How Smoothness Controls Overfitting: Sharp Curvature–Time Laws for Gradient Descent

Anonymous authors
Paper under double-blind review

Abstract

How does the population risk of gradient descent evolve as optimization continues on the same finite training sample? Classical stability bounds can overstate the statistical cost of longer training. We establish sharp curvature–time laws for projected gradient descent on the unit ball with convex, 1-Lipschitz, L-smooth losses for every $L \geq 1$. At unit smoothness, the worst-case expected excess risk is $\Theta(\min\{1, q^{-1} + n^{-1/2} + qn^{-3/2}\})$ for sample size n and cumulative step size q. Although the classical stability certificate is already constant-order at $q \asymp n$, the worst-case risk remains at the optimal $\Theta(n^{-1/2})$ scale and reaches constant order only at $q \asymp n^{3/2}$. The key is an interaction–dissipation principle that balances higher-order sample interactions against the stabilizing effect of losses unchanged by sample replacement. Beyond convexity, sharp risk laws show that even vanishing negative curvature can separate the last iterate from the full parameter average. On a single weakly nonconvex instance, the last iterate, a uniformly sampled iterate, and the full parameter average all have $O(n^{-1/2})$ risk at an earlier horizon, but at a common later horizon their risks are $\Theta(1)$, $\Theta((\log n)^{-1})$, and $\Theta((\log n)^{-2})$, respectively. Further applications yield sharp guarantees for weakly regularized pairwise learning, randomized coordinate descent, and fixed-source learning.

1 Introduction

Training time is a statistical choice, not merely an optimization budget. Continued optimization on a fixed dataset can improve empirical fit, yet its consequences for population risk are much less clear. Algorithmic stability relates generalization to sensitivity to replacing one training observation (Bousquet & Elisseeff, 2002; Hardt et al., 2016). However, a guarantee that worsens with training time does not establish that population risk itself has worsened. Once the guarantee becomes uninformative, it leaves open whether the learner has overfit or the analysis has simply lost precision. Resolving this distinction is necessary to understand how long a finite sample can support accurate learning.

How do training time and smoothness govern actual risk? We study projected gradient descent (PGD) on the Euclidean unit ball with globally convex, 1-Lipschitz, L-smooth losses. Convexity isolates the statistical question from nonconvex optimization effects, but does not ensure that fitting the training sample yields good population performance (Shalev-Shwartz et al., 2010; Feldman, 2016; Amir et al., 2021b). Smoothness controls how rapidly gradients can change, with smaller L imposing tighter curvature constraints. Let n be the sample size and $q = \eta T$ the cumulative step size for T updates with step size η. We determine the sharp worst-case expected excess risk for every $L \geq 1$ and both the last iterate and full parameter average in Theorem 2.1.

At unit smoothness, the risk is $\Theta(\min\{1, q^{-1} + n^{-1/2} + qn^{-3/2}\})$, combining optimization error q^{-1}, finite-sample statistical error $n^{-1/2}$, and the cost of sample reuse $qn^{-3/2}$. The risk reaches the optimal $\Theta(n^{-1/2})$ scale at $q \asymp n^{1/2}$, stays at this scale through $q \asymp n$, and then grows to constant order at $q \asymp n^{3/2}$. Thus the classical qn^{-1} certificate in Hardt et al. (2016) is already constant while the risk is still statistically optimal (Figure 1(a)). The full law determines how allowing larger curvature shortens the optimal-accuracy interval and accelerates later risk growth (Figure 1(b,c)). For our globally convex, Lipschitz loss class, this resolves the open problem posed by Livni (2024) on smooth GD's sample complexity when classical stability bounds cannot certify generalization.

What accounts for the gap between risk and the classical certificate? The classical qn^{-1} argument accumulates worst-case perturbations across updates. But replacing one observation leaves all other

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

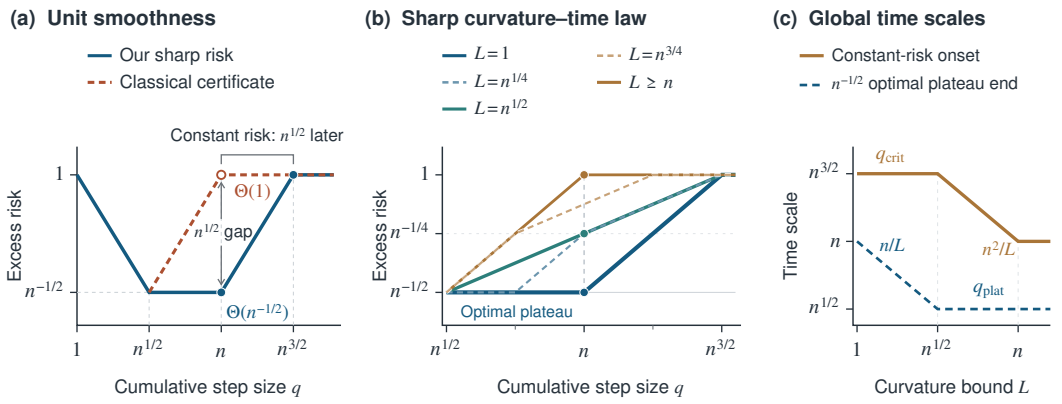

Figure 1: **Smoothness controls overfitting** (Theorem 2.1). (a) At $L = 1$, sharp risk remains $\Theta(n^{-1/2})$ when the classical certificate (Hardt et al., 2016) is constant at $q = n$, reaching constant order only at $q \asymp n^{3/2}$. (b) Larger L shortens the optimal-risk plateau and accelerates risk growth. (c) Optimal-plateau endpoint q_{plat} and constant-risk onset q_{crit} across L.

sample losses unchanged. These shared losses do more than prevent perturbation growth: smooth convexity supplies negative squared-gradient terms. Our *interaction–dissipation principle* pairs this dissipation with signed terms arising from joint dependence on multiple observations, before taking absolute values. Combined with optimization and smooth conversion from average squared sensitivity to excess risk (Lei & Ying, 2020; Nikolakakis et al., 2023), this sharply controls the cost of sample reuse. Matching lower-bound trajectories establish persistent risk deterioration on a single instance. In key regimes, the weighted-factor extension yields asymptotically strictly sharper bounds than the published guarantees for the same learners (Lei et al., 2020; Wu et al., 2025; Nikolakakis et al., 2023).

How much of this picture survives beyond convexity? The convex law gives the last iterate and the full parameter average the same worst-case risk order. Even a small amount of negative curvature can break this equivalence. Keeping global 1-Lipschitz continuity and unit smoothness, we allow ρ-weakly convex losses, whose negative curvature is bounded by ρ. Beyond existing stability and averaging guarantees (Richards & Rabbat, 2021; Neu & Rosasco, 2018; Wang et al., 2024), we establish sharp laws for the last iterate, an independent uniform checkpoint, and the full parameter average in Theorem 3.1. On one $O(n)$-dimensional instance with $\rho = 1/n$, the three outputs of the same trajectory all have $O(n^{-1/2})$ risk at $q \asymp n^{1/2}$. At $q = (n/4)\log n$, their risks become $\Theta(1)$, $\Theta((\log n)^{-1})$, and $\Theta((\log n)^{-2})$, respectively (Corollary 3.2). Both averaging and uniform selection use the full prefix. Thus parameter averaging can retain vanishing excess risk at a horizon when the last iterate incurs constant-order risk. Our contributions are as follows.

- **Sharp curvature–time laws under convexity.** We establish the sharp worst-case expected excess-risk law for PGD throughout the full (L, q) plane for every $L \geq 1$, covering both the last iterate and the full parameter average. Explicit constructions match the dimension-independent upper bounds, and the same law holds for constrained gradient flow. At unit smoothness, the risk remains $\Theta(n^{-1/2})$ when the classical certificate is already constant at $q \asymp n$, and reaches constant order only at $q \asymp n^{3/2}$, resolving an open question on smooth GD (Livni, 2024).
- **Sharp output laws under weak negative curvature.** For unit-smooth ρ-weakly convex losses, we establish sharp worst-case expected excess-risk laws for the last iterate, an independent uniform checkpoint, and the full parameter average. Even vanishing negative curvature can separate these outputs on a single trajectory: all have $O(n^{-1/2})$ risk at an earlier horizon, while at a common later horizon their risks become $\Theta(1)$, $\Theta((\log n)^{-1})$, and $\Theta((\log n)^{-2})$, respectively.
- **A general interaction–dissipation framework.** We develop a weighted-factor formulation extending interaction–dissipation to randomized coordinate updates and regularized minimization. It proves optimal $\Theta(n^{-1/2})$ risk for $1/n$-scale pairwise regularization, a sharp first-projection-adaptive pairwise law, and sharp source-dependent growth and saturation. In key regimes, these rates are asymptotically strictly *sharper* than the corresponding published upper bounds of Lei et al. (2020), Wu et al. (2025), and Nikolakakis et al. (2023).

2 CONVEX LOSSES: THE SHARP CURVATURE–TIME LAW

2.1 LEARNING SETUP AND OUTPUTS

For an i.i.d. training sample $S = (Z_1, \dots, Z_n) \sim P^n$, define the empirical risk $F_S = n^{-1}\sum_i f(\cdot, Z_i)$ and the population risk $F = \mathbb{E}_{Z\sim P} f(\cdot, Z)$ on a fresh observation. Unless indicated otherwise, $\mathbb{E}$ includes the sample, independent algorithmic randomness, and any independent replacement observations. Losses and gradients are finite and jointly measurable. Subtracting $f(0, z)$ ensures integrability without changing gradients or excess risk, but need not preserve nonnegativity. On the Euclidean unit ball $\mathbb{B}^d$, let $F^* = \min_{\mathbb{B}^d} F$ be the optimal population risk and run PGD with Euclidean projection $\Pi_{\mathbb{B}^d}$:

$$\boldsymbol{w}_0 = 0, \qquad \boldsymbol{w}_j = \Pi_{\mathbb{B}^d}(\boldsymbol{w}_{j-1} - \eta \nabla F_S(\boldsymbol{w}_{j-1})). \tag{1}$$

Let $q = \eta T$ be the cumulative step size. The outputs are the last point $\boldsymbol{w}_T$, the full post-update parameter average $\bar{\boldsymbol{w}}_T = T^{-1}\sum_{j=1}^{T} \boldsymbol{w}_j$, and, where specified, $\boldsymbol{w}_J$ with J uniform on $\{1, \dots, T\}$ and independent of S. Write $\widehat{\boldsymbol{w}}(S)$ for the chosen output, suppressing the independent random seed. Replacement runs use the same seed, including the same J. Constrained flow starts at zero and solves $\dot{\boldsymbol{w}} + \nabla F_S(\boldsymbol{w}) + N_{\mathbb{B}^d}(\boldsymbol{w}) \ni 0$, with normal cone $N_{\mathbb{B}^d}$. Its outputs are $\boldsymbol{w}(q)$, $q^{-1}\int_0^q \boldsymbol{w}(t)\, dt$, and an independent uniform-time checkpoint. Appendix C gives well-posedness and measurability.

2.2 THE COMPLETE CURVATURE–TIME LAW

For globally convex, 1-Lipschitz, L-smooth sample losses, let $\mathcal{R}_{n,L}(\widehat{\boldsymbol{w}})$ be the supremum of the expected excess risk $\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*]$ over instances in dimension at most d, with PGD parameters η, T and output rule $\widehat{\boldsymbol{w}} \in \{\boldsymbol{w}_T, \bar{\boldsymbol{w}}_T\}$ fixed. The output rule carries the dependence on η, T, and the dimension d is suppressed. There is no infimum over algorithms.

Theorem 2.1 (Complete curvature–time law)**.** *Let $q = \eta T$. There are universal $c, C, C_d > 0$ such that, for $n \geq 2$, $L \geq 1$, every integer $T \geq 1$ and every constant $0 < \eta \leq 1/L$, the upper bound below holds in every dimension. For $d \geq \lceil C_d n^3 \log^2(2n) \rceil$, both $\widehat{\boldsymbol{w}} = \boldsymbol{w}_T$ and $\widehat{\boldsymbol{w}} = \bar{\boldsymbol{w}}_T$ satisfy the matching law*

$$\mathcal{R}_{n,L}(\widehat{\boldsymbol{w}}) = \Theta\big(\min\{1, q^{-1} + n^{-1/2} + g_{n,L}(q)\}\big), \tag{2}$$

where

$$g_{n,L}(q) = \begin{cases} \min\left\{\dfrac{Lq}{n^{3/2}}, \dfrac{\sqrt{q}}{n^{3/4}}\right\}, & 1 \leq L \leq \sqrt{n}, \\ \min\left\{\dfrac{q}{n}, \dfrac{\sqrt{Lq}}{n}\right\}, & L \geq \sqrt{n}. \end{cases} \tag{3}$$

For $1 \leq L \leq \sqrt{n}$, $d \geq C_d n$ already suffices. The same bounds hold for constrained flow and its full time average, for every $q > 0$. All comparison constants are independent of the safe step.

Remark 2.2. **Optimal-risk plateau and constant-risk onset.** The unit-smooth case makes the training-time picture especially transparent. Writing $R_n(q)$ for the worst-case risk of either fixed output at $L = 1$, Theorem 2.1 gives

$$R_n(q) = \Theta\Big(\min\big\{1, \underbrace{q^{-1}}_{\text{optimization}} + \underbrace{n^{-1/2}}_{\text{statistical}} + \underbrace{q\, n^{-3/2}}_{\text{sample reuse}}\big\}\Big). \tag{4}$$

The three terms correspond to under-optimization at short times, the irreducible statistical scale, and the cost of repeatedly fitting the same finite sample. Training therefore passes through three regimes. It is optimization-limited before $q \asymp \sqrt{n}$, statistically optimal for $\sqrt{n} \lesssim q \lesssim n$, and increasingly limited by sample reuse thereafter. Constant-order risk appears only much later, at $q \asymp n^{3/2}$, as illustrated in Figure 1(a).

More generally, in the sufficient dimension of Theorem 2.1, the statistically optimal training window is $\sqrt{n} \lesssim q \lesssim n/\min\{L, \sqrt{n}\}$. Increasing curvature progressively shortens this plateau. For $1 \leq L \leq \sqrt{n}$, its endpoint moves from n to n/L, while for $L \geq \sqrt{n}$ the window contracts to $q \asymp \sqrt{n}$. Leaving the optimal plateau is nevertheless distinct from reaching constant-order risk. For flow or a fixed safe GD grid, let q_{crit} be the infimum over $q \geq n$ at which the worst-case risk reaches a fixed sufficiently small constant $r_0 > 0$. Then Theorem M.1 gives $q_{\text{crit}} \asymp \max\{n, n^2/(\sqrt{n} + L)\}$.

The two time scales can therefore be polynomially separated. At $L = 1$, optimal accuracy ends around $q \asymp n$, whereas constant-order risk appears only at $q \asymp n^{3/2}$. Figure 1(b,c) summarizes how curvature compresses the optimal-risk window while separately advancing the onset of constant-order risk. The restriction $q \geq n$ excludes large early optimization error and does not assume risk monotonicity.

Remark 2.3. **Why classical stability is loose: interaction–dissipation.** In Theorem 2.1, we improve on the classical q/n stability certificate by absorbing signed sample interactions before accumulating the remaining perturbation costs. For any measurable output $\widehat{\boldsymbol{w}}(S) \in \mathbb{B}^d$, let $D^2_{\widehat{\boldsymbol{w}}} = n^{-1}\sum_i \mathbb{E}\|\widehat{\boldsymbol{w}}(S) - \widehat{\boldsymbol{w}}(S^{(i)})\|^2$ denote its average squared replacement influence. Here $S^{(i)}$ replaces observation i by an independent copy, and the two runs share the output rule and data-independent random seed. For unit-smooth convex losses, fix a deterministic population minimizer $\boldsymbol{w}^*$ and suppose the same output satisfies $\mathbb{E}[F_S(\widehat{\boldsymbol{w}}) - F_S(\boldsymbol{w}^*)] \leq \delta$. Then

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] \leq 2\delta + 3n^{-1/2} + D^2_{\widehat{\boldsymbol{w}}}. \tag{5}$$

This smooth conversion builds on on-average and squared stability (Lei & Ying, 2020; Nikolakakis et al., 2023). For PGD with $0 < \eta \leq 1/L$, the projected descent inequality gives the budget $\delta = (2q)^{-1}$ for both the last iterate and the full parameter average.

The key gain is therefore sharper control of $D^2_{\widehat{\boldsymbol{w}}}$. Replacing one observation leaves the other sample losses unchanged. Smooth convexity supplies negative squared-gradient terms from these shared losses. Pairing them with signed sample interactions before taking absolute values absorbs higher-order dependence before the remaining costs accumulate (Lemma C.1). For the actual projected trajectory, both complete outputs satisfy

$$D^2_{\widehat{\boldsymbol{w}}} \leq q\big(4n^{-3/2} + 20Ln^{-2}\big) \lesssim q(n^{-3/2} + Ln^{-2}). \tag{6}$$

The same order holds for flow. At $L = 1$, combining (5)–(6) yields the $q/n^{3/2}$ sample-reuse cost in Theorem 2.1, rather than the classical q/n certificate. For general L, smooth and Lipschitz risk conversions, together with the classical stability bound, give the branches in $g_{n,L}(q)$ (Appendix C). Section 4 abstracts the same absorption principle to weighted factors and uses it to obtain the sharper application-specific laws.

From envelopes to trajectories. Theorem 2.1 characterizes the sharp worst-case risk envelope, but this pointwise envelope may be realized by different instances at different training horizons. It therefore does not by itself show that one fixed learning problem actually deteriorates as training continues. The next section closes this gap by constructing single instances whose population risk remains persistently large beyond a prescribed horizon, simultaneously for the last iterate, the full parameter average, every safe constant step, and constrained flow.

2.3 Matching overfitting trajectories

We next establish matching lower bounds on a single fixed instance across all later training horizons.

Theorem 2.4 (Matching overfitting lower bounds)**.** *There are universal $c, C, C_d > 0$ such that, for every $n \geq 2$, $L \geq 1$ and prescribed $q_0 \geq C\sqrt{n}$, there exist a distribution P and globally convex, 1-Lipschitz, L-smooth losses $f(\cdot, z)$, fixed before drawing $S \sim P^n$, with the following property. For every $0 < \eta \leq 1/L$ and integer $T \geq 1$ with $q = \eta T \geq q_0$, PGD (1) satisfies*

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] \geq c\min\{1, g_{n,L}(q_0)\}, \qquad \widehat{\boldsymbol{w}} \in \{\boldsymbol{w}_T, \bar{\boldsymbol{w}}_T\}. \tag{7}$$

The same instance satisfies this bound for constrained flow and its full time average at every $q \geq q_0$. The dimension is at most $C_d n$ for $1 \leq L \leq \sqrt{n}$, and at most $C_d n^3 \log^2(2n)$ for $L > \sqrt{n}$.

Remark 2.5. **A concrete persistent trajectory.** For all sufficiently large n, a fixed unit-smooth, $O(n)$-dimensional instance starts at a population minimizer, has expected excess risk $O(n^{-1/2})$ throughout $\sqrt{n} \leq q \leq n$, and has constant-order risk at every $q \geq Cn^{3/2}$. This holds for both outputs, every safe constant step, and constrained flow (Corollary M.6), showing persistent deterioration from statistically optimal risk. Initialization at a population minimizer does not imply samplewise realizability. Taking $q_0 = q \geq C\sqrt{n}$ in Theorem 2.4 recovers the growth term in Theorem 2.1, and affine witnesses supply the optimization and statistical terms (Appendix C.7).

Remark 2.6. **Low and high curvature.** At low curvature, a memory identifies unseen labels and a decoder converts them into population risk. A code-driving force b and target radius θ satisfy the margin constraint $b\theta \lesssim n^{-3/2}$ with travel time $q_0 \asymp \theta/b$. Choosing $\theta \asymp \min\{1, \sqrt{q_0/n^{3/2}}\}$ yields the readout $\min\{L\theta^2, \theta\}$, giving both low-curvature growth branches. At high curvature, a subset construction indexed by an integer $1 \le k \le n$ targets risk $r_k = \sqrt{k/n}$. It attains this risk once $q_0 \gtrsim \max\{nr_k, n^2 r_k^2/L\}$: the two costs are travel time and curvature-limited tracking time. Optimizing k gives both high-curvature growth branches.

3 Beyond convexity: output-dependent risk

For convex losses, the last iterate and the full parameter average have the same worst-case risk order. Even vanishing negative curvature can separate them. We retain global 1-Lipschitzness and unit smoothness, but assume only ρ-weak convexity, so $f(\cdot, z) + (\rho/2)\|\cdot\|^2$ is convex. The checkpoint index J is uniform on $\{1, \dots, T\}$, independent of the sample and shared across replacement runs. Existing replacement perturbations can be amplified by up to $1 + \eta\rho$ per step. We track their accumulated effect through

$$K_j^{(r)} = \eta \sum_{s=0}^{j-1} (1 + \eta\rho)^{rs} \quad (r = 1, 2), \qquad d_j = \min\{1, K_j^{(1)}/n, \sqrt{K_j^{(2)}/n^{3/2}}\}. \tag{8}$$

Here $K_j^{(1)}$ tracks replacement distance and $K_j^{(2)}$ squared influence. The three output rules differ in how they aggregate d_j: $V(\boldsymbol{w}_T) = d_T^2$, $V(\boldsymbol{w}_J) = T^{-1}\sum_{j=1}^T d_j^2$, and $V(\bar{\boldsymbol{w}}_T) = (T^{-1}\sum_{j=1}^T d_j)^2$, so $V(\bar{\boldsymbol{w}}_T) \le V(\boldsymbol{w}_J) \le V(\boldsymbol{w}_T)$. These definitions also apply at $\rho = 0$. As in the convex setting, let $\mathcal{R}^{\mathrm{wc}}_{n,\rho}(\widehat{\boldsymbol{w}})$ denote the worst-case expected excess risk for the fixed output rule $\widehat{\boldsymbol{w}}$ and parameters η, T over this weakly convex class, suppressing dimension.

Theorem 3.1 (Output laws under weak negative curvature)**.** *For $n \ge 2$, $0 \le \rho \le 1$, zero initialization, every $0 < \eta \le 1$ and integer $T \ge 1$, let $q = \eta T$. For each $\widehat{\boldsymbol{w}} \in \{\boldsymbol{w}_T, \boldsymbol{w}_J, \bar{\boldsymbol{w}}_T\}$, the upper bound holds in every dimension, and for $d \ge Cn$ the matching law is*

$$\mathcal{R}^{\mathrm{wc}}_{n,\rho}(\widehat{\boldsymbol{w}}) = \Theta\big(\min\{1, q^{-1} + n^{-1/2} + \rho + V(\widehat{\boldsymbol{w}})\}\big). \tag{9}$$

Constants are universal and independent of the safe step. The same risk and dimension bounds hold for constrained flow with $K^{(r)}(t) = \int_0^t e^{r\rho s}\, ds$, $r = 1, 2$, and normalized time averages on $[0, q]$.

