, or a new human-subject dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development does not involve LLMs as important, original, or non-standard components. Qwen2.5-1.5B-Instruct is used only as one experimental model family, as described in Appendix B.5.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.