All three outputs share the optimization, statistical, and curvature terms, but the last iterate retains the final response, the checkpoint averages squared responses, and the parameter average squares the mean response. At $\rho = 0$, all three laws recover the convex order. The next result realizes all three rates on one fixed instance and trajectory (Appendix I.8).

Corollary 3.2 (Common-instance output separation under vanishing curvature)**.** *For all sufficiently large n, let $\rho_n = 1/n$ and $q_n = (n/4)\log n$. There exists a fixed $O(n)$-dimensional ρ_n-weakly convex instance in this loss class, chosen before sampling, such that, for every $0 < \eta \le 1$, at $T = \lceil q_n/\eta \rceil$, the last iterate $\boldsymbol{w}_T$, an independent uniform checkpoint $\boldsymbol{w}_J$, and the full parameter average $\bar{\boldsymbol{w}}_T$ of the same PGD trajectory satisfy*

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] = \begin{cases} \Theta(1), & \widehat{\boldsymbol{w}} = \boldsymbol{w}_T, \\ \Theta((\log n)^{-1}), & \widehat{\boldsymbol{w}} = \boldsymbol{w}_J, \\ \Theta((\log n)^{-2}), & \widehat{\boldsymbol{w}} = \bar{\boldsymbol{w}}_T. \end{cases} \tag{10}$$

At the earlier horizon $T_0 = \lceil \sqrt{n}/\eta \rceil$, all three outputs have expected excess risk $O(n^{-1/2})$. The same early and late orders hold for constrained flow at times $\sqrt{n}$ and q_n, with an independent uniform-time checkpoint and the full time average. All constants are independent of the safe step.

The matching construction plants a small seed in a bounded code component and amplifies it along a common ray, with the exact growth factor $1 + \eta\rho$ before braking. This lets a single population readout realize all three output rates on the same trajectory. Appendix I gives the details.

4 A General Interaction–Dissipation Framework

In this section, we extend the interaction–dissipation framework to weighted objectives and develop a general influence-to-risk interface for the three applications below.

4.1 From sample interactions to excess risk

The goal of this section is to turn interaction–dissipation into a reusable interface for weighted objectives. Let $\mathcal{E}$ be a finite nonempty family of nonempty factors $e \subseteq [n]$, with fixed weights $\pi_e \geq 0$ and $\sum_e \pi_e = 1$. A factor loss $f_e(\boldsymbol{w}; Z_e)$ uses $Z_e = (Z_i)_{i \in e}$, such as $e = \{i\}$ for pointwise learning or $e = \{i, j\}$ for pairwise learning. For independent $Z_1, \ldots, Z_n$, define on a fixed closed convex domain $0 \in K \subseteq \mathbb{B}^d$

$$F_S(\boldsymbol{w}) = \sum_{e \in \mathcal{E}} \pi_e f_e(\boldsymbol{w}; Z_e), \qquad J_S = C_0 + \tau F_S, \qquad J = C_0 + \tau F, \qquad 0 < \tau \leq 1, \tag{11}$$

where $F(\boldsymbol{w}) = \mathbb{E}[F_S(\boldsymbol{w})]$ is population risk and C_0 is fixed, finite and convex. Losses are finite, jointly measurable, globally convex, 1-Lipschitz and 1-smooth, with integrable reference values.

The incidence weight $p_i = \sum_{e \ni i} \pi_e$ measures the total weight of factors using observation i, equaling $1/n$ for uniform singletons and $2/n$ for complete pairs. For an output $\widehat{\boldsymbol{w}}$, let $\mathcal{I}(\widehat{\boldsymbol{w}}) = \sum_i p_i \mathbb{E}\|\widehat{\boldsymbol{w}}(S) - \widehat{\boldsymbol{w}}(S^{(i)})\|^2$ be its replace-one squared influence. For uniform singletons this is $D^2_{\widehat{\boldsymbol{w}}}$. Replacement runs share the output rule and independent seed, including the coordinate sequence for randomized coordinate descent (RCD). The gain over classical stability comes from pairing signed sample interactions with dissipation from unchanged factors before taking absolute values.

Theorem 4.1 (Weighted-factor interaction–dissipation)**.** *For any measurable parameter $\boldsymbol{X} = \boldsymbol{X}(S, \omega)$ depending on the sample and an independent seed ω, with $\|\boldsymbol{X}\| \leq 1$ almost surely,*

$$\mathcal{C}(\boldsymbol{X}) - c\mathcal{D}(\boldsymbol{X}) \leq \sqrt{H_0} + \frac{H_1}{4c}, \qquad c > 0, \tag{12}$$

where $\mathcal{C}(\boldsymbol{X})$ is the signed sample interaction and $\mathcal{D}(\boldsymbol{X}) \geq 0$ is the dissipation from unchanged factors in (D.2), and H_0, H_1 in (D.3) depend only on factor incidences and weights. No trajectory or optimality condition on $\boldsymbol{X}$ is required. For uniform singletons and complete pairs, $\sqrt{H_0} + H_1 = \Theta(n^{-3/2})$, and the coefficient $\mathfrak{a}_\pi$ defined in Appendix D.1 has the same order. Appendix D gives the precise definitions, proof, and anchored refinement.

Corollary 4.2 (From absorption to influence and risk)**.** *Consider zero-initialized PGD and uniform scalar RCD with a fixed step $0 < \eta \leq 1$ and integer horizon $T \geq 1$, assuming C_0 is L_C-smooth with $L_C + \tau \leq 1$. RCD uses a product domain and i.i.d. uniform coordinates independent of the data. For ridge minimization, assume instead that C_0 is differentiable and let $\lambda > 0$. Then the last iterate or any deterministic convex parameter average of PGD or RCD, and the ridge minimizer $\boldsymbol{A}_\lambda = \arg\min_{\boldsymbol{w} \in K}\{J_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2/2\}$, satisfy*

$$\mathcal{I}(\widehat{\boldsymbol{w}}) \lesssim \tau \mathfrak{a}_\pi \begin{cases} q, & \text{PGD or uniform scalar RCD}, \\ \lambda^{-1}, & \widehat{\boldsymbol{w}} = \boldsymbol{A}_\lambda, \end{cases} \tag{13}$$

where $q = \eta T$ for PGD and $q = \eta T/D$ for D-dimensional scalar RCD. Moreover, for any measurable K-valued output $\widehat{\boldsymbol{w}}$, fix a deterministic $\boldsymbol{w}_\circ \in \arg\min_K J$ and set the gradient fluctuation $\sigma_0 = \|\nabla F_S(\boldsymbol{w}_\circ) - \nabla F(\boldsymbol{w}_\circ)\|_{L^2}$. If the same output satisfies $\mathbb{E}[J_S(\widehat{\boldsymbol{w}}) - J_S(\boldsymbol{w}_\circ)] \leq \delta$, then

$$\mathbb{E}[J(\widehat{\boldsymbol{w}}) - J(\boldsymbol{w}_\circ)] \leq 2\delta + 2\tau\sigma_0 + \tau\mathcal{I}(\widehat{\boldsymbol{w}}). \tag{14}$$

For i.i.d. uniform singleton losses with $C_0 = 0$ and $\tau = 1$, $\mathfrak{a}_\pi = \Theta(n^{-3/2})$ and $\sigma_0 \leq n^{-1/2}$, while PGD's last iterate and full parameter average satisfy $\delta \leq (2q)^{-1}$. Corollary 4.2 and the Lipschitz cap therefore recover the unit-smooth upper law of Theorem 2.1,

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] \lesssim \min\{1, q^{-1} + n^{-1/2} + q/n^{3/2}\}. \tag{15}$$

Section 4.2 uses ridge optimality with regularization bias in δ for pairwise learning, and the anchored refinement in Appendix D.3 for first-projection adaptation. Fixed-source learning applies the interface conditionally with $\tau = N/n$, where N counts the remaining conditionally i.i.d. observations, then recovers the original population risk (Appendix K).

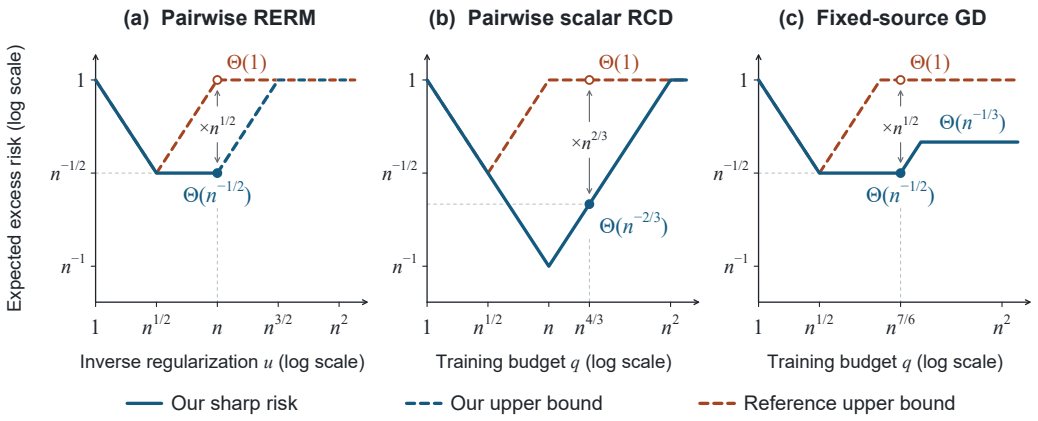

Figure 2: **Our sharp risks versus published upper-bound certificates.** Blue shows our sharp risk, with a dashed continuation in (a) where only an upper bound is known. Red shows the corresponding published certificate where applicable, capped by the trivial constant risk bound. **(a)** Pairwise regularized empirical risk minimization (RERM) with $u = c/\lambda$ and fixed $c \geq 8$, compared with Lei et al. (2020, Theorem 6) for α-smooth pair losses, $\alpha \leq 1$. This theorem applies when $n\lambda \geq 8\alpha$. Outside this range, the red continuation denotes only the trivial constant bound. **(b)** Scalar RCD with $q = \eta T/D$, $\eta = 1$, $D = 66n$, vanishing first projection ($\nu = 0$), and $f^* = \min_{\mathbb{R}^D} F = \Theta(1)$, compared with Wu et al. (2025, Theorem 3). **(c)** Last-iterate GD with $q = \eta T$, $\eta = 1/2$, and fixed Zipf source $p_j \propto j^{-3/2}$, compared with Nikolakakis et al. (2023, Theorems 10–11). Arrows indicate certificate-to-risk gaps. Appendix L verifies the required assumptions and inputs for these comparisons.

4.2 Applications

These applications assume globally nonnegative, convex, 1-Lipschitz, 1-smooth losses with integrable reference values. For symmetric pairs, empirical and population risks are $F_S = \binom{n}{2}^{-1} \sum_{i<j} f(\cdot; Z_i, Z_j)$ and $F = \mathbb{E} f(\cdot; Z, Z')$, for i.i.d. $Z, Z' \sim P$. Figure 2 compares our guarantees with published upper-bound certificates for the same learners.

Weak regularization. We first ask how weak regularization can be while retaining optimal statistical accuracy. For $\lambda > 0$, let $\boldsymbol{A}_\lambda = \arg\min_{\boldsymbol{w} \in \mathbb{B}^d} \{F_S(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|^2/2\}$.

Theorem 4.3 (Sharp law for weak pairwise regularization)**.** *For $n \geq 2$ and every fixed $c > 0$, uniformly over $\lambda \geq c/n$, the worst-case expected excess risk satisfies*

$$\sup_{P,f} \mathbb{E}[F(\boldsymbol{A}_\lambda) - F^*] = \Theta_c\big(\min\{1, \lambda + n^{-1/2}\}\big).$$

The upper bound holds in every dimension, with a matching lower bound in dimension one.

Remark 4.4. **Optimal weak regularization and comparison.** For $n^{-1} \lesssim \lambda \lesssim n^{-1/2}$, Theorem 4.3 gives $\Theta(n^{-1/2})$ worst-case excess risk without requiring a small population minimum or a low-noise condition. For smaller penalties, Corollary 4.2 still provides the dimension-independent upper bound

$$\mathbb{E}[F(\boldsymbol{A}_\lambda) - F^*] \lesssim \min\{1, \lambda + n^{-1/2} + (\lambda n^{3/2})^{-1}\}, \qquad \lambda > 0.$$

We do not claim a matching lower bound below the $\lambda \asymp n^{-1}$ scale. At $\lambda = c/n$ with fixed $c \geq 8$, our matching family has actual excess risk $\Theta(n^{-1/2})$, whereas the corresponding capped certificate of Lei et al. (2020, Theorem 6) remains $\Theta(1)$. Both guarantees concern the same estimator and regularization scale. The sharper rate follows from the KKT replacement response and interaction–dissipation in Corollary 4.2, with regularization bias $\delta \leq \lambda/2$. Figure 2(a) illustrates the gap, and Proposition L.1 proves the sharp law and verifies the comparison.

First-projection adaptation. We ask whether interaction–dissipation can exploit statistical cancellation in pairwise objectives that individual pair-gradient variance does not capture.

For a deterministic $\boldsymbol{w}_K^* \in \arg\min_K F$, let $F_K^* = \min_K F$ and define the centered first projection

$$\boldsymbol{h}_K(z) = \mathbb{E}_{Z'} \nabla f(\boldsymbol{w}_K^*; z, Z') - \nabla F(\boldsymbol{w}_K^*), \qquad \|\boldsymbol{h}_K(Z)\|_{L^2} \le \nu \le 1.$$

Theorem 4.5 (Sharp first-projection risk law)**.** *For $n \ge 2$, let $s_\nu = n^{-1} + \nu n^{-1/2}$. Run zero-initialized PGD on $K = \mathbb{B}^d$ or uniform scalar RCD on a closed coordinate box $0 \in K \subseteq \mathbb{B}^D$. For every $0 < \eta \le 1$, integer $T \ge 1$, and either the last iterate or full parameter average,*

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F_K^*] \lesssim \min\{1, q^{-1} + s_\nu + q s_\nu / n\}. \tag{16}$$

Here $q = \eta T$ for PGD and $q = \eta T / D$ for D-dimensional scalar RCD. The upper bound is dimension independent, with universal constants. For $n \ge 64$, matching worst-case lower bounds hold for RCD on the fixed box $\mathcal{K}_n \subseteq \mathbb{B}^{66n}$ in (J.2), and separately for PGD on $\mathbb{B}^{66n}$.

Remark 4.6. **Adaptation and separation from existing bounds.** Theorem 4.5 gives a three-phase picture of worst-case risk. It reaches the statistical scale s_ν around $q \asymp s_\nu^{-1}$, remains at this order through $q \asymp n$, and then grows to constant order around $q \asymp n / s_\nu$. At $\nu = 0$, this gives $\Theta(n^{-1})$ risk near $q = n$, with constant-order risk delayed until $q \asymp n^2$. A small first projection requires neither small pair-gradient variance nor small minimum loss. For i.i.d. symmetric signs, the globally Huberized loss agreeing with $(w - zz')^2/4$ on $[-1, 1]$ has $\nu = 0$, while both the pair-gradient variance and population minimum equal $1/4$ at its optimum (Proposition J.1).

For unprojected scalar RCD with $\eta = 1$, $D = 66n$, $\nu = 0$, and $q \asymp n^{4/3}$, the canonical family in Proposition L.3 has last-iterate risk $\Theta(n^{-2/3})$, whereas the corresponding capped certificate of Wu et al. (2025, Theorem 3) remains constant-order. Holding the learner, horizon, anchor, and population-risk conversion fixed, omitting interaction absorption instead gives a bound of order $\Theta(n^{-1/3})$ in this regime (Lemma L.4). This comparison isolates the gain from absorption within the same analysis. The proof uses anchored refinement (Theorem D.6), with matching PGD and RCD trajectories in Appendix J.

Fixed-source growth and saturation. We return to pointwise losses on a countable observation space, with $P(Z = j) = p_j$ and $F(\boldsymbol{w}) = \sum_j p_j f(\boldsymbol{w}, j)$. The source $p = (p_j)$ is fixed across sample sizes and unknown to the learner. Let $K_n = \big|\{j : p_j \ge 1/n\}\big|$ count observation atoms with probability at least $1/n$, and assume, for a slowly varying function ℓ,

$$\big|\{j : p_j \ge 1/u\}\big| \sim u^\alpha \ell(u), \qquad 0 < \alpha < 1, \qquad u \to \infty.$$

Theorem 4.7 (Sharp fixed-source growth and saturation law)**.** *For every fixed source p satisfying this condition, there exist $n_0(p)$ and $C_p > 0$ such that, for $n \ge n_0(p)$ and $d \ge C_p K_n$, the worst-case expected excess risk over loss families with p fixed satisfies*

$$\mathcal{R}_p(n, t) = \Theta_p\left(\min\left\{1, t^{-1} + n^{-1/2} + \min\left\{\frac{K_n}{n}, \frac{t\sqrt{K_n}}{n^2}\right\}\right\}\right). \tag{17}$$

This holds for zero-initialized unit-ball PGD, for either the last iterate or full parameter average, with every $0 < \eta \le 1$ and integer $T \ge 1$, and for ridge minimization over the same ball with $\lambda > 0$. Here $t = \eta T$ for PGD and $t = \lambda^{-1}$ for ridge. The upper bound holds in every dimension, and constants depend only on p. The loss family may depend on predetermined n, t before sampling.

Remark 4.8. **Growth, saturation, and comparison.** Theorem 4.7 shows how the source profile sets both the speed of risk growth and its saturation level without being known to the learner. The source-dependent contribution grows as $t\sqrt{K_n}/n^2$ and saturates at K_n/n. For the Zipf source $p_j \propto j^{-3/2}$, we have $K_n \asymp n^{2/3}$, so worst-case risk leaves the optimal $n^{-1/2}$ scale around $t \asymp n^{7/6}$ and saturates at $n^{-1/3}$ around $t \asymp n^{4/3}$. Thus prolonged training can sharply worsen statistical accuracy while excess risk still vanishes as $n \to \infty$. More generally, Section K.12 proves consistency for every fixed countable source whenever $t = t_n \to \infty$.

For $K_n \gtrsim_p \sqrt{n}$, the matching last-iterate family at $q_* \asymp_p n^{3/2}/\sqrt{K_n}$, run by unprojected GD with $\eta = 1/2$, has excess risk $\Theta_p(n^{-1/2})$, whereas the capped certificates of Nikolakakis et al. (2023, Theorems 10–11) remain constant-order on the same trajectory (Proposition L.5). To obtain the fixed-source law, we condition on the observed source structure, retain the actual factor coefficient $\tau = N/n$, and reassemble conditional residual bounds into the original population risk. Both learners use the full sample throughout. Figure 2(c) illustrates the gap, and Appendices K and L give the full analysis and comparison.

5 Related Work

We review related work on stability, overfitting, averaging, and structured learning, highlighting how our sharp risk laws differ from existing guarantees.

Stability and generalization. Stability connects sample sensitivity to generalization (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010). For gradient methods, the smooth-update analysis of Hardt et al. (2016) led to refinements using data and optimization geometry (Kuzborskij & Lampert, 2018; Charles & Papailiopoulos, 2018), on-average model sensitivity (Lei & Ying, 2020), and trajectory energy (Nikolakakis et al., 2023). These guarantees do not by themselves determine when population risk deteriorates. At unit smoothness, our sharp risk remains $\Theta(n^{-1/2})$ at $q \asymp n$, when the classical qn^{-1} certificate is already constant, and reaches constant order only at $q \asymp n^{3/2}$. Interaction–dissipation captures this gap by matching signed sample interactions with shared-loss dissipation before accumulating the remaining costs.

Overfitting of gradient methods. Dimension-independent learnability can coexist with poor ERM generalization (Shalev-Shwartz et al., 2010; Feldman, 2016). Separations between stochastic and full-batch methods (Amir et al., 2021b;a) motivated studies of GD's dependence on sample size, dimension, step size and horizon (Livni, 2024; Schliserman et al., 2025; Zhang et al., 2025; Burla & Livni, 2026). Our contribution is a matching law jointly in smoothness and training time. For globally convex, globally Lipschitz, L-smooth losses, we determine worst-case excess risk throughout the (L, q) plane for last and full-average PGD outputs and constrained flow on the unit ball in Theorem 2.1. At constant smoothness it resolves Livni (2024, Open Question 6) for this loss class. Appendix L.7 compares the closest constructions under their respective assumptions.

Iterate averaging and weak convexity. Geometric averaging acts as regularization in least-squares SGD (Neu & Rosasco, 2018). Stability also yields guarantees under weak convexity (Richards & Rabbat, 2021), for nonconvex population stationarity (Lei, 2023), and for stochastic weight averaging (Wang et al., 2024). These results do not establish matching excess-risk laws separating three outputs of one trajectory. We characterize the last iterate, an independent uniform checkpoint and the full parameter average in Theorem 3.1. Under vanishing negative curvature, Corollary 3.2 realizes their later risks $\Theta(1)$, $\Theta((\log n)^{-1})$ and $\Theta((\log n)^{-2})$ simultaneously on one instance.

Pairwise learning and structured objectives. Pairwise generalization has been studied through regularized estimation (Lei et al., 2018), refined stability (Lei et al., 2020), SGD (Yang et al., 2021a; Lei et al., 2021a) and coordinate descent (Wu et al., 2025). Our weighted-factor framework retains the overlap structure when pairing interactions with dissipation. It yields the sharp pairwise RERM, RCD and fixed-source laws in Section 4.2. Figure 2 and Appendix L compare specified published certificates for the same learners and regimes. Appendix B develops these connections further.

6 Conclusion

In this work, we characterize the statistical effect of continued optimization on a fixed sample for smooth convex and weakly convex losses. For smooth convex losses, we establish sharp curvature–time laws that identify both the endpoint of the statistically optimal training regime and the later onset of constant-order worst-case excess risk. Population risk can therefore remain statistically optimal well beyond the regime certified by classical stability bounds. Our interaction–dissipation framework explains this separation by coupling higher-order sample interactions with dissipation from losses unaffected by sample replacement. Matching fixed-instance constructions show that the resulting deterioration can persist along a single optimization trajectory.

Beyond convexity, we show that vanishing negative curvature can separate the last iterate, an independent uniform checkpoint, and the full parameter average. The framework also yields sharp results for weak pairwise regularization, first-projection adaptation, randomized coordinate updates, and heterogeneous source frequencies. Our upper bounds are dimension independent, whereas in the high-curvature regime the current matching constructions for the fastest risk growth require dimension polynomially larger than the sample size.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, presentation refinement, figure preparation, and LaTeX preparation. The research questions, mathematical claims, constructions, theorem statements, proof strategies, and proofs originated with and were developed by the authors. Generative AI tools were also used as auxiliary verification tools to stress-test proofs already developed by the authors and to help identify potential gaps, hidden assumptions, edge cases, or inconsistencies. All mathematical claims and proofs were independently verified, revised where necessary, and finalized by the authors. The authors take full responsibility for the correctness and content of the final manuscript.

ETHICS STATEMENT

This theoretical study uses no personal data or human subjects and makes no empirical claims about deployed systems.

REPRODUCIBILITY STATEMENT

The assumptions and algorithmic conventions are stated in Section 2 and alongside the corresponding results. Complete proofs and explicit lower-bound constructions are provided in the appendices, whose opening roadmap links the main results to their proofs.

REFERENCES

Idan Amir, Yair Carmon, Tomer Koren, and Roi Livni. Never go full batch (in stochastic convex optimization). In *Advances in Neural Information Processing Systems*, volume 34, pp. 25033–25043, 2021a.

Idan Amir, Tomer Koren, and Roi Livni. SGD generalizes better than GD (and regularization doesn't help). In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 63–92, 2021b.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/hash/7fe1f8abaad094e0b5cb1b01d712f708-Abstract.html.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4381–4391, 2020.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2nd edition, 2017. doi: 10.1007/978-3-319-48311-5.

Anna Ben-Hamou, Stéphane Boucheron, and Mesrob I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1): 249–287, 2017. doi: 10.3150/15-BEJ743. URL https://arxiv.org/abs/1412.8652.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 610–626, 2020. URL https://proceedings.mlr.press/v125/bousquet20b.html.

Tal Burla and Roi Livni. All ERMs can fail in stochastic convex optimization: Lower bounds in linear dimension. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.08350.

Daniel Carmon, Roi Livni, and Amir Yehudayoff. The sample complexity of ERMs in stochastic convex optimization. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3799–3807, 2024. URL https://proceedings.mlr.press/v238/carmon24a.html.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 745–754, 2018. URL https://proceedings.mlr.press/v80/charles18a.html.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008. doi: 10.1214/009052607000000910.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. doi: 10.1137/18M1178244.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016. doi: 10.1214/15-AOS1391.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3): 586–596, 1981.

André Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79, 2005. URL https://www.jmlr.org/papers/v6/elisseeff05a.html.

Vitaly Feldman. Generalization of ERM in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/05a624166c8eb8273b8464e8d9cb5bd9-Abstract.html.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1270–1279, 2019. URL https://proceedings.mlr.press/v99/feldman19a.html.

Alexander Gnedin, Ben Hansen, and Jim Pitman. Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probability Surveys*, 4:146–171, 2007. doi: 10.1214/07-PS092.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234, 2016.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Ilja Kuzborskij and Christoph H. Lampert. Data-dependent stability of stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2815–2824, 2018. URL https://proceedings.mlr.press/v80/kuzborskij18a.html.

Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 191–227, 2023. URL https://proceedings.mlr.press/v195/lei23a.html.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5809–5819, 2020.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=r28GdiQF7vM.

Yunwen Lei, Shao-Bo Lin, and Ke Tang. Generalization bounds for regularized pairwise learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 2376–2382, 2018. doi: 10.24963/ijcai.2018/329. URL https://www.ijcai.org/proceedings/2018/329.

Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21236–21246, 2020. URL https://leiyw.github.io/paper/NeurIPS2020.pdf.

Yunwen Lei, Mingrui Liu, and Yiming Ying. Generalization guarantee of SGD for pairwise learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21216–21228, 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/b1301141feffabac455e1f90a7de2054-Abstract.html.

Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6175–6186, 2021b. URL https://proceedings.mlr.press/v139/lei21b.html.

Roi Livni. The sample complexity of gradient descent in stochastic convex optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 64215–64241, 2024.

David McAllester and Luis Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4:895–911, 2003. URL https://www.jmlr.org/papers/v4/mcallester03a.html.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006. doi: 10.1007/s10444-004-7634-z.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 3222–3242, 2018. URL https://proceedings.mlr.press/v75/neu18a.html.

Konstantinos E. Nikolakakis, Farzin Haddadpour, Amin Karbasi, and Dionysios S. Kalogerias. Beyond Lipschitz: Sharp generalization and excess risk bounds for full-batch GD. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=pOyi9KqE56b.

Konstantinos E. Nikolakakis, Amin Karbasi, and Dionysis Kalogerias. Select without fear: Almost all minibatch schedules generalize optimally. *SIAM Journal on Mathematics of Data Science*, 7(3): 965–992, 2025. doi: 10.1137/23M1617096.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 127–239, 2014.

Sejun Park, Umut Şimşekli, and Murat A. Erdogdu. Generalization bounds for stochastic gradient descent via localized ε-covers. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2209.08951.

Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability results in learning theory. *Analysis and Applications*, 3(4):397–417, 2005. doi: 10.1142/S0219530505000650. URL https://www.mit.edu/~rakhlin/papers/stability.pdf.

Dominic Richards and Mike Rabbat. Learning with gradient descent and weakly convex losses. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1990–1998, 2021. URL https://proceedings.mlr.press/v130/richards21a.html.

Michael Eli Sander, Joan Puigcerver, Josip Djolonga, Gabriel Peyré, and Mathieu Blondel. Fast, differentiable and sparse top-k: A convex analysis perspective. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29919–29936, 2023.

Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3380–3394, 2022. URL https://proceedings.mlr.press/v178/schliserman22a.html.

Matan Schliserman, Uri Sherman, and Tomer Koren. The dimension strikes back with gradients: Generalization of gradient methods in stochastic convex optimization. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 1041–1107, 2025.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

Wei Shen, Zhenhuan Yang, Yiming Ying, and Xiaoming Yuan. Stability and optimization error of stochastic gradient descent for pairwise learning. *Analysis and Applications*, 18(5):887–927, 2020. doi: 10.1142/S0219530519400062.

Daniel A. Spielman. Linear-time encodable and decodable error-correcting codes. *IEEE Transactions on Information Theory*, 42(6):1723–1731, 1996. doi: 10.1109/18.556668.

Shira Vansover-Hager, Tomer Koren, and Roi Livni. Rapid overfitting of multi-pass SGD in stochastic convex optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60905–60923, 2025.

Peng Wang, Li Shen, Zerui Tao, Shuaida He, and Dacheng Tao. Generalization analysis of stochastic weight averaging with general sampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51442–51464, 2024. URL https://proceedings.mlr.press/v235/wang24bl.html.

Liang Wu, Ruixi Hu, and Yunwen Lei. Stability-based generalization analysis of randomized coordinate descent for pairwise learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21545–21553, 2025. doi: 10.1609/aaai.v39i20.35457. URL https://arxiv.org/abs/2503.01530.

Zhenhuan Yang, Yunwen Lei, Siwei Lyu, and Yiming Ying. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2026–2034, 2021a. URL https://proceedings.mlr.press/v130/yang21c.html.

Zhenhuan Yang, Yunwen Lei, Puyu Wang, Tianbao Yang, and Yiming Ying. Simple stochastic and online gradient descent algorithms for pairwise learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20160–20171, 2021b. URL https://proceedings.neurips.cc/paper/2021/hash/a87d27f712df362cd22c7a8ef823e987-Abstract.html.

Shuang Zeng and Yunwen Lei. Stochastic gradient methods: Bias, stability and generalization. *Journal of Machine Learning Research*, 27(6):1–55, 2026. URL https://www.jmlr.org/papers/v27/24-0637.html.

Peiyuan Zhang, Jiaye Teng, and Jingzhao Zhang. Generalization lower bounds for GD and SGD in smooth stochastic convex optimization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 334–342, 2025. URL https://proceedings.mlr.press/v258/zhang25a.html.

Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pingali, Chao Chen, and Mayank Goswami. Stability of SGD: Tightness analysis and improved bounds. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 2364–2373, 2022. URL https://proceedings.mlr.press/v180/zhang22b.html.

Contents of Appendix

A Notation Reference **18**

B Extended Related Work **20**

C Shared response and risk bounds **21**

C.1 Well-posedness, measurable outputs, and normalization 21

C.2 Cocoercivity and optimization of the updated average 22

C.3 Product-space identities and the flow budget 22

C.4 The complete projected-step argument 24

C.5 Population-centered risk conversion 24

C.6 Closing the upper rates and threshold converses 25

C.7 Elementary matching optimization and statistical terms 26

C.8 Last iterates and the classical uniform bound 26

D A general interaction–dissipation theorem **27**

D.1 Independent atoms, weighted factors, and replacement operators 27

D.2 The algebraic inequality 28

D.3 Stability bounds for projected and coordinate updates 29

D.4 Stability of regularized minimizers 30

D.5 From stability to population risk 31

D.6 The first-projection-adaptive pairwise upper bound 32

E Efficient smooth lower bounds in linear dimension **34**

E.1 A constructive code family 34

E.2 A convex mask with a one-code proximal oracle 34

E.3 A smooth readout with a global numerical oracle 36

E.4 The fixed losses and their global parameters 37

E.5 Repeated samples and finite identification 37

E.6 Every safe projected-GD step and the full average 38

E.7 Matching last-iterate lower bounds 39

E.8 Implementation of the lower-bound construction 39

F Fast overfitting in polynomial dimension **41**

F.1 Multiscale subset geometry 41

F.2 Simultaneous finite certificates and dimension reduction 42

F.3 A calibrated readout with a nonvanishing margin 43

F.4 A common initial point, the exact trajectory, and expected risk 44

F.5 Upper bounds and the complete nonsmooth formula 45

G Uniform smoothing and the curvature–time threshold **46**

G.1 Envelope facts and an oracle-uniform comparison 46

G.2 A uniform-in-time radial perturbation bound . . . 47
G.3 The exact time-n smoothness threshold . . . 48
G.4 Nonnegative penalties that vanish on a reference path . . . 48
G.5 Completion of the constant-risk threshold . . . 49

H Closing the full curvature–time law **49**
H.1 The high-curvature upper bound . . . 49
H.2 A truncated subset geometry . . . 49
H.3 Polar separation and the small averaged norm . . . 50
H.4 Finite compression, fixed before sampling . . . 51
H.5 The encoder and readout losses . . . 51
H.6 The reference ray and a global population certificate . . . 52
H.7 Samplewise smoothing and all safe discrete steps . . . 53
H.8 Integer degree optimization and completion . . . 54
H.9 Proof of Theorem 2.4 . . . 55

I Proof of the risk law under weak negative curvature **55**
I.1 Assumptions and exact response kernels . . . 56
I.2 Closed forms and the two output laws . . . 57
I.3 Upper bound: retain both responses, then average in norm . . . 58
I.4 A scale-free readout that survives complete averaging . . . 60
I.5 The seed-and-release lower construction . . . 61
I.6 The discrete clock . . . 65
I.7 Uniform checkpoint selection and completion of the main theorem . . . 65
I.8 Proof of Corollary 3.2 . . . 66

J Scalar-coordinate lower bounds for pairwise learning **67**
J.1 A learning interpretation of the first projection . . . 67
J.2 Algorithm, class, and quantifiers . . . 67
J.3 A canonical drive that records unseen labels . . . 68
J.4 A coordinate-invariant decoder . . . 69
J.5 One fixed globally admissible pair kernel . . . 70
J.6 The entire scalar-coordinate trajectory . . . 71
J.7 Random visits, every safe step, and the full average . . . 72
J.8 Optimization and statistical terms on the same box . . . 73
J.9 Full-gradient PGD on the Euclidean unit ball . . . 75

K Fixed-source risk laws **77**
K.1 Statement and quantifiers . . . 77
K.2 The conditional response keeps the coefficient $1/n$ . . . 78
K.3 From stability to population risk . . . 78
K.4 From occupancy to a fixed-source profile . . . 80

K.5 A fixed rare layer with unequal probabilities . . . 80
K.6 Global regularity and population risk . . . 81
K.7 PGD dynamics: identification, averaging, and persistence . . . 82
K.8 The same losses: a separate ridge verification . . . 83
K.9 The remaining terms and the fixed-source completion . . . 83
K.10 Zipf sources and interpretation . . . 84
K.11 A separation with nonvanishing anchor variance . . . 84
K.12 Empirical frequencies and qualitative consistency . . . 86

L Three comparisons with published bounds for the same learners **86**
L.1 Weak regularization in pairwise RERM . . . 86
L.2 Bounded trajectories for unprojected updates . . . 88
L.3 Scalar RCD and the Wu–Hu–Lei expression . . . 88
L.4 An explicitly derived signed-anchor baseline . . . 89
L.5 Full GD and actual individual-gradient path energy . . . 91
L.6 General GD/RCD coverage and broader source horizons . . . 92
L.7 Other smooth-GD comparisons . . . 93

M Further consequences and scope **93**
M.1 Exact threshold conventions . . . 93
M.2 Curvature-dependent statistical and iteration budgets . . . 94
M.3 A near-ERM bound with its confidence parameter integrated out . . . 95
M.4 Accuracy, training windows, and update counts . . . 96
M.5 Persistent deterioration on one instance . . . 96
M.6 Restoring radius and Lipschitz units . . . 96
M.7 The same sharp laws with globally nonnegative losses . . . 96
M.8 Dimension requirements and uniform stability . . . 97

Reading guide. Appendix A collects the main notation, followed by extended related work, proofs, application analyses, and further consequences.

Convex risk and matching trajectories. For Theorem 2.1, read the upper bounds in Appendix C, the low-curvature construction in Appendix E, and the high-curvature construction in Appendix H. The last uses the geometry of Appendix F and the smoothing estimates of Appendix G. Section H.9 assembles the persistent lower bound in Theorem 2.4.

Framework and applications. Appendix D proves the arbitrary-state absorption in Theorem 4.1, the algorithmic interface in Corollary 4.2, and the anchored pairwise refinement. The latter verifies the trajectory-energy and output-comparison budgets separately. Appendices J and K complete the coordinate and fixed-source laws. Appendix L verifies the published certificates on the specified learners and paths.

Weak negative curvature. Appendix I uses the response estimates of Appendix C and the code construction of Appendix E. Section I.8 proves the common-instance output separation in Corollary 3.2. Implementation details appear in Sections E.8 and I.5.5.

A Notation Reference

The table below collects the principal notation used in the main text and the interaction–dissipation framework. Each reference links to the corresponding definition. Replacement runs always share the same independent random seed.

Notation	Meaning	Definition
Learning setup and outputs		
n, d	Sample size and parameter dimension.	Section 2
S, P	Training sample $S = (Z_1, \ldots, Z_n) \sim P^n$ and observation distribution.	Section 2
$f(\boldsymbol{w}, z)$	Sample loss at parameter $\boldsymbol{w}$ and observation z.	Section 2
F_S, F	Empirical and population risks, respectively.	Section 2
$\mathbb{B}^d, \Pi_{\mathbb{B}^d}$	Euclidean unit ball and its Euclidean projection.	(1)
F^*	Minimum population risk on $\mathbb{B}^d$.	Section 2
η, T	Constant step size and number of updates.	(1)
q	Cumulative step size ηT for PGD and elapsed time for constrained flow.	Section 2
$\boldsymbol{w}_j, \boldsymbol{w}_T$	PGD iterate after j updates and the last iterate.	(1)
$\bar{\boldsymbol{w}}_T$	Full post-update parameter average $T^{-1}\sum_{j=1}^{T} \boldsymbol{w}_j$.	Section 2
$J, \boldsymbol{w}_J$	Independent uniform index in $\{1, \ldots, T\}$ and the corresponding checkpoint.	Section 2
$\widehat{\boldsymbol{w}}(S)$	Chosen output rule, with its independent seed suppressed.	Section 2
Convex risks and replacement influence		
L	Global smoothness constant of each sample loss.	Theorem 2.1
$\mathcal{R}_{n,L}(\widehat{\boldsymbol{w}})$	Worst-case expected excess risk for the fixed PGD parameters and output rule.	Section 2
$g_{n,L}(q)$	Sample-reuse growth term in the sharp convex risk law.	(3)
q_{plat}	Endpoint of the statistically optimal training window.	Figure 1
q_{crit}	Infimum over $q \geq n$ where worst-case risk reaches a fixed small constant.	Remark 2.2
$S^{(i)}$	Sample with observation i independently replaced.	Remark 2.3
$D^2_{\widehat{\boldsymbol{w}}}$	Average squared distance between outputs on S and $S^{(i)}$.	Remark 2.3
δ	Expected empirical suboptimality budget relative to a deterministic population minimizer.	(5)
Weakly convex output laws		
ρ	Weak-convexity parameter: $f(\cdot, z) + (\rho/2)\|\cdot\|^2$ is convex.	Section 3
$K_j^{(r)}$	Accumulated perturbation kernel $\eta\sum_{s=0}^{j-1}(1+\eta\rho)^{rs}$, $r \in \{1, 2\}$.	(8)
d_j	Capped replacement-response scale at update j.	(8)
$V(\widehat{\boldsymbol{w}})$	Output-dependent aggregation of d_j: final square, mean square, or square of the mean.	Section 3
$\mathcal{R}^{\mathrm{wc}}_{n,\rho}(\widehat{\boldsymbol{w}})$	Worst-case expected excess risk over the weakly convex class for a fixed output rule.	Section 3

Notation	Meaning	Definition
Weighted-factor interaction–dissipation		
K	Fixed closed convex feasible set with $0 \in K \subseteq \mathbb{B}^d$.	Section 4.1
$\mathcal{E}, e, Z_e$	Factor family, a factor's observation indices, and the observations it uses.	(11)
π_e, p_i	Factor weight and incidence weight $p_i = \sum_{e \ni i} \pi_e$.	Section 4.1
J_S, J, C_0, τ	Empirical and population objectives $C_0 + \tau F_S$ and $C_0 + \tau F$, with fixed convex background C_0.	(11)
$\mathcal{I}(\boldsymbol{X})$	Incidence-weighted squared replacement influence of a measurable state.	Section 4.1
$\mathcal{C}, \mathcal{D}$	Signed sample interaction and nonnegative dissipation from unchanged factors.	(D.2)
$\mathsf{E}_i, \mathsf{Q}_i$	Integration over atom Z_i alone and its centering operator $\mathsf{Q}_i = I - \mathsf{E}_i$.	Section D.1
N, P_3	Weighted centering operator $\sum_i p_i \mathsf{Q}_i$ and cubic incidence sum $\sum_i p_i^3$.	(D.1)
H_0, H_1	Incidence geometry coefficients for zero and positive dissipation components.	(D.3)
$\mathfrak{a}_\pi$	Basic incidence coefficient $\sqrt{H_0} + H_1 + P_3$.	Section D.1
$\boldsymbol{w}_\circ$	Deterministic population minimizer for risk conversion and deterministic anchor in the algebraic refinement.	Corollary 4.2
σ_0	Gradient-fluctuation norm $\|\nabla F_S(\boldsymbol{w}_\circ) - \nabla F(\boldsymbol{w}_\circ)\|_{L^2}$.	Corollary 4.2
$\boldsymbol{\xi}, \sigma_1$	Centered weighted anchor gradient and $\sigma_1 = \|\mathsf{N}\boldsymbol{\xi}\|_{L^2}$.	(D.4)
$r_e, U(\boldsymbol{X})$	Factor loss after subtracting its tangent at the anchor, and its weighted expected residual.	(D.4)
Pairwise learning, coordinate updates, and fixed sources		
$\lambda, \boldsymbol{A}_\lambda$	Ridge penalty and the corresponding constrained regularized minimizer.	Corollary 4.2
D	Scalar RCD dimension, with effective clock $q = \eta T / D$.	Corollary 4.2
$\boldsymbol{h}_K, \nu$	Centered first projection of the pair gradient at $\boldsymbol{w}_K^*$ and an upper bound on its L^2 norm.	Section 4.2
$\boldsymbol{w}_K^*, F_K^*$	Deterministic population minimizer on K and its risk value.	Section 4.2
s_ν	First-projection statistical scale $n^{-1} + \nu n^{-1/2}$.	Theorem 4.5
$p = (p_j)$	Fixed countable source distribution, where p_j is the probability of label j.	Section 4.2
K_n	Number of source labels with probability at least $1/n$.	Section 4.2
N	Number of conditionally i.i.d. observations remaining after conditioning.	Section K.2
α, ℓ	Regular-variation exponent in $(0, 1)$ and slowly varying function for source frequencies.	Section 4.2
$\mathcal{R}_p(n, t)$	Worst-case expected excess risk for the fixed source p.	(17)
t	Fixed-source effort parameter: ηT for PGD and λ^{-1} for ridge minimization.	Theorem 4.7

B Extended Related Work

Stability and generalization. Bousquet & Elisseeff (2002) connect algorithmic stability to generalization (Rakhlin et al., 2005; Mukherjee et al., 2006; Elisseeff et al., 2005), while Shalev-Shwartz et al. (2010) relate stability to learnability beyond uniform convergence. High-probability conversion from uniform stability was sharpened by Feldman & Vondrak (2018; 2019) and Bousquet et al. (2020). For iterative learning, Hardt et al. (2016) propagate sample perturbations through smooth updates (Zhang et al., 2022). Nikolakakis et al. (2025) extend classical SGD generalization bounds to arbitrary data-independent minibatch schedules for smooth Lipschitz losses. Finer analyses incorporate initialization risk and local curvature (Kuzborskij & Lampert, 2018), convergence and objective geometry (Charles & Papailiopoulos, 2018; Lei & Ying, 2021; Schliserman & Koren, 2022), and on-average squared sensitivity with risk-dependent bounds (Lei & Ying, 2020). Bassily et al. (2020) establish upper and lower stability bounds for nonsmooth convex losses. Extensions cover minimax learning (Lei et al., 2021b) and biased gradient estimators (Zeng & Lei, 2026). Trajectory-localized covering arguments provide another route to algorithm-dependent generalization (Park et al., 2022).

For full-batch GD, Nikolakakis et al. (2023) connect squared sensitivity, optimization error and individual-gradient path energy under smooth nonnegative losses without global Lipschitzness. Our interaction–dissipation mechanism retains signed sample interactions and absorbs them against dissipation from unchanged losses before accumulating costs. Combined with empirical comparison for the same output, this determines matching risk laws for bounded trajectories. Appendix L compares published certificates on the same learners, separating certificate looseness from actual excess risk.

Overfitting of gradient methods. Failures of ERM and uniform convergence can coexist with dimension-independent learnability (Shalev-Shwartz et al., 2010; Feldman, 2016). Carmon et al. (2024) sharpen ERM sample complexity, and Burla & Livni (2026) establish lower bounds in linear dimension. Amir et al. (2021b) separate SGD from GD, while Amir et al. (2021a) prove limitations for broader full-batch first-order methods. Training-parameter dependence is central to Livni (2024), Schliserman et al. (2025), and Zhang et al. (2025). For nonsmooth losses, Vansover-Hager et al. (2025) show that a second SGD pass can already raise excess population risk to constant order at a step size optimal for one pass.

Among the closest results, Burla & Livni (2026) obtain a projected-GD lower bound of order $\sqrt{q/n^{3/2}}$ without imposing a fixed smoothness budget. Our result shows that constant global smoothness changes this dependence: at $L = 1$, the sharp sample-reuse term is instead $q/n^{3/2}$. The gap is polynomial before saturation. For example, at $q \asymp n$, the Burla–Livni construction gives an $n^{-1/4}$ lower-bound scale, whereas in our 1-smooth class the worst-case excess risk remains $\Theta(n^{-1/2})$. Zhang et al. (2025), by contrast, obtain the classical q/n growth using a smooth non-realizable quadratic witness, but their trajectory is unprojected and may leave the unit ball, and their loss is not globally Lipschitz. Thus neither result determines how training time interacts with a fixed smoothness budget under bounded trajectories and global regularity. We close this gap by determining matching expected excess-risk laws throughout the (L, q) plane for projected GD with globally convex, 1-Lipschitz, L-smooth losses. At $L = 1$, the sharp law is $\Theta(\min 1, q^{-1} + n^{-1/2} + q/n^{3/2})$. Consequently, the risk remains $\Theta(n^{-1/2})$ at $q \asymp n$, when the classical q/n stability certificate is already constant-order, and reaches constant order only at $q \asymp n^{3/2}$. This resolves Livni (2024, Open Question 6) for this loss class. Appendix L.7 gives detailed comparisons with the closest constructions.

Iterate averaging and weak convexity. Neu & Rosasco (2018) relate geometrically weighted averaging of least-squares SGD iterates to ridge regularization. Richards & Rabbat (2021) derive population-risk guarantees for a uniformly selected GD iterate under empirical weak convexity. Wang et al. (2024) establish stability bounds for stochastic weight averaging with sampling both with and without replacement. Lei (2023) studies population-gradient guarantees for smooth nonconvex problems and Moreau-envelope guarantees for weakly convex nonsmooth problems. These works quantify benefits of averaging and curvature through risk or stationarity guarantees (Bach & Moulines, 2013; Dieuleveut & Bach, 2016; Davis & Drusvyatskiy, 2019).

Our weakly convex results determine matching excess-risk laws for the last iterate, a uniform checkpoint and the full parameter average under common samplewise curvature assumptions. Beyond the worst-case envelopes in Theorem 3.1, Corollary 3.2 realizes their separation on one trajectory with vanishing negative curvature. All three outputs have $O(n^{-1/2})$ risk earlier, but their later risks are respectively $\Theta(1)$, $\Theta((\log n)^{-1})$ and $\Theta((\log n)^{-2})$.

Pairwise learning and structured objectives. Complete-pair objectives are U-statistics with overlapping observations (Clémençon et al., 2008), whose orthogonal decomposition has classical foundations (Hoeffding, 1948; Efron & Stein, 1981). Lei et al. (2018) analyze regularized pairwise learning using localized U-process tools. Lei et al. (2020) sharpen stability bounds for regularized learning and pairwise SGD, including optimistic rates. Pairwise SGD analyses cover smooth and nonsmooth losses (Shen et al., 2020; Yang et al., 2021a;b) and uniform sampling of pairs of distinct examples at each update (Lei et al., 2021a). Most directly, Wu et al. (2025) study pairwise randomized coordinate descent through on-average argument stability, early stopping and low-noise conditions.

Our framework matches shared-observation interactions with dissipation using an absorption theorem for arbitrary bounded measurable states. Learner-specific update or KKT conditions supply the response, while the anchored refinement in Appendix D isolates the first-projection contribution. This yields sharp laws for weak pairwise regularization and coordinate updates, compared in Figure 2. For fixed-source learning, occupancy and missing-mass theory connect source frequencies to unseen observations (McAllester & Ortiz, 2003; Gnedin et al., 2007; Ben-Hamou et al., 2017). Appendix K couples this profile to replacement response, determining risk growth and saturation over training time without source knowledge in the algorithm.

C Shared response and risk bounds

This appendix closes the upper side of Theorem 2.1. The proof combines three estimates: optimization along actual projected updates, absorption of sample interactions (Lemma C.1), and conversion of squared influence to population risk. The last subsection records the last-iterate and classical comparisons.

C.1 Well-posedness, measurable outputs, and normalization

The constrained flow and its full time average are

$$\boldsymbol{u}_S' + \nabla F_S(\boldsymbol{u}_S) + N_{\mathbb{B}^d}(\boldsymbol{u}_S) \ni \mathbf{0}, \qquad \boldsymbol{u}_S(0) = \mathbf{0}, \qquad \boldsymbol{A}_q = q^{-1}\int_0^q \boldsymbol{u}_S(t)\,dt. \tag{C.1}$$

Here $N_{\mathbb{B}^d}$ is the outward normal cone. With dimension explicit, write $\mathcal{R}_{n,d,L}(\widehat{\boldsymbol{w}})$ for the supremum of $\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*]$ over the loss and distribution class of Section 2. The learner, its time parameters, and its output rule are fixed. For flow these rules are $\widehat{\boldsymbol{w}} = \boldsymbol{A}_q$ and $\widehat{\boldsymbol{w}} = \boldsymbol{u}_S(q)$. Both risks have the full curvature law of Theorem 2.1 at every $q > 0$.

Subtracting $f(\mathbf{0}, z)$ changes neither gradients nor excess risk. After this normalization, $|f(\boldsymbol{w}, z)| \le 1$ on the unit ball, so all risk differences and product-space calculations are integrable. A population minimizer exists by continuity and compactness. All sample functions and gradients are assumed jointly measurable in their displayed variables.

For a fixed sample, $\Phi = F_S + I_{\mathbb{B}^d}$ is a proper closed convex function. Its subdifferential is maximal monotone, so the finite-dimensional subgradient-flow theorem yields a unique locally absolutely continuous solution from 0 (Bauschke & Combettes, 2017). In the smooth case its velocity is $-\nabla F_S(\boldsymbol{u})$ projected onto the tangent cone. In particular $\|\boldsymbol{u}'(t)\| \le 1$ almost everywhere. The same speed bound holds for a globally Lipschitz convex F_S without smoothness. The evolution variational inequality is

$$\langle \boldsymbol{u}'(t), \boldsymbol{u}(t) - \boldsymbol{w}_\circ \rangle \le F_S(\boldsymbol{w}_\circ) - F_S(\boldsymbol{u}(t)), \qquad \boldsymbol{w}_\circ \in \mathbb{B}^d. \tag{C.2}$$

To see uniqueness directly, use the other solution as the comparator in each EVI and add. The squared distance has nonpositive derivative and vanishes initially. Measurability can be obtained from the measurable implicit Euler resolvents, or from the convergent explicit projected schemes in Lemma G.1. Thus the expectations in the paper involve well-defined measurable outputs. No choice of a normal-cone element is made by the loss oracle.

The common initial point makes the initial replacement influence zero. The orthogonal absorption inequality itself is valid for any bounded measurable state. It does not assume that state is reachable. Reachability enters through the empirical gradient equation and the common initial condition.

C.2 Cocoercivity and optimization of the updated average

Global convex L-smoothness gives cocoercivity (Bauschke & Combettes, 2017):

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq L^{-1} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2. \tag{C.3}$$

This uses no Hessian or interior constrained optimum.

For $\boldsymbol{x}^+ = \Pi_{\mathbb{B}^d}(\boldsymbol{x} - \eta \nabla F_S(\boldsymbol{x}))$, projection optimality says

$$\langle \nabla F_S(\boldsymbol{x}), \boldsymbol{x}^+ - \boldsymbol{w}_\circ \rangle \leq \frac{1}{\eta} \langle \boldsymbol{x} - \boldsymbol{x}^+, \boldsymbol{x}^+ - \boldsymbol{w}_\circ \rangle.$$

Convexity at $\boldsymbol{x}$ and the smooth upper inequality at $\boldsymbol{x}^+$ then give

$$\begin{aligned} F_S(\boldsymbol{x}^+) - F_S(\boldsymbol{w}_\circ) &\leq \langle \nabla F_S(\boldsymbol{x}), \boldsymbol{x}^+ - \boldsymbol{w}_\circ \rangle + \frac{L}{2} \|\boldsymbol{x}^+ - \boldsymbol{x}\|^2 \\ &\leq \frac{\|\boldsymbol{x} - \boldsymbol{w}_\circ\|^2 - \|\boldsymbol{x}^+ - \boldsymbol{w}_\circ\|^2}{2\eta} - \frac{1 - \eta L}{2\eta} \|\boldsymbol{x}^+ - \boldsymbol{x}\|^2. \end{aligned} \tag{C.4}$$

For $\eta \leq 1/L$, sum with fixed $\boldsymbol{w}_\circ = \boldsymbol{w}^*$ and apply convexity to the post-update average. Since $\boldsymbol{w}_0 = \mathbf{0}$ and $\|\boldsymbol{w}^*\| \leq 1$,

$$F_S(\bar{\boldsymbol{w}}_T) - F_S(\boldsymbol{w}^*) \leq \frac{1}{2\eta T}. \tag{C.5}$$

For flow, integrating (C.2) gives

$$F_S(\boldsymbol{A}_q) - F_S(\boldsymbol{w}^*) \leq \frac{\|\boldsymbol{w}^*\|^2 - \|\boldsymbol{u}_S(q) - \boldsymbol{w}^*\|^2}{2q} \leq \frac{1}{2q}. \tag{C.6}$$

For either averaged output $\boldsymbol{X}$, let $\Gamma(\boldsymbol{X}) = \mathbb{E}[F(\boldsymbol{X}) - F_S(\boldsymbol{X})]$ denote its expected generalization gap. Taking expectations yields $\mathbb{E}[F(\boldsymbol{X}) - F^*] \leq \Gamma(\boldsymbol{X}) + 1/(2q)$. This compares to a population minimizer, not a sample-dependent empirical minimizer.

C.3 Product-space identities and the flow budget

The conditional expectation $\mathsf{E}_i \boldsymbol{Y} = \mathbb{E}[\boldsymbol{Y} \mid S_{-i}]$ is an orthogonal projection on $\mathcal{H} = L^2(P^n; \mathbb{R}^d)$. If $i \neq j$, Fubini's theorem on the product probability space gives $\mathsf{E}_i \mathsf{E}_j = \mathsf{E}_j \mathsf{E}_i$. The operators

$$\mathsf{Q}_A = \prod_{i \in A} (I - \mathsf{E}_i) \prod_{i \notin A} \mathsf{E}_i$$

are orthogonal projections for $A \subseteq [n]$: if $A \neq B$, one coordinate in their symmetric difference contributes $\mathsf{E}_i(I - \mathsf{E}_i) = 0$. Expanding $\prod_i (\mathsf{E}_i + (I - \mathsf{E}_i))$ shows that they sum to the identity. Write $\boldsymbol{Y}_A = \mathsf{Q}_A \boldsymbol{Y}$, with the same component convention for other random vectors. Therefore

$$\|(I - \mathsf{E}_i)\boldsymbol{Y}\|_{L^2}^2 = \sum_{A \ni i} \|\boldsymbol{Y}_A\|_{L^2}^2, \qquad \sum_{A \neq \varnothing} \|\boldsymbol{Y}_A\|_{L^2}^2 = \|\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y}\|_{L^2}^2.$$

The identities used in the pointwise argument are

$$\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{Y}^{(i)}\|^2 = 2 \sum_{A \ni i} \|\boldsymbol{Y}_A\|_{L^2}^2, \tag{C.7}$$

$$\mathbb{E}\langle \boldsymbol{B}_i, \boldsymbol{X} - \boldsymbol{X}^{(i)} \rangle = \sum_{A \ni i} \langle \boldsymbol{B}_{i,A}, \boldsymbol{X}_A \rangle_{L^2}. \tag{C.8}$$

Conditional on S_{-i}, $\boldsymbol{Y}$ and $\boldsymbol{Y}^{(i)}$ are independent identically distributed copies. Their conditional squared distance is twice their conditional variance. This proves (C.7). Also

$$\mathbb{E}[\langle \boldsymbol{B}_i, \boldsymbol{X}^{(i)} \rangle \mid S] = \langle \boldsymbol{B}_i, \mathsf{E}_i \boldsymbol{X} \rangle,$$

which proves (C.8) without any restriction on the sample dependence of $\boldsymbol{B}_i$.

Lemma C.1 (Absorption of higher-order interactions). *If $\|\boldsymbol{X}\|_{L^2} \le R$ and $\|\boldsymbol{B}_i\|_{L^2} \le G$, then*

$$-\frac{4}{n^2}\sum_i \langle \boldsymbol{B}_i, (I-\mathsf{E}_i)\boldsymbol{X}\rangle_{L^2} - \frac{2}{Ln^2}\sum_{i\neq j}\mathbb{E}\|\boldsymbol{B}_j - \boldsymbol{B}_j^{(i)}\|^2 \le \frac{4GR}{n^{3/2}} + \frac{2LR^2}{n^2}. \tag{C.9}$$

Proof of Lemma C.1. Using (C.7)–(C.8), the left side is

$$-\frac{4}{n^2}\sum_A\sum_{i\in A}\langle \boldsymbol{B}_{i,A}, \boldsymbol{X}_A\rangle_{L^2} - \frac{4}{Ln^2}\sum_A\sum_j(|A| - \mathbf{1}_{\{j\in A\}})\|\boldsymbol{B}_{j,A}\|_{L^2}^2. \tag{C.10}$$

For $|A| = k \ge 2$, discard only the nonpositive terms with $j \notin A$. Each remaining term obeys

$$-4\langle \boldsymbol{b}, \boldsymbol{x}\rangle_{L^2} - \frac{4(k-1)}{L}\|\boldsymbol{b}\|_{L^2}^2 \le \frac{L}{k-1}\|\boldsymbol{x}\|_{L^2}^2,$$

by completing the square. Summing over $i \in A$ costs at most $Lk\|\boldsymbol{X}_A\|_{L^2}^2/(k-1) \le 2L\|\boldsymbol{X}_A\|_{L^2}^2$. Orthogonality bounds the total high-order contribution by $2LR^2/n^2$.

For $A = \{i\}$, Cauchy–Schwarz gives

$$\frac{4}{n^2}\sum_i \|\boldsymbol{B}_{i,\{i\}}\|_{L^2}\|\boldsymbol{X}_{\{i\}}\|_{L^2} \le \frac{4}{n^2}(nG^2)^{1/2}R.$$

The constant component does not occur. Combining the singleton and higher-order terms proves the claim. □

This proof also yields, with $\mathcal{E}_k = \sum_{|A|=k}\|\boldsymbol{X}_A\|_{L^2}^2$, the refined expression

$$\frac{4G}{n^{3/2}}\sqrt{\mathcal{E}_1} + \frac{L}{n^2}\sum_{k=2}^{n}\frac{k}{k-1}\mathcal{E}_k. \tag{C.11}$$

This explicitly separates first-order statistics from higher-order interactions. No assumption makes the higher-order energies vanish.

For the flow, put $v(t) = n^{-1}\sum_i \mathbb{E}\|\boldsymbol{u}_S(t) - \boldsymbol{u}_{S^{(i)}}(t)\|^2$ and let $D_q^2 = n^{-1}\sum_i \mathbb{E}\|\boldsymbol{A}_q(S) - \boldsymbol{A}_q(S^{(i)})\|^2$. The continuous counterpart of Theorem C.2 is

$$v'(t) \le \frac{4}{n^{3/2}} + \frac{2L}{n^2} \quad \text{a.e.,} \qquad D_q^2 \le q\left(\frac{2}{n^{3/2}} + \frac{L}{n^2}\right). \tag{C.12}$$

At a fixed time take $\boldsymbol{X} = \boldsymbol{u}_S(t)$ and $\boldsymbol{B}_j(S) = \nabla f(\boldsymbol{X}, Z_j)$. Normal-cone monotonicity bounds the replaced squared-distance derivative by $-2n^{-1}\sum_j \mathbb{E}\langle \boldsymbol{B}_j(S) - \boldsymbol{B}_j(S^{(i)}), \boldsymbol{X} - \boldsymbol{X}^{(i)}\rangle$. For $j \neq i$, cocoercivity gives the negative square of the common-loss gradient difference. For $j = i$, exchanging the original and fresh observation doubles $\mathbb{E}\langle \boldsymbol{B}_i(S), \boldsymbol{X} - \boldsymbol{X}^{(i)}\rangle$. Averaging over i therefore produces the left side of Lemma C.1, with $R = G = 1$.

States and velocities have norm at most one, so the squared-distance derivative is bounded. Fubini applied to integral increments justifies differentiation under expectation almost everywhere. Integrating the resulting bound and applying Jensen gives

$$\frac{1}{n}\sum_i \mathbb{E}\left\|\frac{1}{q}\int_0^q (\boldsymbol{u}_S(t) - \boldsymbol{u}_{S^{(i)}}(t))\,dt\right\|^2 \le \frac{1}{q}\int_0^q v(t)\,dt.$$

The factor $1/2$ in the time integral gives the constants in the averaged bound.

C.4 THE COMPLETE PROJECTED-STEP ARGUMENT

Theorem C.2 (Squared influence along GD). *Let $v_j = n^{-1}\sum_i \mathbb{E}\|\boldsymbol{w}_j(S) - \boldsymbol{w}_j(S^{(i)})\|^2$. For every admissible L-smooth instance and $0 < \eta \le 1/L$,*

$$v_{j+1} - v_j \le \frac{4\eta}{n^{3/2}} + \frac{8L\eta}{n^2} + \frac{12\eta^2}{n^2}. \tag{C.13}$$

For $L \ge 1$, either the last iterate or the complete post-update average has

$$D_{\boldsymbol{X}}^2 \le q(4n^{-3/2} + 20Ln^{-2}).$$

Proof. Fix an update and a replacement i. With $\boldsymbol{X} = \boldsymbol{w}_j(S)$, $\boldsymbol{Y} = \boldsymbol{w}_j(S^{(i)})$, $\boldsymbol{\Delta} = \boldsymbol{X} - \boldsymbol{Y}$, define

$$\boldsymbol{B}_k(S) = \nabla f(\boldsymbol{X}, Z_k), \quad \boldsymbol{A} = \frac{1}{n}\sum_{k\ne i}(\boldsymbol{B}_k(S) - \boldsymbol{B}_k(S^{(i)})), \quad \boldsymbol{B} = \frac{1}{n}(\boldsymbol{B}_i(S) - \boldsymbol{B}_i(S^{(i)})).$$

By Cauchy–Schwarz for the average and (C.3),

$$\|\boldsymbol{A}\|^2 \le \frac{n-1}{n^2}\sum_{k\ne i}\|\boldsymbol{B}_k(S) - \boldsymbol{B}_k(S^{(i)})\|^2 \le L\langle \boldsymbol{A}, \boldsymbol{\Delta}\rangle.$$

The last inequality remains valid for $n = 1$, when $\boldsymbol{A} = \boldsymbol{0}$. Since $\|\boldsymbol{B}\| \le 2/n$, projection and $\|\boldsymbol{u}+\boldsymbol{v}\|^2 \le \frac{3}{2}\|\boldsymbol{u}\|^2 + 3\|\boldsymbol{v}\|^2$ imply

$$\begin{aligned}\|\boldsymbol{\Delta}^+\|^2 &\le \|\boldsymbol{\Delta}\|^2 - 2\eta\langle \boldsymbol{A}, \boldsymbol{\Delta}\rangle - 2\eta\langle \boldsymbol{B}, \boldsymbol{\Delta}\rangle + \tfrac{3}{2}\eta^2\|\boldsymbol{A}\|^2 + 3\eta^2\|\boldsymbol{B}\|^2 \\ &\le \|\boldsymbol{\Delta}\|^2 - \tfrac{\eta}{2}\langle \boldsymbol{A}, \boldsymbol{\Delta}\rangle - 2\eta\langle \boldsymbol{B}, \boldsymbol{\Delta}\rangle + 12\eta^2/n^2. \end{aligned}\tag{C.14}$$

Use cocoercivity once more on $\langle \boldsymbol{A}, \boldsymbol{\Delta}\rangle$, average over the replaced coordinate, and exchange the changed sample. The increment in average squared distance is at most

$$\begin{aligned}&-\frac{4\eta}{n^2}\sum_A\sum_{i\in A}\langle \boldsymbol{B}_{i,A}, \boldsymbol{X}_A\rangle_{L^2} \\ &\quad -\frac{\eta}{Ln^2}\sum_A\sum_k(|A| - \mathbf{1}_{\{k\in A\}})\|\boldsymbol{B}_{k,A}\|_{L^2}^2 + 12\eta^2/n^2.\end{aligned}$$

The first two terms are η times the left side of Lemma C.1 with its parameter replaced by $4L$. Applying that lemma gives $4\eta n^{-3/2} + 8L\eta n^{-2} + 12\eta^2 n^{-2}$, which proves (C.13).

For $L \ge 1$ and $\eta \le 1/L$, $8L + 12\eta \le 20L$. Sum the recurrence from $v_0 = 0$ and use Jensen over $j = 1, \dots, T$. Since $\eta(T+1)/2 \le \eta T = q$,

$$\frac{1}{n}\sum_i \mathbb{E}\|\bar{\boldsymbol{w}}_T(S) - \bar{\boldsymbol{w}}_T(S^{(i)})\|^2 \le q(4n^{-3/2} + 20Ln^{-2}).$$

The proof covers the largest allowed step and the updated average without a limiting argument. □

C.5 POPULATION-CENTERED RISK CONVERSION

For a deterministic $\boldsymbol{w}^* \in \arg\min_{\mathbb{B}^d} F$, put

$$\boldsymbol{a}_z = \nabla f(\boldsymbol{w}^*, z), \quad \boldsymbol{\mu} = \mathbb{E}\boldsymbol{a}_z, \quad r_z(\boldsymbol{w}) = f(\boldsymbol{w}, z) - f(\boldsymbol{w}^*, z) - \langle \boldsymbol{a}_z, \boldsymbol{w} - \boldsymbol{w}^*\rangle.$$

We prove that any measurable output $\boldsymbol{X}(S) \in \mathbb{B}^d$ satisfying $\mathbb{E}[F_S(\boldsymbol{X}) - F_S(\boldsymbol{w}^*)] \le \delta$ obeys

$$\Gamma(\boldsymbol{X}) = \mathbb{E}[F(\boldsymbol{X}) - F_S(\boldsymbol{X})] \le n^{-1/2} + \sqrt{2L\delta + 4L/\sqrt{n}}\, D_{\boldsymbol{X}} + \tfrac{L}{2}D_{\boldsymbol{X}}^2. \tag{C.15}$$

Since sample gradients are bounded, dominated differentiation gives $\boldsymbol{\mu} = \nabla F(\boldsymbol{w}^*)$. For any feasible $\boldsymbol{w}$, optimality of $\boldsymbol{w}^*$ along the line segment to $\boldsymbol{w}$ gives $\langle \boldsymbol{\mu}, \boldsymbol{w} - \boldsymbol{w}^*\rangle \ge 0$.

The residual r_z is nonnegative on the entire ambient space. A smooth descent step gives $\|\nabla r_z(\boldsymbol{w})\|^2 \le 2Lr_z(\boldsymbol{w})$. With $\bar{\boldsymbol{a}}_S = n^{-1}\sum_i \boldsymbol{a}_{Z_i}$ and $r_S = n^{-1}\sum_i r_{Z_i}$,

$$\begin{aligned}\mathbb{E}r_S(\boldsymbol{X}) &= \mathbb{E}[F_S(\boldsymbol{X}) - F_S(\boldsymbol{w}^*) - \langle \bar{\boldsymbol{a}}_S, \boldsymbol{X} - \boldsymbol{w}^*\rangle] \\ &\le \delta + 2\mathbb{E}\|\bar{\boldsymbol{a}}_S - \boldsymbol{\mu}\|.\end{aligned}$$

Independence and $\|\boldsymbol{a}_Z\| \le 1$ give

$$\mathbb{E}\|\bar{\boldsymbol{a}}_S - \boldsymbol{\mu}\|^2 = \frac{1}{n}\big(\mathbb{E}\|\boldsymbol{a}_Z\|^2 - \|\boldsymbol{\mu}\|^2\big) \le \frac{1}{n}.$$

Consequently,

$$\frac{1}{n}\sum_i \mathbb{E}\|\nabla r_{Z_i}(\boldsymbol{X})\|^2 \le 2L\delta + 4L/\sqrt{n}. \tag{C.16}$$

The exact replace-one identity is

$$\Gamma(\boldsymbol{X}) = \frac{1}{n}\sum_i \mathbb{E}[f(\boldsymbol{X}^{(i)}, Z_i) - f(\boldsymbol{X}, Z_i)].$$

Here $\boldsymbol{X}^{(i)}$ is independent of Z_i and has the same marginal law as $\boldsymbol{X}$. Its affine contribution equals

$$\begin{aligned}\frac{1}{n}\sum_i \mathbb{E}\langle \boldsymbol{a}_{Z_i}, \boldsymbol{X}^{(i)} - \boldsymbol{X}\rangle &= \langle \boldsymbol{\mu}, \mathbb{E}\boldsymbol{X}\rangle - \mathbb{E}\langle \bar{\boldsymbol{a}}_S, \boldsymbol{X}\rangle \\ &\le \mathbb{E}\|\bar{\boldsymbol{a}}_S - \boldsymbol{\mu}\| \le n^{-1/2}.\end{aligned}$$

Apply Taylor's upper inequality to the residual at $\boldsymbol{X}$ and use Cauchy–Schwarz across both the coordinate i and the probability space. The residual linear contribution is at most $\sqrt{2L\delta + 4L/\sqrt{n}}\, D_{\boldsymbol{X}}$, and the quadratic term is $LD_{\boldsymbol{X}}^2/2$. This proves (C.15). Direct Lipschitz reading also gives $\Gamma(\boldsymbol{X}) \le D_{\boldsymbol{X}}$. No symmetry of the output beyond using the same algorithm on a replaced sample is needed for these averaged identities. At $L = 1$, Young's inequality bounds the cross term by $\delta + 2n^{-1/2} + D_{\boldsymbol{X}}^2/2$. Adding the empirical comparison gives (5).

C.6 Closing the upper rates and threshold converses

For $L = 1$, the preceding bounds imply $D_{\boldsymbol{X}}^2 \le 24q/n^{3/2}$. The terms in (C.15) are bounded by a universal multiple of

$$n^{-1/2} + n^{-3/4} + \sqrt{q}/n + q/n^{3/2} \le C(n^{-1/2} + q/n^{3/2}).$$

Add (C.5) or (C.6), and cap the population excess by the loss diameter two. This proves the unit-smooth specialization of (2).

When $1 \le L \le \sqrt{n}$, we again have $D_{\boldsymbol{X}}^2 \le Cq/n^{3/2}$. The term involving L/q in (C.15) is at most $C\sqrt{L}/n^{3/4} \le Cn^{-1/2}$. Its other cross term is at most $C\sqrt{Lq}/n \le C(n^{-1/2} + Lq/n^{3/2})$. Thus smooth reading gives $C(n^{-1/2} + Lq/n^{3/2})$. Lipschitz reading gives $C\sqrt{q/n^{3/2}}$. The minimum of these bounds is at most

$$C\left(n^{-1/2} + \min\{Lq/n^{3/2}, \sqrt{q/n^{3/2}}\}\right).$$

This proves the low-curvature upper formula.

For $L \ge \sqrt{n}$, taking the minimum of Lipschitz reading and the classical uniform bound closes the high-curvature upper law. Section H.1 gives the full calculation. For general $L \ge 1$, Lipschitz reading gives the necessary direction of (M.1). Fix $r_0 > 0$ and take n sufficiently large that $1/(2n) \le r_0/2$. If $q \ge n$ and expected excess is at least r_0, then

$$q \ge \frac{c(r_0)}{n^{-3/2} + Ln^{-2}} = c(r_0)\frac{n^2}{\sqrt{n} + L}.$$

Together with $q \ge n$, this is the necessary threshold. At exact time $q = n$, the continuous bound reads

$$\mathcal{R}_{n,d,L}(\boldsymbol{A}_n) \le \frac{1}{2n} + \sqrt{2/\sqrt{n} + L/n}.$$

Hence fixed positive risk requires $L \ge c(r_0)n$ for sufficiently large n. The discrete bound has larger universal constants and the same conclusion. These converse statements hold in every finite dimension.

C.7 Elementary matching optimization and statistical terms

On $[-1,1]$, take the deterministic affine loss $f(w) = -cw$, where $c = \min\{1, 1/(2q)\}$. Both flow and GD move right at speed c, with right-endpoint projection if needed. At the horizon q, every iterate is at most $cq \le 1/2$ when $q \ge 1/2$. For $q < 1/2$, the same conclusion holds because $c = 1$. The optimum is at 1, so expected excess is at least $c/2$, a universal multiple of $\min\{1, q^{-1}\}$. Affine losses are admissible for every positive smoothness upper bound.

For the statistical term, let $f(w,z) = zw$, $z \in \{-1,1\}$, with $\mathbb{E}z = 1/(4\sqrt{n})$. Its population optimum is -1. On the event $\bar{z} \le 0$, all iterates from zero are nonnegative, so the excess is at least $1/(4\sqrt{n})$ at every positive horizon. Relative to the unbiased product distribution, the chi-square divergence is

$$(1 + 1/(16n))^n - 1 \le e^{1/16} - 1 < 1/8.$$

Total variation is at most half the square root of chi-square, hence less than $1/4$. The unbiased event $\bar{z} \le 0$ has probability at least $1/2$, so the biased event has probability at least $1/4$. This proves expected excess at least $1/(16\sqrt{n})$ for every horizon and every step. Padding by zero coordinates embeds either example in any larger dimension. Lower bounds for different terms are combined by taking a maximum over instances, using $\max_j \min\{1, a_j\} \ge \frac{1}{3}\min\{1, \sum_{j=1}^3 a_j\}$.

C.8 Last iterates and the classical uniform bound

For $n \ge 1$ and any finite dimension, write $\boldsymbol{X}_{\mathrm{last}} = \boldsymbol{u}_S(q)$ for flow at $q > 0$, and $\boldsymbol{X}_{\mathrm{last}} = \boldsymbol{w}_T$ for GD with $0 < \eta \le 1/L$, integer $T \ge 1$ and $q = \eta T$. The remaining conventions are those of Section 2.

Lemma C.3 (Last-iterate optimization and influence)**.** *For every admissible smooth loss family,*

$$F_S(\boldsymbol{X}_{\mathrm{last}}) - F_S(\boldsymbol{w}^*) \le (2q)^{-1} \quad \textit{for either dynamics}, \tag{C.17}$$

$$\frac{1}{n}\sum_i \mathbb{E}\|\boldsymbol{u}_S(q) - \boldsymbol{u}_{S^{(i)}}(q)\|^2 \le q(4n^{-3/2} + 2Ln^{-2}), \tag{C.18}$$

$$\frac{1}{n}\sum_i \mathbb{E}\|\boldsymbol{w}_T(S) - \boldsymbol{w}_T(S^{(i)})\|^2 \le q(4n^{-3/2} + 20Ln^{-2}) \qquad (L \ge 1). \tag{C.19}$$

Consequently every smooth upper risk law and every necessary curvature–time bound stated in the main text also holds for the last iterate.

Proof. Constrained gradient flow decreases F_S: its energy identity gives $\frac{d}{dt}F_S(\boldsymbol{u}_S(t)) = -\|\boldsymbol{u}_S'(t)\|^2$ almost everywhere. Therefore the EVI, integrated with the fixed comparator $\boldsymbol{w}^*$, implies

$$F_S(\boldsymbol{u}_S(q)) - F_S(\boldsymbol{w}^*) \le \frac{1}{q}\int_0^q [F_S(\boldsymbol{u}_S(t)) - F_S(\boldsymbol{w}^*)]dt \le \frac{1}{2q}.$$

For GD, (C.4) with $\boldsymbol{w}_\circ = \boldsymbol{w}_{j-1}$ shows monotonicity of $F_S(\boldsymbol{w}_j)$ for every safe step. Summing that same inequality with $\boldsymbol{w}_\circ = \boldsymbol{w}^*$ yields

$$F_S(\boldsymbol{w}_T) - F_S(\boldsymbol{w}^*) \le \frac{1}{T}\sum_{j=1}^T [F_S(\boldsymbol{w}_j) - F_S(\boldsymbol{w}^*)] \le \frac{1}{2\eta T}.$$

This uses an average of objective values, not a comparison between $F_S(\boldsymbol{w}_T)$ and $F_S(\bar{\boldsymbol{w}}_T)$. Integrating the flow influence inequality, or summing its discrete counterpart from the common initial state, gives (C.18)–(C.19). The population-centered readout (C.15) applies to any measurable output with these two properties. Its substitution and the rate simplifications in Appendix C therefore prove the assertion, with the same universal orders. □

For later use, the classical uniform bound holds for both smooth outputs:

$$\mathbb{E}[F(\boldsymbol{X}) - F^*] \le \min\{2, (2q)^{-1} + 2q/n\}. \tag{C.20}$$

For GD, the map $\boldsymbol{w} \mapsto \Pi_{\mathbb{B}}(\boldsymbol{w} - \eta\nabla F_S(\boldsymbol{w}))$ is nonexpansive when $\eta \le 1/L$, by cocoercivity and nonexpansiveness of projection. Changing one sample at the same state changes one update by at most $2\eta/n$. Induction gives distance at most $2q/n$ for the last iterate and hence for its full average. Replace-one symmetrization and Lipschitzness finish the proof. For flow, the two EVIs give distance at most $2t/n$ at time t, so the same bound applies. Combining this with (C.17), or with the existing average-output optimization bound, proves (C.20).

D A GENERAL INTERACTION–DISSIPATION THEOREM

Theorem 4.1 is the $R = G = 1$ case of the unanchored inequality in Theorem D.1. Corollary 4.2 combines the iterative response in Corollary D.3, the regularized response in Lemma D.4, and the population-risk conversion in Lemma D.5. In this setting, the bounds (D.17) and (D.20) imply $\mathcal{I}(\widehat{\boldsymbol{w}}) \le 12q\tau\mathfrak{a}_\pi$ and $\mathcal{I}(\widehat{\boldsymbol{w}}) \le 2\tau\mathfrak{a}_\pi/\lambda$, respectively, using $0 < \eta, \tau \le 1$ for the iterative bound. The coefficient $\mathfrak{a}_\pi$ is defined below. This appendix also retains the anchored refinement (D.7), its statistical inputs, and the trajectory-energy response used by the first-projection application.

D.1 INDEPENDENT ATOMS, WEIGHTED FACTORS, AND REPLACEMENT OPERATORS

Let $Z_1, \dots, Z_N$ be independent random variables, with arbitrary individual laws, and let ω be independent algorithmic randomness. Every replacement below keeps ω fixed. For a square-integrable vector-valued random variable $\boldsymbol{Y} = \boldsymbol{Y}(Z_1, \dots, Z_N, \omega)$, let $\mathsf{E}_i\boldsymbol{Y}$ integrate Z_i alone, and put $\mathsf{Q}_i = I - \mathsf{E}_i$. Inner products and norms with subscript L^2 refer to L^2 over both the atoms and ω. Let $\mathcal{E}$ be a finite family of nonempty subsets of $[N]$, with deterministic weights $\pi_e \ge 0$ and $\sum_e \pi_e = 1$. Define

$$p_i = \sum_{e \ni i} \pi_e, \qquad \mathsf{N} = \sum_{i=1}^N p_i \mathsf{Q}_i, \qquad P_3 = \sum_{i=1}^N p_i^3. \tag{D.1}$$

Atoms with $p_i = 0$ can be deleted. For each e, the factor loss $f_e(\boldsymbol{w}; Z_e)$ is jointly measurable, finite, convex, and 1-smooth on $\mathbb{R}^d$. All gradients evaluated at the states under consideration have norm at most G. When risk values are used, assume integrability of f_e at a deterministic reference point. Gradient bounds on the relevant bounded domain then ensure that its risk differences are integrable. The constants R, G below need not equal one.

For any measurable state $\boldsymbol{X} = \boldsymbol{X}(Z_1, \dots, Z_N, \omega)$ with $\|\boldsymbol{X}\| \le R$ almost surely, write $\boldsymbol{G}_e = \nabla f_e(\boldsymbol{X}; Z_e)$. An independent replacement of Z_i gives $\boldsymbol{X}^{(i)}$, using the same output rule and seed. Set

$$\begin{aligned}
\mathcal{I}(\boldsymbol{X}) &= \sum_i p_i \mathbb{E}\|\boldsymbol{X} - \boldsymbol{X}^{(i)}\|^2, \\
\mathcal{C}(\boldsymbol{X}) &= -\sum_e \pi_e \sum_{i \in e} p_i \langle \boldsymbol{G}_e, \mathsf{Q}_i \boldsymbol{X} \rangle_{L^2}, \\
\mathcal{D}(\boldsymbol{X}) &= \sum_e \pi_e \sum_{i \notin e} p_i \|\mathsf{Q}_i \boldsymbol{G}_e\|_{L^2}^2.
\end{aligned} \tag{D.2}$$

The sign of $\mathcal{C}$ is part of its definition. It must be retained when combining the interaction with the nonnegative dissipation. For $\varnothing \ne A \subseteq [N]$, define

$$\begin{aligned}
t_{Ae} &= \sum_{i \in A \cap e} p_i, & d_{Ae} &= \sum_{i \in A \setminus e} p_i, \\
H_0 &= \max_{A \ne \varnothing} \sum_{e : d_{Ae} = 0} \pi_e t_{Ae}^2, & H_1 &= \max_{A \ne \varnothing} \sum_{e : d_{Ae} > 0} \pi_e \frac{t_{Ae}^2}{d_{Ae}}.
\end{aligned} \tag{D.3}$$

An empty sum is zero. These are explicit quantities determined by the incidence pattern and its actual weights. The basic incidence coefficient in Corollary 4.2 is

$$\mathfrak{a}_\pi := \sqrt{H_0} + H_1 + P_3.$$

For an anchored bound, fix a deterministic $\boldsymbol{w}_\circ \in \mathbb{R}^d$ independent of the atoms and seed, and put

$$\begin{aligned}
\boldsymbol{a}_e &= \nabla f_e(\boldsymbol{w}_\circ; Z_e), & \bar{\boldsymbol{a}} &= \sum_e \pi_e \boldsymbol{a}_e, & \boldsymbol{\xi} &= \bar{\boldsymbol{a}} - \mathbb{E}\bar{\boldsymbol{a}}, \\
\sigma_0 &= \|\boldsymbol{\xi}\|_{L^2}, & \sigma_1 &= \|\mathsf{N}\boldsymbol{\xi}\|_{L^2}, & & \\
r_e(\boldsymbol{w}) &= f_e(\boldsymbol{w}; Z_e) - f_e(\boldsymbol{w}_\circ; Z_e) - \langle \boldsymbol{a}_e, \boldsymbol{w} - \boldsymbol{w}_\circ \rangle, & U(\boldsymbol{X}) &= \sum_e \pi_e \mathbb{E} r_e(\boldsymbol{X}). & &
\end{aligned} \tag{D.4}$$

Convexity and global smoothness imply

$$r_e(\boldsymbol{w}) \geq 0, \qquad \|\nabla r_e(\boldsymbol{w})\|^2 \leq 2r_e(\boldsymbol{w}). \tag{D.5}$$

Indeed, r_e is nonnegative everywhere, and its smoothness inequality at $\boldsymbol{w} - \nabla r_e(\boldsymbol{w})$ gives the second assertion. No assertion that $\boldsymbol{w}_\circ$ minimizes a population objective is needed yet.

D.2 The algebraic inequality

Theorem D.1 (Weighted-factor interaction–dissipation)**.** *Under the preceding assumptions, for every $c > 0$,*

$$\mathcal{C}(\boldsymbol{X}) - c\mathcal{D}(\boldsymbol{X}) \leq RG\sqrt{H_0} + \frac{R^2 H_1}{4c}, \tag{D.6}$$

$$\mathcal{C}(\boldsymbol{X}) - c\mathcal{D}(\boldsymbol{X}) \leq R\sigma_1 + R\sqrt{2H_0 U(\boldsymbol{X})} + \frac{R^2 H_1}{4c}. \tag{D.7}$$

The second inequality uses the additional anchored inputs in (D.4). Both hold for arbitrary measurable states. No algorithmic reachability assumption is needed.

Proof. The operators E_i are commuting orthogonal projections. Thus

$$\boldsymbol{Y}_A = \prod_{i \in A} \mathsf{Q}_i \prod_{i \notin A} \mathsf{E}_i \boldsymbol{Y}, \qquad A \subseteq [N], \tag{D.8}$$

give orthogonal components of $\boldsymbol{Y}$ (Hoeffding, 1948; Efron & Stein, 1981). Their coefficients may depend on ω, which has not been integrated by any E_i. Orthogonality still holds after integration over ω. In particular,

$$\sum_{A \neq \varnothing} \|\boldsymbol{X}_A\|_{L^2}^2 = \mathbb{E}\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X} \mid \omega]\|^2 \leq R^2.$$

Expanding the two terms in (D.2) gives the exact identity

$$\mathcal{C} - c\mathcal{D} = \sum_{A \neq \varnothing} \sum_e \pi_e \left(-t_{Ae} \langle \boldsymbol{G}_{e,A}, \boldsymbol{X}_A \rangle_{L^2} - c d_{Ae} \|\boldsymbol{G}_{e,A}\|_{L^2}^2 \right). \tag{D.9}$$

If $d_{Ae} > 0$, completing a square in the Hilbert space yields

$$-t_{Ae} \langle \boldsymbol{G}_{e,A}, \boldsymbol{X}_A \rangle_{L^2} - c d_{Ae} \|\boldsymbol{G}_{e,A}\|_{L^2}^2 \leq \frac{t_{Ae}^2}{4c d_{Ae}} \|\boldsymbol{X}_A\|_{L^2}^2.$$

The sum of these terms is at most $R^2 H_1/(4c)$. For the remaining terms, Cauchy–Schwarz first within each A, and then over A, gives

$$\left| \sum_{A \neq \varnothing} \sum_{e : d_{Ae} = 0} \pi_e t_{Ae} \langle \boldsymbol{G}_{e,A}, \boldsymbol{X}_A \rangle_{L^2} \right|$$
$$\leq \sqrt{H_0} \left(\sum_{A \neq \varnothing} \|\boldsymbol{X}_A\|_{L^2}^2 \right)^{1/2} \left(\sum_{A \neq \varnothing} \sum_{e : d_{Ae} = 0} \pi_e \|\boldsymbol{G}_{e,A}\|_{L^2}^2 \right)^{1/2} \leq RG\sqrt{H_0}.$$

This proves (D.6).

For (D.7), write $\boldsymbol{G}_e = \boldsymbol{a}_e + \boldsymbol{b}_e$, where $\boldsymbol{b}_e = \nabla r_e(\boldsymbol{X})$. The variable $\boldsymbol{a}_e$ depends only on Z_e, so its A-component is zero unless $A \subseteq e$. In particular, $\mathsf{Q}_i \boldsymbol{a}_e = \boldsymbol{0}$ for $i \notin e$, and replacing $\boldsymbol{G}_e$ by $\boldsymbol{b}_e$ leaves $\mathcal{D}$ unchanged. The entire anchor interaction equals

$$-\sum_e \pi_e \sum_{i \in e} p_i \langle \boldsymbol{a}_e, \mathsf{Q}_i \boldsymbol{X} \rangle_{L^2} = -\left\langle \sum_i p_i \mathsf{Q}_i \bar{\boldsymbol{a}}, \boldsymbol{X} \right\rangle_{L^2}$$
$$= -\langle \mathsf{N}\boldsymbol{\xi}, \boldsymbol{X} \rangle_{L^2} \leq R\sigma_1.$$

Apply the preceding square completion to the $\boldsymbol{b}_e$ terms. Their zero-dissipation energy is at most $\sum_e \pi_e \mathbb{E}\|\boldsymbol{b}_e\|^2 \leq 2U(\boldsymbol{X})$ by (D.5). This proves the stated bound. □

All statements can also be used conditionally on a background random variable, provided the remaining atoms have a product conditional law, the conditional weights are fixed, and the background is unchanged by every replacement. The anchor in (D.7) must then be fixed under this conditioning and these replacements. Conditioning on an arbitrary data-dependent iterate does not in general meet these conditions. The unanchored inequality avoids the need to select a conditional population optimizer.

D.3 Stability bounds for projected and coordinate updates

Let the actual empirical objective be

$$J_S(\boldsymbol{w}) = C_0(\boldsymbol{w}) + \tau \sum_e \pi_e f_e(\boldsymbol{w}; Z_e), \qquad 0 < \tau \le 1, \tag{D.10}$$

where C_0 is fixed relative to the atoms being replaced. In this subsection C_0 is convex and L_C-smooth, with $L_C + \tau \le 1$. Let K be a fixed closed convex domain containing zero. Consider projected GD with constant $0 < \eta \le 1$, initialized at $\boldsymbol{X}_0 = \mathbf{0}$. Alternatively, partition the parameter space into D orthogonal blocks and take $K = K_1 \times \cdots \times K_D$. At step t, choose one block uniformly and independently of the data and past, take its partial gradient step, and project only that block onto K_j. The same block is chosen in the original and replacement runs. Set

$$\kappa = 1 \quad \text{for PGD}, \qquad \kappa = 1/D \quad \text{for coordinate updates}, \qquad h = \kappa\eta, \qquad q = hT. \tag{D.11}$$

Scalar coordinates are the special case of one-dimensional blocks. Assume that all original and replacement states have norm at most R, and all factor gradients at those states have norm at most G. This holds, for example, when $K \subseteq R\mathbb{B}^d$ and those gradients are bounded on K.

Lemma D.2 (Replacement stability for projected and coordinate updates)**.** *For either algorithm, with the quantities in* (D.2) *evaluated at* $\boldsymbol{X}_t$*,*

$$\mathcal{I}(\boldsymbol{X}_{t+1}) - \mathcal{I}(\boldsymbol{X}_t) \le 4h\tau\big(\mathcal{C}(\boldsymbol{X}_t) - \mathcal{D}(\boldsymbol{X}_t)/4\big) + 12h\eta\tau^2 G^2 P_3. \tag{D.12}$$

Proof. Fix an atom replacement i and the two states $\boldsymbol{x} = \boldsymbol{X}_t$ and $\boldsymbol{y} = \boldsymbol{X}_t^{(i)}$. Write $\boldsymbol{\Delta} = \boldsymbol{x} - \boldsymbol{y}$ and split $\nabla J_S(\boldsymbol{x}) - \nabla J_{S^{(i)}}(\boldsymbol{y}) = \boldsymbol{A}_i + \tau \boldsymbol{D}_i$, where

$$\boldsymbol{A}_i = \nabla C_0(\boldsymbol{x}) - \nabla C_0(\boldsymbol{y}) + \tau \sum_{e \not\ni i} \pi_e [\nabla f_e(\boldsymbol{x}; Z_e) - \nabla f_e(\boldsymbol{y}; Z_e)],$$

$$\boldsymbol{D}_i = \sum_{e \ni i} \pi_e [\nabla f_e(\boldsymbol{x}; Z_e) - \nabla f_e(\boldsymbol{y}; Z_e^{(i)})].$$

The common objective defining $\boldsymbol{A}_i$ is convex and has smoothness at most $L_C + \tau \le 1$. Cocoercivity therefore gives $\|\boldsymbol{A}_i\|^2 \le \langle \boldsymbol{A}_i, \boldsymbol{\Delta}\rangle$. This also holds if the common objective is affine. Also $\|\boldsymbol{D}_i\| \le 2Gp_i$. Nonexpansiveness of projection onto the same K gives

$$\begin{aligned}\|\boldsymbol{x}^+ - \boldsymbol{y}^+\|^2 - \|\boldsymbol{\Delta}\|^2 &\le -2\eta\langle \boldsymbol{A}_i, \boldsymbol{\Delta}\rangle - 2\eta\tau\langle \boldsymbol{D}_i, \boldsymbol{\Delta}\rangle + \tfrac{3}{2}\eta^2\|\boldsymbol{A}_i\|^2 + 3\eta^2\tau^2\|\boldsymbol{D}_i\|^2 \\ &\le -\tfrac{\eta}{2}\langle \boldsymbol{A}_i, \boldsymbol{\Delta}\rangle - 2\eta\tau\langle \boldsymbol{D}_i, \boldsymbol{\Delta}\rangle + 12\eta^2\tau^2 G^2 p_i^2.\end{aligned}$$

We used $\eta \le 1$ and the elementary inequality $\|\boldsymbol{u} + \boldsymbol{v}\|^2 \le \frac{3}{2}\|\boldsymbol{u}\|^2 + 3\|\boldsymbol{v}\|^2$. Exchanging Z_i and its independent replacement, with the seed fixed, gives

$$\sum_i p_i \mathbb{E}\langle \boldsymbol{D}_i, \boldsymbol{X}_t - \boldsymbol{X}_t^{(i)}\rangle = 2\sum_e \pi_e \sum_{i \in e} p_i \langle \boldsymbol{G}_e, \mathsf{Q}_i \boldsymbol{X}_t\rangle_{L^2} = -2\mathcal{C}(\boldsymbol{X}_t). \tag{D.13}$$

For a common factor $e \not\ni i$, its gradient in the replacement world is exactly $\boldsymbol{G}_e^{(i)}$. Factorwise cocoercivity, monotonicity of the background, and the conditional-variance identity yield

$$\begin{aligned}\sum_i p_i \mathbb{E}\langle \boldsymbol{A}_i, \boldsymbol{X}_t - \boldsymbol{X}_t^{(i)}\rangle &\ge \tau \sum_i p_i \sum_{e \not\ni i} \pi_e \mathbb{E}\|\boldsymbol{G}_e - \boldsymbol{G}_e^{(i)}\|^2 \\ &= 2\tau \sum_e \pi_e \sum_{i \notin e} p_i \|\mathsf{Q}_i \boldsymbol{G}_e\|_{L^2}^2 = 2\tau\mathcal{D}(\boldsymbol{X}_t).\end{aligned}$$

Summing the full-step estimate with weights p_i proves (D.12) for PGD.

For coordinate updates let $T_S(\boldsymbol{x}) = \Pi_K(\boldsymbol{x} - \eta \nabla J_S(\boldsymbol{x}))$ be the full projected candidate, and let M_t be the uniformly selected block, shared by the two runs. Product separability gives the exact conditional identity

$$\mathbb{E}_{M_t}\|\boldsymbol{X}_{t+1} - \boldsymbol{X}^{(i)}_{t+1}\|^2 = (1-\kappa)\|\boldsymbol{x} - \boldsymbol{y}\|^2 + \kappa\|T_S(\boldsymbol{x}) - T_{S^{(i)}}(\boldsymbol{y})\|^2. \tag{D.14}$$

Every increment in the full-step estimate is multiplied by κ. This proves the assertion for coordinates. In particular, no nonexpansiveness claim about a fixed partial-gradient map is used. □

Corollary D.3 (Response from a trajectory energy budget)**.** *Suppose, in the anchored setting, that a deterministic number $B \geq 0$ satisfies*

$$\sum_{t=0}^{T-1} hU(\boldsymbol{X}_t) \leq B. \tag{D.15}$$

For $\widehat{\boldsymbol{w}} = \boldsymbol{X}_j$ with any $0 \leq j \leq T$, and for every convex average $\widehat{\boldsymbol{w}} = \sum_{j=0}^{T} \alpha_j \boldsymbol{X}_j$ with deterministic weights,

$$\mathcal{I}(\widehat{\boldsymbol{w}}) \leq 4\tau\left(qR\sigma_1 + R\sqrt{2H_0 qB} + qR^2 H_1\right) + 12\eta q\tau^2 G^2 P_3. \tag{D.16}$$

Without an anchor or trajectory energy budget, the corresponding bound is

$$\mathcal{I}(\widehat{\boldsymbol{w}}) \leq 4q\tau(RG\sqrt{H_0} + R^2 H_1) + 12\eta q\tau^2 G^2 P_3. \tag{D.17}$$

Proof. Use Theorem D.1 with $c = 1/4$ in Lemma D.2. Since $\mathcal{I}(\boldsymbol{X}_0) = 0$, summing through time $j-1$ gives the bound for $\boldsymbol{X}_j$: the prefix query energy is at most B because each $U(\boldsymbol{X}_t)$ is nonnegative, and

$$\sum_{t<j} h\sqrt{U(\boldsymbol{X}_t)} \leq \sqrt{\left(\sum_{t<j} h\right)\left(\sum_{t<j} hU(\boldsymbol{X}_t)\right)} \leq \sqrt{qB}.$$

Jensen's inequality applied to every paired replacement gives $\mathcal{I}(\sum_j \alpha_j \boldsymbol{X}_j) \leq \sum_j \alpha_j \mathcal{I}(\boldsymbol{X}_j)$. The unanchored proof uses (D.6) instead. □

The averaging conclusion uses the same deterministic weights in both sample worlds. For data-dependent weights, the exact difference is

$$\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{w}}^{(i)} = \sum_t \alpha_t(S)(\boldsymbol{X}_t - \boldsymbol{X}^{(i)}_t) + \sum_t [\alpha_t(S) - \alpha_t(S^{(i)})]\boldsymbol{X}^{(i)}_t.$$

The additional weight-change term is not controlled by the state influence bound. Even a deterministic average needs its own empirical output budget before Lemma D.5 can turn influence into excess risk.

The same conclusions hold for unprojected updates when boundedness of all original and replacement trajectories has been established separately. One then uses the identity map in place of projection. A bounded comparator alone does not establish this hypothesis.

D.4 Stability of regularized minimizers

Lemma D.4 (KKT replacement response)**.** *In (D.10), now assume only that C_0 is finite, differentiable and convex. Let $\boldsymbol{X}$ minimize $J_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2/2$ over the same fixed closed convex K, where $\lambda > 0$. Suppose solutions exist, have norm at most R, and the preceding factor-gradient bounds hold. Then*

$$\lambda\mathcal{I}(\boldsymbol{X}) \leq 2\tau\big(\mathcal{C}(\boldsymbol{X}) - \mathcal{D}(\boldsymbol{X})\big), \tag{D.18}$$

$$\mathcal{I}(\boldsymbol{X}) \leq \frac{2\tau}{\lambda}\left(R\sigma_1 + R\sqrt{2H_0 U(\boldsymbol{X})} + R^2 H_1/4\right), \tag{D.19}$$

$$\mathcal{I}(\boldsymbol{X}) \leq \frac{2\tau}{\lambda}(RG\sqrt{H_0} + R^2 H_1/4). \tag{D.20}$$

Proof. For a replaced atom, let $\boldsymbol{x}, \boldsymbol{y}$ be the two minimizers. Their variational inequalities, tested at each other, imply

$$\lambda\|\boldsymbol{x}-\boldsymbol{y}\|^2 \le -\langle \nabla J_S(\boldsymbol{x}) - \nabla J_{S^{(i)}}(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle.$$

The common normal cone has thus been accounted for, including minimizers on the boundary. Split the gradients as in Lemma D.2. Monotonicity of C_0 and factorwise cocoercivity again give a common pairing of at least $2\tau\mathcal{D}$ after weighted expectation. Exchange gives the changed pairing $-2\tau\mathcal{C}$. This proves (D.18). The other inequalities follow from Theorem D.1 with $c = 1$. □

D.5 From stability to population risk

Define the population factor objective $F(\boldsymbol{w}) = \sum_e \pi_e \mathbb{E} f_e(\boldsymbol{w}; Z_e)$ and its residual $r_F(\boldsymbol{w}) = \sum_e \pi_e \mathbb{E} r_e(\boldsymbol{w})$, where the expectations in these definitions use fresh independent atoms. Let $\widehat{\boldsymbol{w}}^{(e)}$ replace all atoms in e simultaneously, keeping the same output rule and seed, and define

$$W(\widehat{\boldsymbol{w}}) = \sum_e \pi_e \mathbb{E}\|\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{w}}^{(e)}\|^2.$$

Lemma D.5 (Factor replacement and population risk)**.** *For every bounded measurable output $\widehat{\boldsymbol{w}}$,*

$$\mathbb{E} r_F(\widehat{\boldsymbol{w}}) \le 2U(\widehat{\boldsymbol{w}}) + W(\widehat{\boldsymbol{w}}), \qquad W(\widehat{\boldsymbol{w}}) \le \mathcal{I}(\widehat{\boldsymbol{w}}). \tag{D.21}$$

If $\widehat{\boldsymbol{w}} \in K$ almost surely, $\boldsymbol{w}_\circ \in \arg\min_K (C_0 + \tau F)$, and

$$\mathbb{E}[J_S(\widehat{\boldsymbol{w}}) - J_S(\boldsymbol{w}_\circ)] \le \delta. \tag{D.22}$$

Then

$$\mathbb{E}[(C_0+\tau F)(\widehat{\boldsymbol{w}}) - (C_0+\tau F)(\boldsymbol{w}_\circ)] \le 2\delta + 2\tau R\sigma_0 + \tau\mathcal{I}(\widehat{\boldsymbol{w}}). \tag{D.23}$$

Proof. Smoothness, (D.5), and Young's inequality imply, for all $\boldsymbol{x}, \boldsymbol{y}$,

$$r_e(\boldsymbol{y}) \le r_e(\boldsymbol{x}) + \sqrt{2r_e(\boldsymbol{x})}\|\boldsymbol{y}-\boldsymbol{x}\| + \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \le 2r_e(\boldsymbol{x}) + \|\boldsymbol{y}-\boldsymbol{x}\|^2.$$

The random output $\widehat{\boldsymbol{w}}^{(e)}$ is independent of the original atoms Z_e and has the same marginal law as $\widehat{\boldsymbol{w}}$. Consequently

$$\mathbb{E} r_F(\widehat{\boldsymbol{w}}) = \sum_e \pi_e \mathbb{E} r_e(\widehat{\boldsymbol{w}}^{(e)}; Z_e) \le 2U(\widehat{\boldsymbol{w}}) + W(\widehat{\boldsymbol{w}}).$$

In terms of the atom components of $\widehat{\boldsymbol{w}}$, exact conditional-variance identities give

$$W(\widehat{\boldsymbol{w}}) = 2\sum_{A\neq\varnothing}\left(\sum_e \pi_e \mathbf{1}_{\{A\cap e\neq\varnothing\}}\right)\|\widehat{\boldsymbol{w}}_A\|_{L^2}^2,$$

$$\mathcal{I}(\widehat{\boldsymbol{w}}) = 2\sum_{A\neq\varnothing}\left(\sum_{i\in A} p_i\right)\|\widehat{\boldsymbol{w}}_A\|_{L^2}^2.$$

For every A, the coefficient in the first display is no larger than that in the second, proving $W \le \mathcal{I}$.

Put $\boldsymbol{\mu} = \mathbb{E}\bar{\boldsymbol{a}}$ and $A(\boldsymbol{w}) = C_0(\boldsymbol{w}) - C_0(\boldsymbol{w}_\circ) + \tau\langle\boldsymbol{\mu}, \boldsymbol{w}-\boldsymbol{w}_\circ\rangle$. Convexity of C_0 and constrained first-order optimality of $\boldsymbol{w}_\circ$ give $A(\boldsymbol{w}) \ge 0$ for every $\boldsymbol{w} \in K$. This includes boundary optimizers and does not require $\boldsymbol{\mu} = \mathbf{0}$. The exact empirical decomposition is

$$J_S(\widehat{\boldsymbol{w}}) - J_S(\boldsymbol{w}_\circ) = A(\widehat{\boldsymbol{w}}) + \tau\langle\boldsymbol{\xi}, \widehat{\boldsymbol{w}}-\boldsymbol{w}_\circ\rangle + \tau\sum_e \pi_e r_e(\widehat{\boldsymbol{w}}).$$

Because $\boldsymbol{w}_\circ$ is deterministic and $\mathbb{E}\boldsymbol{\xi} = \mathbf{0}$, $|\mathbb{E}\langle\boldsymbol{\xi}, \widehat{\boldsymbol{w}}-\boldsymbol{w}_\circ\rangle| = |\mathbb{E}\langle\boldsymbol{\xi}, \widehat{\boldsymbol{w}}\rangle| \le R\sigma_0$. Applying (D.21), and using $A(\widehat{\boldsymbol{w}}) \ge 0$, yields

$$\begin{aligned}\mathbb{E}[(C_0+\tau F)(\widehat{\boldsymbol{w}}) - (C_0+\tau F)(\boldsymbol{w}_\circ)] &= \mathbb{E}A(\widehat{\boldsymbol{w}}) + \tau\mathbb{E} r_F(\widehat{\boldsymbol{w}}) \\ &\le 2[\mathbb{E}A(\widehat{\boldsymbol{w}}) + \tau U(\widehat{\boldsymbol{w}})] + \tau W(\widehat{\boldsymbol{w}}) \\ &\le 2\delta + 2\tau R\sigma_0 + \tau\mathcal{I}(\widehat{\boldsymbol{w}}).\end{aligned}$$

□

Only (D.21), rather than (D.23), may be invoked when a conditional anchor is not a minimizer of the conditional population objective. Original empirical contributions must then be reassembled before using an optimization estimate for the original objective.

D.6 THE FIRST-PROJECTION-ADAPTIVE PAIRWISE UPPER BOUND

Let $n \geq 2$, let $Z_1, \dots, Z_n$ be i.i.d., and let $f(\boldsymbol{w}; z, z') = f(\boldsymbol{w}; z', z)$ be a symmetric pair loss satisfying the global convexity and unit smoothness assumptions. Its gradient has norm at most one on a fixed closed convex $K \subseteq \mathbb{B}^d$ containing zero. Define

$$F_S(\boldsymbol{w}) = \binom{n}{2}^{-1} \sum_{i<j} f(\boldsymbol{w}; Z_i, Z_j), \qquad F(\boldsymbol{w}) = \mathbb{E} f(\boldsymbol{w}; Z, Z'),$$

where Z, Z' are independent. Let $\boldsymbol{w}^* \in K$ be a deterministic population minimizer, and let $\boldsymbol{a}(Z, Z') = \nabla f(\boldsymbol{w}^*; Z, Z')$. Its orthogonal decomposition is

$$\begin{aligned} \boldsymbol{\mu} &= \mathbb{E}\boldsymbol{a}(Z, Z'), \qquad \boldsymbol{h}_1(Z) = \mathbb{E}[\boldsymbol{a}(Z, Z') \mid Z] - \boldsymbol{\mu}, \\ \boldsymbol{h}_2(Z, Z') &= \boldsymbol{a}(Z, Z') - \boldsymbol{\mu} - \boldsymbol{h}_1(Z) - \boldsymbol{h}_1(Z'), \\ \nu_1^2 &= \mathbb{E}\|\boldsymbol{h}_1(Z)\|^2, \qquad \nu_2^2 = \mathbb{E}\|\boldsymbol{h}_2(Z, Z')\|^2. \end{aligned} \tag{D.24}$$

In particular, $\mathbb{E}[\boldsymbol{h}_2(Z, Z') \mid Z] = \boldsymbol{0}$ and $2\nu_1^2 + \nu_2^2 = \mathbb{E}\|\boldsymbol{a} - \boldsymbol{\mu}\|^2 \leq 1$.

Theorem D.6 (First-projection-adaptive PGD and coordinate risk)**.** *Suppose $\nu_1 \leq \nu \leq 1$, and write $s_\nu = n^{-1} + \nu n^{-1/2}$. Run either projected GD on K, or uniform block-coordinate projected GD on a product domain K, with $0 < \eta \leq 1$, zero initialization, and clock $q = \kappa\eta T$ as in (D.11). For the last iterate $\widehat{\boldsymbol{w}} = \boldsymbol{X}_T$ and the full post-update average $\widehat{\boldsymbol{w}} = T^{-1}\sum_{t=1}^{T} \boldsymbol{X}_t$,*

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F(\boldsymbol{w}^*)] \leq C \min\left\{1, q^{-1} + s_\nu + \frac{q s_\nu}{n}\right\}, \tag{D.25}$$

where C is a universal constant, independent of dimension, block sizes, n, T, η, and ν. The algorithm is not given ν or the population minimizer.

Here $F(\boldsymbol{w}^*) = F_K^* := \min_K F$: the anchor and risk benchmark use the algorithm's actual feasible set. If $K \subseteq \mathbb{B}$ and risk is instead measured against $F^* = \min_{\mathbb{B}} F$, add the approximation error $F_K^* - F^*$ to the right-hand side of (D.25).

Proof. We compute the incidence and anchor inputs, and then establish the query and output budgets for the actual algorithms.

Incidence calculation. For comparison, $n \geq 2$ singleton factors have $p_i = 1/n$. The only zero-dissipation term for a given nonempty A occurs when $A = e = \{i\}$, giving $H_0 = n^{-3}$. For $|A| = k \geq 2$, only the k factors inside A contribute, and their sum is $k/[n^2(k-1)]$. Thus $H_1 = 2/n^2$ and $P_3 = 1/n^2$.

Now take all unordered pairs as factors, with $\pi_e = \binom{n}{2}^{-1}$, so $p_i = 2/n$ and $P_3 = 8/n^2$. If $|A| = k$, then $t_{Ae} = (2/n)|A \cap e|$ and $d_{Ae} = (2/n)(k - |A \cap e|)$. The only nonzero zero-dissipation contributions occur for $k = 1, 2$, and their sums are respectively $8/n^3$ and $32/[n^3(n-1)]$. For $k = 2$, the positive-dissipation sum is $8(n-2)/[n^2(n-1)]$. For $k \geq 3$, it is

$$\frac{2}{n}\left(\frac{2k(n-k)}{n(n-1)(k-1)} + \frac{4k(k-1)}{n(n-1)(k-2)}\right) \leq \frac{30}{n^2}.$$

The two terms are at most $6/n^2$ and $24/n^2$, using $k/(k-1) \leq 3/2$, $k/(k-2) \leq 3$, and $k \leq n$. Thus, including $n = 2$,

$$H_0 \leq 32/n^3, \qquad H_1 \leq 32/n^2, \qquad P_3 = 8/n^2. \tag{D.26}$$

For singleton factors with $n \geq 2$, these calculations give $\mathfrak{a}_\pi = n^{-3/2} + 3n^{-2}$. For complete pair factors, the $k = 1$ contribution gives $H_0 \geq 8/n^3$, so (D.26) yields $\mathfrak{a}_\pi = \Theta(n^{-3/2})$. Thus both designs have the same order of the basic incidence coefficient.

Anchor variances. The empirical anchor fluctuation is exactly

$$\boldsymbol{\xi} = \frac{2}{n}\sum_i \boldsymbol{h}_1(Z_i) + \binom{n}{2}^{-1}\sum_{i<j}\boldsymbol{h}_2(Z_i, Z_j).$$

Degeneracy makes distinct pair terms orthogonal even when they share an atom. The two orders are also orthogonal. Moreover N has eigenvalue $2/n$ on order one and $4/n$ on order two. Therefore

$$\sigma_0^2 = \frac{4\nu_1^2}{n} + \frac{2\nu_2^2}{n(n-1)},$$
$$\sigma_1^2 = \frac{16\nu_1^2}{n^3} + \frac{32\nu_2^2}{n^3(n-1)}. \tag{D.27}$$

Since $\nu_2 \le 1$, these give $\sigma_0 \le 2s_\nu$ and $\sigma_1 \le 8s_\nu/n$.

The actual optimization budget. Fix the sample and current state $\boldsymbol{x}$, write $\boldsymbol{g} = \nabla F_S(\boldsymbol{x})$, and let $\boldsymbol{y} = \Pi_K(\boldsymbol{x} - \eta\boldsymbol{g})$ be the full projected candidate. For coordinate updates put $\boldsymbol{x}^+ = \boldsymbol{x} + P_j(\boldsymbol{y} - \boldsymbol{x})$, where P_j is the selected block projection. Projection optimality gives

$$\tfrac{1}{2}(\|\boldsymbol{x} - \boldsymbol{w}^*\|^2 - \|\boldsymbol{y} - \boldsymbol{w}^*\|^2) \ge \eta\langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{w}^*\rangle + \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{D.28}$$

Smoothness along each actual block update gives

$$\mathbb{E}_j F_S(\boldsymbol{x}^+) \le F_S(\boldsymbol{x}) + \kappa\langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}\rangle + \tfrac{\kappa}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

Also $\mathbb{E}_j\|\boldsymbol{x}^+ - \boldsymbol{w}^*\|^2 = (1-\kappa)\|\boldsymbol{x} - \boldsymbol{w}^*\|^2 + \kappa\|\boldsymbol{y} - \boldsymbol{w}^*\|^2$. Multiply (D.28) by κ, add the expected objective decrease times η, and use $\eta \le 1$ and convexity. The result is

$$h[F_S(\boldsymbol{x}) - F_S(\boldsymbol{w}^*)] \le \tfrac{1}{2}(\|\boldsymbol{x} - \boldsymbol{w}^*\|^2 - \mathbb{E}_j\|\boldsymbol{x}^+ - \boldsymbol{w}^*\|^2) + \eta[F_S(\boldsymbol{x}) - \mathbb{E}_j F_S(\boldsymbol{x}^+)]. \tag{D.29}$$

For full PGD the same proof has $\kappa = 1$ and no mask expectation. The projection condition applied blockwise with comparator $\boldsymbol{x}_j$ gives $\langle \boldsymbol{g}_j, \boldsymbol{y}_j - \boldsymbol{x}_j\rangle \le -\|\boldsymbol{y}_j - \boldsymbol{x}_j\|^2/\eta$. Consequently every actual step decreases F_S, for either algorithm and for every seed.

Sum (D.29) and average over data and masks. The initialization is zero, $\|\boldsymbol{w}^*\| \le 1$, and $F_S(\mathbf{0}) - F_S(\boldsymbol{X}_T) \le \|\boldsymbol{X}_T\| \le 1$. Hence

$$\sum_{t=0}^{T-1} h\mathbb{E}[F_S(\boldsymbol{X}_t) - F_S(\boldsymbol{w}^*)] \le \tfrac{3}{2}. \tag{D.30}$$

Pathwise descent and convexity give the two comparisons

$$F_S(\boldsymbol{X}_T),\ F_S\left(T^{-1}\sum_{t=1}^{T}\boldsymbol{X}_t\right) \le \frac{1}{T}\sum_{t=1}^{T}F_S(\boldsymbol{X}_t) \le \frac{1}{T}\sum_{t=0}^{T-1}F_S(\boldsymbol{X}_t).$$

Thus the pre-update query estimate gives, for both actual outputs,

$$\mathbb{E}[F_S(\widehat{\boldsymbol{w}}) - F_S(\boldsymbol{w}^*)] \le \frac{3}{2q}. \tag{D.31}$$

The population first-order pairing $\langle \boldsymbol{\mu}, \boldsymbol{w} - \boldsymbol{w}^*\rangle$ is nonnegative on K. The residual decomposition and $\mathbb{E}\langle \boldsymbol{\xi}, \boldsymbol{w}^*\rangle = 0$ give

$$U(\boldsymbol{X}_t) \le \mathbb{E}[F_S(\boldsymbol{X}_t) - F_S(\boldsymbol{w}^*)] - \mathbb{E}\langle \boldsymbol{\xi}, \boldsymbol{X}_t\rangle \le \mathbb{E}[F_S(\boldsymbol{X}_t) - F_S(\boldsymbol{w}^*)] + \sigma_0.$$

Summing the genuine pre-update queries therefore yields

$$B := \sum_{t=0}^{T-1} hU(\boldsymbol{X}_t) \le \tfrac{3}{2} + q\sigma_0. \tag{D.32}$$

Closing the risk. Apply Corollary D.3 with $R = G = \tau = 1$ and $C_0 = 0$, and insert (D.26)–(D.32). Up to universal constants the resulting response is bounded by

$$\frac{qs_\nu}{n} + \sqrt{\frac{q}{n^3}} + \frac{q\sqrt{s_\nu}}{n^{3/2}} + \frac{q}{n^2}.$$

Since $s_\nu \geq 1/n$,

$$\sqrt{q/n^3} \leq \tfrac{1}{2}(n^{-1} + q/n^2), \qquad \frac{q\sqrt{s_\nu}}{n^{3/2}} \leq \frac{qs_\nu}{n}, \qquad q/n^2 \leq qs_\nu/n.$$

Lemma D.5, with the actual output estimate (D.31), proves $\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F(\boldsymbol{w}^*)] \leq C(q^{-1} + s_\nu + qs_\nu/n)$. Finally, the Lipschitz diameter of K bounds the excess by two, giving (D.25). □

Corollary D.7 (Pairwise regularized minimization)**.** *Under the statistical assumptions of Theorem D.6, let $\widehat{\boldsymbol{w}}$ minimize $F_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2/2$ over K, for $\lambda > 0$. Then*

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F(\boldsymbol{w}^*)] \leq C \min\{1, \lambda + s_\nu + s_\nu/(\lambda n)\}. \tag{D.33}$$

Proof. Optimality against $\boldsymbol{w}^*$ gives the pathwise comparison $F_S(\widehat{\boldsymbol{w}}) - F_S(\boldsymbol{w}^*) \leq \lambda/2$, so $U(\widehat{\boldsymbol{w}}) \leq \lambda/2 + \sigma_0$ by the same anchor decomposition. Use Lemma D.4, (D.26), and (D.27). With $q = 1/\lambda$ the response terms have the same forms qs_ν/n, $\sqrt{q/n^3}$, $q\sqrt{s_\nu}/n^{3/2}$ and q/n^2 as above. Lemma D.5 with $\delta = \lambda/2$ proves the assertion. □

E Efficient smooth lower bounds in linear dimension

This appendix proves the low-curvature lower side of Theorem 2.1, including fixed-instance persistence. No regularity property is assumed only along the trajectory: all sample losses are globally convex, Lipschitz, and smooth. The two $O(n)$-dimensional blocks have distinct roles: memory identifies unseen labels, and the code turns that information into risk. The proof first establishes a globally valid smooth loss and its proximal representation, then verifies identification, actual safe-step updates, and the complete averaged prefix. Section E.8 collects the numerical implementation.

E.1 A constructive code family

Fix a binary code with message length k, block length $m \leq Ck$, positive relative distance, a polynomial-time encoder, and a bounded-distance decoder correcting a fixed fraction $r_{\mathrm{dec}} > 0$ of coordinate errors. Such constructive asymptotically good families exist (Spielman, 1996). Constant-factor padding handles every message length. Normalize each codeword to a unit vector $\boldsymbol{c_v} \in \{-1/\sqrt{m}, 1/\sqrt{m}\}^m$, indexed by $\boldsymbol{v} \in \{-1,1\}^k$. There is a fixed $0 \leq \rho < 1$ with $\langle \boldsymbol{c_u}, \boldsymbol{c_v}\rangle \leq \rho$ for $\boldsymbol{u} \neq \boldsymbol{v}$.

Choose a fixed $\kappa < 1$ sufficiently close to one that

$$2\kappa^2 > 1 + \rho, \qquad 2(1-\kappa) < r_{\mathrm{dec}}, \qquad \zeta = (1+\kappa)/2, \quad \Delta = \zeta - \kappa > 0. \tag{E.1}$$

For a unit vector $\boldsymbol{y}$, correlation $\langle \boldsymbol{c_v}, \boldsymbol{y}\rangle > \kappa$ implies $\|\boldsymbol{y} - \boldsymbol{c_v}\|^2 < 2(1-\kappa)$. Every sign mismatch costs at least $1/m$ in this squared distance, so rounding the signs of $\boldsymbol{y}$ and applying the decoder finds $\boldsymbol{v}$. At most one codeword has correlation at least κ: otherwise $2\kappa \leq \|\boldsymbol{c_v} + \boldsymbol{c_u}\| \leq \sqrt{2+2\rho}$. This efficiently readable geometry is closely related to the bad-ERM readout of Feldman (2016). Below we construct a smooth optimization dynamics that reaches it.

E.2 A convex mask with a one-code proximal oracle

Let $k \geq 2$, $\boldsymbol{c} : \{-1,1\}^k \to \mathbb{B}^m$, and $a, b > 0$. Put

$$T_{k-1}(\boldsymbol{x}) = \sum_{j=1}^{k-1} |\boldsymbol{x}|_{(j)} = \|\boldsymbol{x}\|_1 - \min_j |x_j|,$$

where the order is decreasing, and define

$$\begin{aligned} H(\boldsymbol{x}, \boldsymbol{y}) &= aT_{k-1}(\boldsymbol{x}) + b\|\boldsymbol{y}\|, \\ L_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{y}) &= a\langle \boldsymbol{v}, \boldsymbol{x}\rangle - b\langle \boldsymbol{c}(\boldsymbol{v}), \boldsymbol{y}\rangle, \\ P(\boldsymbol{x}, \boldsymbol{y}) &= \max\{H(\boldsymbol{x}, \boldsymbol{y}), \max_{\boldsymbol{v}\in\{-1,1\}^k} L_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{y})\}. \end{aligned} \tag{E.2}$$

The top-sum norm and its proximal machinery are classical. See, for example, the convex-analytic top-k literature Sander et al. (2023). What matters here is the exact mask of the arbitrary code map.

Lemma E.1 (Value masking)**.** *Let $\boldsymbol{\sigma} = \operatorname{sign}(\boldsymbol{x})$, with a fixed convention at zero. Then*

$$P(\boldsymbol{x}, \boldsymbol{y}) = \max\{H(\boldsymbol{x}, \boldsymbol{y}), L_{\boldsymbol{\sigma}}(\boldsymbol{x}, \boldsymbol{y})\}.$$

Moreover $P \geq 0$, $P(0) = 0$, and P is globally C_P-Lipschitz for $C_P = \sqrt{ka^2 + b^2}$ when the codewords are unit vectors. Every subgradient has memory coordinates bounded by a in absolute value and a code-block norm bounded by b.

Proof. If $\boldsymbol{v} \neq \boldsymbol{\sigma}$ changes a nonzero coordinate, its memory score loses $2a\sum_{j:v_j\neq\sigma_j} |x_j|$. This is at least $2a \min_j |x_j|$. Consequently

$$L_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{y}) \leq a\|\boldsymbol{x}\|_1 - 2a \min_j |x_j| + b\|\boldsymbol{y}\| \leq H(\boldsymbol{x}, \boldsymbol{y}).$$

If the only changed coordinates are zero, then $\min_j |x_j| = 0$ and the same domination holds. Every affine branch and every supporting gradient of H has the stated block bounds. Taking their convex hull preserves the bounds. Since $H \geq 0$ and every term vanishes at zero, the remaining claims follow. □

A sign formula for values alone would not establish efficient smooth gradients. The proximal reduction is stronger.

Theorem E.2 (One-code proximal reduction)**.** *Fix an input $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{y})$ and $\lambda > 0$, and let $\boldsymbol{\sigma} = \operatorname{sign}(\boldsymbol{x})$. Then*

$$\operatorname{prox}_{\lambda P}(\boldsymbol{w}) = \operatorname{prox}_{\lambda \max\{H, L_{\boldsymbol{\sigma}}\}}(\boldsymbol{w}).$$

The right side requires one evaluation of $\boldsymbol{c}(\boldsymbol{\sigma})$ and a one-dimensional concave maximization. In particular the Moreau envelope P_λ has a polynomial-accuracy gradient oracle whenever the code map can be evaluated efficiently.

Proof. Set $R = \max\{H, L_{\boldsymbol{\sigma}}\}$ and let $\boldsymbol{z}$ minimize $\lambda R(\boldsymbol{z}) + \|\boldsymbol{z} - \boldsymbol{w}\|^2/2$. Suppose some nonzero memory coordinate z_j has sign opposite to x_j. Then

$$H(\boldsymbol{z}) - L_{\boldsymbol{\sigma}}(\boldsymbol{z}) \geq 2a|z_j| - a \min_{\ell} |z_\ell| > 0.$$

Thus $R = H$ locally at $\boldsymbol{z}$. Every memory subgradient of H in coordinate j has the sign of z_j or is zero. The proximal condition $x_j = z_j + \lambda g_j$ then contradicts the opposite-sign assumption. If $x_j = 0$, the same condition rules out $z_j \neq 0$: when its sign opposes the chosen sign convention, H is strictly active, and when its sign agrees, both active gradients have that sign. Therefore the proximal point is in the input orthant, with zero coordinates allowed.

If there are no zero coordinates at $\boldsymbol{z}$, Lemma E.1 gives $P(\boldsymbol{z}) = R(\boldsymbol{z})$. If there are zero coordinates, every code branch is bounded by $H(\boldsymbol{z})$, so equality again holds. Since $P \geq R$ everywhere, $\boldsymbol{z}$ is also the minimizer for P.

For $\boldsymbol{g} = (a\boldsymbol{\sigma}, -b\boldsymbol{c}(\boldsymbol{\sigma}))$ and $t \in [0, 1]$, put $\boldsymbol{z}(t) = \operatorname{prox}_{\lambda(1-t)H}(\boldsymbol{w} - \lambda t\boldsymbol{g})$. The scalar concave dual has derivative $\lambda[L_{\boldsymbol{\sigma}}(\boldsymbol{z}(t)) - H(\boldsymbol{z}(t))]$. The proximal map of H separates into Euclidean shrinkage and the memory prox, which subtracts the projection onto $\{\boldsymbol{v} : \|\boldsymbol{v}\|_\infty \leq \lambda(1-t)a, \|\boldsymbol{v}\|_1 \leq \lambda(1-t)a(k-1)\}$. Sorting and scalar thresholding compute this projection. Lemma E.7 supplies the certified scalar search and error bound, including endpoints and vanishing derivatives. Thus the reduced prox has polynomial arithmetic accuracy, and $\nabla P_\lambda(\boldsymbol{w}) = (\boldsymbol{w} - \operatorname{prox}_{\lambda P}(\boldsymbol{w}))/\lambda$. □

Lemma E.3 (Exact branch certificate after smoothing)**.** *Assume the codewords have norm one. If* $\boldsymbol{\sigma} = \operatorname{sign}(\boldsymbol{x})$ *and*

$$a\big(\min_i |x_i| - \lambda a\big) > 2b(\|\boldsymbol{y}\| + \lambda b), \tag{E.3}$$

then

$$\nabla P_\lambda(\boldsymbol{x}, \boldsymbol{y}) = (a\boldsymbol{\sigma}, -b\boldsymbol{c}(\boldsymbol{\sigma})).$$

Proof. Let $\boldsymbol{g} = (a\boldsymbol{\sigma}, -b\boldsymbol{c}(\boldsymbol{\sigma}))$ and evaluate P at $\boldsymbol{w} - \lambda\boldsymbol{g}$. Its memory sign is still $\boldsymbol{\sigma}$. At that point,

$$L_{\boldsymbol{\sigma}} - H = a \min_i |x_i - \lambda a\sigma_i| - b\langle \boldsymbol{c}(\boldsymbol{\sigma}), \boldsymbol{y} + \lambda b\boldsymbol{c}(\boldsymbol{\sigma})\rangle - b\|\boldsymbol{y} + \lambda b\boldsymbol{c}(\boldsymbol{\sigma})\| > 0.$$

Lemma E.1 then implies $\boldsymbol{g} \in \partial P(\boldsymbol{w} - \lambda\boldsymbol{g})$. The proximal optimality condition gives $\operatorname{prox}_{\lambda P}(\boldsymbol{w}) = \boldsymbol{w} - \lambda\boldsymbol{g}$. □

E.3 A SMOOTH READOUT WITH A GLOBAL NUMERICAL ORACLE

For $0 < \theta \le 1$ define

$$H_\theta(\boldsymbol{y}) = (\|\boldsymbol{y}\| - \theta)_+, \qquad h_{i,\theta}(\boldsymbol{y}) = \max\left\{H_\theta(\boldsymbol{y}), \max_{\boldsymbol{v}: v_i = +1} (\langle \boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{y}\rangle - \kappa\theta)\right\}. \tag{E.4}$$

For the smoothness parameter $L \ge 1$ let $R_{i,\theta,L} = (h_{i,\theta})_{1/L}$ be its Moreau envelope on the entire ambient space. The function $h_{i,\theta}$ is convex, nonnegative, globally 1-Lipschitz, and zero at zero. Thus $R_{i,\theta,L}$ is convex, globally 1-Lipschitz and L-smooth, with the same zero set. At every zero its unique gradient is zero.

Lemma E.4 (Radial-mask reduction and exact reading)**.** *The value and Moreau gradient of* $h_{i,\theta}$ *can be approximated with polynomial arithmetic complexity, using at most one bounded-distance decoding call and a scalar search. If* $\|\boldsymbol{y}\| \le \theta$ *and an admissible codeword* $\boldsymbol{c}$ *has positive margin* $s = \langle \boldsymbol{c}, \boldsymbol{y}\rangle - \kappa\theta$*, that word is unique and*

$$R_{i,\theta,L}(\boldsymbol{y}) = \begin{cases} Ls^2/2, & s \le 1/L, \\ s - 1/(2L), & s \ge 1/L. \end{cases} \tag{E.5}$$

If no admissible word has positive margin in this ball, the value and gradient are zero.

Proof. A code branch exceeding $H_\theta(\boldsymbol{y})$ must have $\langle \boldsymbol{c}, \boldsymbol{y}\rangle > \kappa\|\boldsymbol{y}\|$. For $\|\boldsymbol{y}\| \le \theta$ this follows from its positive value. For $\|\boldsymbol{y}\| \ge \theta$, the inequality gives $\langle \boldsymbol{c}, \boldsymbol{y}\rangle > \|\boldsymbol{y}\| - (1-\kappa)\theta \ge \kappa\|\boldsymbol{y}\|$. The angular cap in (E.1) contains at most one codeword. Decode the signs of $\boldsymbol{y}$ (with a fixed convention at zero). Retain its code branch if the returned message has bit $v_i = +1$. Otherwise retain no code branch. No angular threshold comparison is needed.

If no word is retained, any branch exceeding the radial mask at the radial proximal point would have input-direction correlation greater than κ and would have been decoded with the required bit. Thus that radial point is the full proximal point. If a word $\boldsymbol{c}$ is retained, compute the prox of $\max\{H_\theta, \langle \boldsymbol{c}, \cdot\rangle - \kappa\theta\}$. At its minimizer $\boldsymbol{z}$, either the radial branch is strictly active, in which case $\boldsymbol{z}$ is the radial proximal point and any omitted dominant branch would have been decoded instead, or the code branch is at least as large as the radial branch. In the latter case, if $\boldsymbol{z} \ne 0$, its angular correlation with $\boldsymbol{c}$ is at least κ and all other code branches are smaller. At $\boldsymbol{z} = 0$ all code branches are negative. Thus no omitted branch changes this prox. Mixing the radial function with one affine branch gives the same one-dimensional concave dual as in Theorem E.2. The prox of tH_θ preserves direction and maps $r = \|\boldsymbol{y}\|$ to r for $r \le \theta$, to θ for $\theta < r < \theta + t$, and to $r - t$ for $r \ge \theta + t$. The dual gap and strong convexity certify numerical accuracy.

For the exact formula within the radius-θ ball, put $u = \min\{s, 1/L\}$ and $\boldsymbol{z} = \boldsymbol{y} - u\boldsymbol{c}$. Since $u \le s < \langle \boldsymbol{c}, \boldsymbol{y}\rangle$, one has $\|\boldsymbol{z}\| \le \|\boldsymbol{y}\| \le \theta$. If $u = s$, the code branch is zero at $\boldsymbol{z}$, the radial branch is zero, and every other code branch is strictly negative. The vector $Ls\boldsymbol{c}$ is a convex combination of the zero gradient and $\boldsymbol{c}$. Hence $\boldsymbol{y} - \boldsymbol{z} = (Ls\boldsymbol{c})/L$ is the proximal condition. If $u = 1/L < s$, the code branch remains uniquely positive at $\boldsymbol{z}$ and $\boldsymbol{y} - \boldsymbol{z} = \boldsymbol{c}/L$ is its proximal condition. Evaluating the proximal objective gives the two values in (E.5). If no positive branch exists, the nonnegative function is minimized at $\boldsymbol{y}$, so its envelope and gradient vanish there. □

E.4 THE FIXED LOSSES AND THEIR GLOBAL PARAMETERS

Choose constants

$$A \geq \max\{20, 20/(\kappa - \rho\zeta)\}, \quad K = 128A, \quad \lambda = (100A)^{-1}, \quad C_0 \geq \max\{10, 16/\Delta\},$$
$$0 < \theta_0^2 \leq \min\{1/16, 1/(1000AC_0)\}. \tag{E.6}$$

For a prescribed starting horizon $q_0 \geq KC_0\sqrt{n}$, put $Q_0 = q_0/K$, $k = 2n$, and

$$a = \frac{1}{4n}, \quad \lambda_m = \frac{A}{\sqrt{n}}, \quad \theta = \theta_0 \min\{\sqrt{Q_0/n^{3/2}}, 1\}, \quad \lambda_c = \frac{C_0}{Q_0}, \quad b = \zeta\theta\lambda_c. \tag{E.7}$$

Use the decoder P from (E.2) and the fixed radial Huber function

$$J(\boldsymbol{z}) = \begin{cases} \|\boldsymbol{z}\|^2/2, & \|\boldsymbol{z}\| \leq 1, \\ \|\boldsymbol{z}\| - 1/2, & \|\boldsymbol{z}\| > 1. \end{cases}$$

Its gradient is the Euclidean projection onto the unit ball, so it is convex, nonnegative, globally 1-Lipschitz and 1-smooth.

Draw i uniformly from $[2n]$, and set $\boldsymbol{\delta}_i = n^{-1}\mathbf{1} - 2\boldsymbol{e}_i$. Then $\mathbb{E}_i\boldsymbol{\delta}_i = 0$ and $\|\boldsymbol{\delta}_i\| \leq 2$. The sample losses are

$$f_i(\boldsymbol{x}, \boldsymbol{y}) = K^{-1}\left[P_\lambda(\boldsymbol{x}, \boldsymbol{y}) + \lambda_m J(\boldsymbol{x}) + \lambda_c J(\boldsymbol{y}) + R_{i,\theta,L}(\boldsymbol{y}) - \langle \boldsymbol{\delta}_i, \boldsymbol{x}\rangle\right]. \tag{E.8}$$

Every term is fixed before sampling. The only data argument is i. The parameter dimension is $k + m = O(n)$.

The decoder's Lipschitz constant is $\sqrt{1/(8n) + b^2} < 1$. Also $\lambda_c \leq 1$, $\lambda_m \leq A$, and the readout is 1-Lipschitz. The raw Lipschitz constant is at most $A + 5 < K$. Its smoothness is at most $\lambda^{-1} + \lambda_m + \lambda_c + L \leq 101A + 1 + L \leq KL$ for $L \geq 1$. Thus the normalized losses meet all global assumptions. All nonlinear terms are nonnegative and vanish at zero, while the linear term has zero population mean. Consequently,

$$F^* = F(0) = 0, \qquad F(\boldsymbol{x}, \boldsymbol{y}) \geq K^{-1}\mathbb{E}_i R_{i,\theta,L}(\boldsymbol{y}) \quad \text{for every } (\boldsymbol{x}, \boldsymbol{y}). \tag{E.9}$$

This is a pointwise inequality, so it remains valid at data-dependent outputs.

E.5 REPEATED SAMPLES AND FINITE IDENTIFICATION

Let N_i count label i, and define

$$\mu_i = (1 - 2N_i)/n, \qquad \sigma_i = \operatorname{sign}(\mu_i), \qquad \mathcal{E} = \{\|\boldsymbol{\mu}\| \leq 4/\sqrt{n}\}.$$

Each μ_i is nonzero, $|\mu_i| \geq 1/n$, and $\sigma_i = +1$ exactly for unseen labels. The multinomial second moment is

$$\mathbb{E}\|\boldsymbol{\mu}\|^2 = 4/n - 2/n^2 \leq 4/n,$$

so $\Pr(\mathcal{E}) \geq 3/4$. In particular no distinct-label event is required.

Analyze the raw loss Kf_i, with raw time t. Actual time q corresponds to $Q = q/K$. Stop at the first time the trajectory reaches the unit sphere or a training readout becomes positive. Before that time, both Huber terms are quadratic, and all training readout gradients are zero. Every decoder-envelope gradient is a subgradient of P at its proximal point. Lemma E.1 therefore bounds each memory component by a and its code component in norm by b. Variation of constants gives, on $\mathcal{E}$,

$$|x_i(t)| \geq \frac{3}{4n\lambda_m}(1 - e^{-\lambda_m t}), \qquad \operatorname{sign} x_i(t) = \sigma_i \quad (t > 0), \tag{E.10}$$

$$\|\boldsymbol{x}(t)\| \leq \frac{\|\boldsymbol{\mu}\| + a\sqrt{k}}{\lambda_m} < 5/A, \qquad \|\boldsymbol{y}(t)\| \leq \zeta\theta(1 - e^{-\lambda_c t}) \leq \min\{bt, \zeta\theta\}. \tag{E.11}$$

The sign claim follows from the positive signed forcing $|\mu_i| - a \geq 3/(4n)$, not from an oracle convention.

Set $t_0 = 8\lambda$. The parameters imply

$$b^2 \leq C_0\theta_0^2/n^2 \leq 1/(1000An^2), \qquad b\zeta\theta \leq C_0\theta_0^2/n^{3/2}. \tag{E.12}$$

We now verify the exact identification certificate (E.3) for all $t \geq t_0$. If $t \leq 1/\lambda_m$, then

$$\min_i |x_i(t)| - \lambda a \geq \frac{3t}{8n} - \frac{\lambda}{4n} \geq \frac{t}{4n}.$$

The left side of the certificate is at least $t/(16n^2)$, whereas its right side is at most $2b^2(t+\lambda) \leq (9/4)b^2 t < t/(16n^2)$. If $t \geq 1/\lambda_m$, the memory bound gives

$$\min_i |x_i(t)| - \lambda a \geq \frac{3(1-e^{-1})}{4A\sqrt{n}} - \frac{1}{400An} > \frac{1}{3A\sqrt{n}}.$$

Thus the left side is at least $1/(12An^{3/2})$, while the right side is at most

$$2b(\zeta\theta + \lambda b) \leq 3C_0\theta_0^2/n^{3/2} < 1/(12An^{3/2}).$$

It follows that, for $t \geq t_0$,

$$\nabla P_\lambda(\boldsymbol{x}(t), \boldsymbol{y}(t)) = (a\boldsymbol{\sigma}, -b\boldsymbol{c}_{\boldsymbol{\sigma}}), \quad \boldsymbol{x}' = \boldsymbol{\mu} - a\boldsymbol{\sigma} - \lambda_m \boldsymbol{x}, \quad \boldsymbol{y}' = b\boldsymbol{c}_{\boldsymbol{\sigma}} - \lambda_c \boldsymbol{y}. \tag{E.13}$$

In particular,

$$\boldsymbol{y}(t) = \zeta\theta(1 - e^{-\lambda_c(t-t_0)})\boldsymbol{c}_{\boldsymbol{\sigma}} + e^{-\lambda_c(t-t_0)}\boldsymbol{y}(t_0). \tag{E.14}$$

The stopped argument now closes. Before t_0, $\|\boldsymbol{y}(t)\|/\theta \leq 8/(100A) < \kappa$, so every training readout is zero. Afterwards, a training label has $\sigma_i = -1$, so every codeword admissible in its readout differs from $\boldsymbol{c}_{\boldsymbol{\sigma}}$. Its correlation divided by θ is at most $\rho\zeta + 8/(100A) < \kappa$. Also $\|\boldsymbol{y}\| \leq \zeta\theta < \theta$, so the radial readout is zero. Finally

$$\|\boldsymbol{x}(t)\|^2 + \|\boldsymbol{y}(t)\|^2 < (5/A)^2 + \zeta^2\theta_0^2 < 1.$$

Neither stopping event can occur. These are therefore the actual constrained-flow bounds for all time on $\mathcal{E}$.

E.6 Every safe projected-GD step and the full average

For an actual step $0 < \eta \leq 1/L$, put $h = \eta/K$. Then $h \leq \lambda$, $h\lambda_m < 1$, and $h\lambda_c < 1$. An induction before identification gives

$$|x_{j,i}| \geq \frac{3}{4n\lambda_m}[1 - (1 - h\lambda_m)^j], \tag{E.15}$$

$$\|\boldsymbol{x}_j\| < 5/A, \qquad \|\boldsymbol{y}_j\| \leq \zeta\theta[1 - (1 - h\lambda_c)^j] \leq \min\{bjh, \zeta\theta\}. \tag{E.16}$$

The unprojected next candidate satisfies the same norm bounds, hence lies strictly in the ball on $\mathcal{E}$. Let $J = \lceil 8\lambda/h \rceil$ and $u = Jh \leq 9\lambda$. Since $1 - (1 - h\lambda_m)^j \geq 1 - e^{-\lambda_m jh}$, the two-case certificate above holds whenever $jh \geq 8\lambda$. Thus for all $j \geq J$ the actual updates are

$$\boldsymbol{x}_{j+1} = (1 - h\lambda_m)\boldsymbol{x}_j + h(\boldsymbol{\mu} - a\boldsymbol{\sigma}), \qquad \boldsymbol{y}_{j+1} = (1 - h\lambda_c)\boldsymbol{y}_j + hb\boldsymbol{c}_{\boldsymbol{\sigma}}. \tag{E.17}$$

Before J, $\|\boldsymbol{y}_j\|/\theta \leq 9/(100A) < \kappa$. Afterwards, with $\alpha = 1 - h\lambda_c$,

$$\boldsymbol{y}_{J+r} = \zeta\theta(1 - \alpha^r)\boldsymbol{c}_{\boldsymbol{\sigma}} + \alpha^r \boldsymbol{y}_J.$$

The same correlation calculation, now with $9/(100A)$, keeps all training readouts zero. This completes the induction for all iterates. There is no selector ambiguity.

Let $\tau_c = 1/\lambda_c = Q_0/C_0 \geq \sqrt{n}$, which is greater than 9λ. For either dynamics and every raw horizon $Q \geq Q_0$,

$$\langle \boldsymbol{c}_{\boldsymbol{\sigma}}, \bar{\boldsymbol{y}} \rangle \geq \zeta\theta\left(1 - \frac{4\tau_c}{Q}\right) \geq \kappa\theta + \Delta\theta/2. \tag{E.18}$$

For clarity, the discrete sum is explicit. The first J correlations sum to at least $-Jbu$. The subsequent sum is at least

$$\zeta\theta(T - J) - (\zeta\theta + bu)\sum_{r=1}^{T-J} \alpha^r \geq \zeta\theta(T - J) - \frac{\zeta\theta + bu}{h\lambda_c}.$$

Divide by T, use $Q = hT$, $b\tau_c = \zeta\theta$, and $u \le \tau_c$. The loss relative to $\zeta\theta$ is at most $\zeta\theta(\tau_c + 2u + u^2/\tau_c)/Q \le 4\zeta\theta\tau_c/Q$. For flow, integrate (E.14). The preheating integral has absolute value at most $bt_0^2/2$, and the same bound follows. The last inequality in (E.18) uses $C_0 \ge 16/\Delta$.

At least half of the labels are unseen. For each such label, $\boldsymbol{c}_{\boldsymbol{\sigma}}$ is an admissible readout code. The average stays in the radius-θ ball, so Lemma E.4 gives

$$R_{i,\theta,L}(\bar{\boldsymbol{y}}) \ge c \min\{L\theta^2, \theta\}.$$

Use (E.9), the unseen fraction, and $\Pr(\mathcal{E}) \ge 3/4$. The same fixed instance therefore satisfies, for all $q \ge q_0$ and every safe step,

$$\mathbb{E}[F(\boldsymbol{X}) - F^*] \ge c \min\left\{1, \frac{Lq_0}{n^{3/2}}, \sqrt{\frac{q_0}{n^{3/2}}}\right\}. \tag{E.19}$$

For the worst-case bound, take $q_0 = q$ when $q \ge KC_0\sqrt{n}$. If q is smaller and $L \le \sqrt{n}$, the additional term in (2) is at most a constant multiple of $n^{-1/2}$ and is absorbed by Appendix C.7. The optimization term there completes the lower side of Theorem 2.1. For $L = 1$, a fixed $q_0 = Cn^{3/2}$ gives a constant lower bound at every later time. Since a 1-smooth loss is admissible for every $L \ge 1$, this also provides the low-curvature sufficient side of Theorem M.1.

E.7 Matching last-iterate lower bounds

Lemma E.5 (Pointwise readout after the prescribed horizon)**.** *For the fixed losses (E.8), with any $n \ge 1$, $L \ge 1$ and prescribed $q_0 \ge KC_0\sqrt{n}$, on $\mathcal{E}$ every flow time $q \ge q_0$ and every GD pair $0 < \eta \le 1/L$, $T \ge 1$, $\eta T \ge q_0$ satisfy*

$$\langle \boldsymbol{c}_{\boldsymbol{\sigma}}, \boldsymbol{y}_{\mathrm{last}} \rangle \ge \kappa\theta + \Delta\theta/2, \qquad \|\boldsymbol{y}_{\mathrm{last}}\| \le \zeta\theta < \theta. \tag{E.20}$$

Thus the same single instance obeys (E.19) for its last iterate as well as its full average, simultaneously over those times and steps.

Proof. Write $\tau_c = 1/\lambda_c = Q_0/C_0$ and $Q = q/K$. For flow let $u = t_0$, and for GD let $u = Jh$. The preceding trajectory proof gives $u \le 9\lambda \le \tau_c$, $b\tau_c = \zeta\theta$, and $\|\boldsymbol{y}(u)\| \le bu$. In continuous time (E.14), and in discrete time (E.17) with $(1 - h/\tau_c)^{T-J} \le e^{-(Q-u)/\tau_c}$, imply

$$\langle \boldsymbol{c}_{\boldsymbol{\sigma}}, \boldsymbol{y}_{\mathrm{last}} \rangle \ge \zeta\theta - (\zeta\theta + bu)e^{-(Q-u)/\tau_c} \ge \zeta\theta[1 - 2e^{-(C_0-1)}].$$

Since $0 < \Delta < 1$ and $C_0 \ge 16/\Delta$, one has $2\zeta e^{-(C_0-1)} \le \Delta/2$. For example $e^{C_0-1} \ge C_0 - 1 \ge 15/\Delta$ already suffices. The norm bound is (E.11) or (E.16). This proves (E.20).

The margin in (E.20) permits the same unseen-label readout and parameter substitution used in (E.19), now for the last iterate. Choosing $q_0 = q$ and using the affine witnesses from Appendix C.7 at smaller horizons completes its low-curvature law, while fixing $q_0 = Cn^{3/2}$ gives persistence. □

E.8 Implementation of the lower-bound construction

For the family (E.8), fix the constructive code and rational constants satisfying (E.1) and (E.6) with strict slack. Such choices exist. The displayed square roots specify the remaining parameters exactly.

Theorem E.6 (Computing loss values and gradients)**.** *Let $n \ge 1$, let L, q_0 be positive rational parameters satisfying the conditions of (E.7), and let $\boldsymbol{w} \in \mathbb{B}^{O(n)}$ be rational. Suppose each rational input coordinate or parameter has at most B bits in its numerator and denominator. Given a label i and an integer $p \ge 1$, an algorithm returns dyadic $\widetilde{f}_i(\boldsymbol{w})$ and $\widetilde{\boldsymbol{g}}_i(\boldsymbol{w})$ satisfying*

$$|\widetilde{f}_i(\boldsymbol{w}) - f_i(\boldsymbol{w})| \le 2^{-p}, \qquad \|\widetilde{\boldsymbol{g}}_i(\boldsymbol{w}) - \nabla f_i(\boldsymbol{w})\| \le 2^{-p}, \tag{E.21}$$

in $\mathrm{poly}(n, B, p)$ bit operations and space. This holds at every such input, including branch boundaries, with at most one codeword encoding and one bounded-distance decoding call and no access to the training set or good event $\mathcal{E}$.

Lemma E.7 (Scalar search)**.** *Let H be convex and G-Lipschitz, with its proximal map and values computable to prescribed accuracy, and let $\ell(\boldsymbol{z}) = \langle \boldsymbol{g}, \boldsymbol{z} \rangle + c$, $\|\boldsymbol{g}\| \le G$. For $\lambda > 0$, $t \in [0,1]$, put*

$$\boldsymbol{z}(t) = \operatorname{prox}_{\lambda(1-t)H}(\boldsymbol{w} - \lambda t \boldsymbol{g}), \qquad \phi(t) = \min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{w}\|^2 + \lambda[(1-t)H(\boldsymbol{z}) + t\ell(\boldsymbol{z})] \right\}.$$

Then ϕ is concave, $\phi'(t) = \lambda[\ell(\boldsymbol{z}(t)) - H(\boldsymbol{z}(t))]$, and

$$\|\boldsymbol{z}(t) - \boldsymbol{z}(s)\| \le 2\lambda G|t - s|. \tag{E.22}$$

For proximal error $\delta > 0$, certified bisection with a dual-gap stopping test takes $O(1+\log(1+\lambda G/\delta))$ iterations, including nonunique and endpoint maximizers. Precision depends polynomially on logarithms of parameter magnitudes and $1/\delta$, not on a smallest nonzero derivative.

Proof. Strong convexity gives unique $\boldsymbol{z}(t)$. Compactness in t and coercivity in $\boldsymbol{z}$ justify the minimax identity. Differentiating the minimum gives the stated derivative. Comparing optimality inclusions proves (E.22): monotonicity absorbs the common coefficient, while its change and the affine shift each cost at most $\lambda G|t - s|$.

For $J(\boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{w}\|^2 + \lambda \max\{H(\boldsymbol{z}), \ell(\boldsymbol{z})\}$,

$$0 \le J(\boldsymbol{z}(t)) - \max_{u \in [0,1]} \phi(u) \le J(\boldsymbol{z}(t)) - \phi(t) \le |\phi'(t)|.$$

Its strong convexity bounds proximal error by $\sqrt{2|\phi'(t)|}$. Evaluate ϕ' with certified error $\delta^2/32$. A sign-definite interval maintains the bisection bracket, including one-sided endpoint tests. If it contains zero, $|\phi'| \le \delta^2/16$, giving error $\delta/\sqrt{8}$. Approximating $\boldsymbol{z}(t)$ within $\delta/4$ keeps total error below δ.

Otherwise stop at bracket width $\delta/(4\lambda G)$. The bracket contains a maximizer t^*, so (E.22) gives error $\delta/2$ from $\boldsymbol{z}(t^*)$, again allowing $\delta/4$ evaluation error. Derivative tolerance $\delta^2/32$ follows by computing $\boldsymbol{z}(t)$ within $\delta^2/(128\lambda(G+1))$ and the scalar values sufficiently accurately. For $G = 0$, the proximal point is $\boldsymbol{w}$ without search. □

Proof of Theorem E.6. Theorem E.2 uses the input's exact memory signs. For the readout, decode the exact input signs and retain the affine branch alongside H_θ only when its message has bit $i = +1$. Otherwise retain only H_θ. Lemma E.4 certifies this reduction without an angular-threshold comparison.

The radial masks use norms and clipping. For penalty scale $s \ge 0$, the memory mask subtracts projection onto $\{\|\boldsymbol{u}\|_\infty \le s,\ \|\boldsymbol{u}\|_1 \le s(k-1)\}$, whose absolute coordinates are

$$u_j = \min\{s, (|x_j| - \alpha)_+\}.$$

Use $\alpha = 0$ if feasible. Otherwise choose $\alpha \ge 0$ enforcing $\sum_j u_j = s(k-1)$ by sorting the $2k$ breakpoints. Nonexpansiveness controls input-rounding error even at tied breakpoints. The case $s = 0$ is immediate.

Apply Lemma E.7. For proximal point $\boldsymbol{z}^*$ and envelope radius λ_e,

$$\nabla h_{\lambda_e}(\boldsymbol{w}) = (\boldsymbol{w} - \boldsymbol{z}^*)/\lambda_e, \qquad h_{\lambda_e}(\boldsymbol{w}) = h(\boldsymbol{z}^*) + \|\boldsymbol{z}^* - \boldsymbol{w}\|^2/(2\lambda_e).$$

Proximal tolerance $O(2^{-p}\lambda_e)$ controls gradients. Evaluate values using the reduced objective, whose minimum equals the full objective's, avoiding an exponential maximum. If the reduced function is G-Lipschitz, $\|\boldsymbol{z}^* - \boldsymbol{w}\| \le \lambda_e G$, so proximal error δ changes that objective by at most

$$2G\delta + \delta^2/(2\lambda_e).$$

Reduce the tolerance accordingly and divide error among the fixed summands of (E.8). Huber and linear terms are explicit.

All positive parameter magnitudes and reciprocals in (E.6)–(E.7) have logarithms polynomial in $B + \log(2n)$. Dyadic intervals enclose their square roots and $1/\sqrt{m}$ to b bits in polynomial time. Proximal points lie within $\lambda_e G$ of shifted inputs, giving comparable logarithmic norm bounds. Coefficient error τ perturbs objectives on this region by τ times a polynomial in these magnitude bounds. Strong convexity bounds minimizer displacement by the square root of that perturbation. Thus $b = \operatorname{poly}(n, B, p)$ certifies values, derivatives, gap tests, and uncertain algebraic comparisons. Code operations are polynomial in n. □

Remark (Numerical robustness). Suppose the exact instance has expected excess at least $c > 0$ for both outputs whenever $q \geq q_0$, and $0 < \eta \leq 1/L$. Let implemented empirical-gradient norm errors be at most ε_j, and projection/output-rounding errors at most δ_j. If $\sum_{j\geq 1}(\eta\varepsilon_j + \delta_j) \leq c/2$, the last iterate and full post-update average retain expected excess at least $c/2$ at every such horizon. Cocoercivity and projection make exact updates nonexpansive, so distances from the common initial state accumulate by this sum. Averaging and population 1-Lipschitzness preserve the bound. Keep rational iterates feasible by rounding projections inward, charging this to δ_j. A schedule proportional to $(j+1)^{-2}$ suffices, requiring $O(\log(j+1))$ additional accuracy bits for fixed parameters. Each empirical gradient uses n sample-oracle calls. These complexity bounds are per update.

F Fast overfitting in polynomial dimension

We prove the q/n lower mechanism in Proposition M.2. The construction is independent of the final time and step size. The proof proceeds from subset geometry to finite compression, then to the actual trajectory. Lemma F.1 supplies geometry shared with Appendix H. The linear-dimensional construction in Appendix E is separate.

F.1 Multiscale subset geometry

Fix $n \geq 1$, $M = 4n^2, \mathcal{U} = \binom{[M]}{n}$, and $N = |\mathcal{U}|$. Let

$$\mathcal{K} = \{2^j : j \in \mathbb{Z}_{\geq 0},\ 2^j \leq n/2\} \cup \{n\}.$$

For $n = 1$, this is $\{1\}$ and all higher-layer sums below are empty. Consecutive levels $r < k$ satisfy $k \leq 4r$ and $r \leq n/2$, and

$$\sum_{r\in\mathcal{K}} \sqrt{r/n} < 4. \tag{F.1}$$

The uncompressed rth block has coordinates indexed by r-subsets of $[M]$. Define

$$\boldsymbol{v}_r(U) = \binom{|U|}{r}^{-1/2} \sum_{A\subseteq U,\ |A|=r} \boldsymbol{e}_A, \qquad \bar{\boldsymbol{v}}_r = N^{-1}\sum_{U\in\mathcal{U}} \boldsymbol{v}_r(U), \qquad \boldsymbol{p}_r(U) = \boldsymbol{v}_r(U) - \bar{\boldsymbol{v}}_r.$$

Lemma F.1 (Shared subset geometry)**.** *For $1 \leq r \leq n$ and $U \in \mathcal{U}$,*

$$\bar{\boldsymbol{v}}_r = \frac{\sqrt{\binom{n}{r}}}{\binom{M}{r}}\mathbf{1}, \quad \langle \boldsymbol{v}_r(U), \bar{\boldsymbol{v}}_r\rangle = \|\bar{\boldsymbol{v}}_r\|^2 = \frac{\binom{n}{r}}{\binom{M}{r}}, \quad \|\boldsymbol{p}_r(U)\| \leq 1. \tag{F.2}$$

For $1 \leq r < n$ and $i \in U$, define

$$\boldsymbol{E}_{r,i}(U) = \sqrt{\frac{n}{r}}\left[\boldsymbol{v}_r(U\setminus\{i\}) - \sqrt{\frac{n-r}{n}}\boldsymbol{v}_r(U)\right]. \tag{F.3}$$

Then

$$\|\boldsymbol{E}_{r,i}(U)\| = 1, \qquad \frac{1}{n}\sum_{i\in U}\boldsymbol{E}_{r,i}(U) = 0, \qquad \langle \boldsymbol{E}_{r,i}(U), \boldsymbol{v}_r(U)\rangle = \langle \boldsymbol{E}_{r,i}(U), \bar{\boldsymbol{v}}_r\rangle = 0. \tag{F.4}$$

For $S, U \in \mathcal{U}$, $i \in S \cap U$ and $\ell = |S\cap U| \geq 1$, set $\rho_r(\ell) = \binom{\ell}{r}/\binom{n}{r}$, with value zero for $r > \ell$. Then

$$\langle \boldsymbol{v}_r(U), \boldsymbol{v}_r(S)\rangle = \rho_r(\ell), \qquad -\langle \boldsymbol{E}_{r,i}(U), \boldsymbol{v}_r(S)\rangle = m_r(\ell) := \sqrt{\frac{r}{n-r}}\frac{n-\ell}{\ell}\rho_r(\ell). \tag{F.5}$$

The second identity and m_r are used only for $r < n$.

Proof. Each r-subset of $[M]$ belongs to $\binom{M-r}{n-r}$ members of $\mathcal{U}$. Counting its coefficient in the average gives (F.2). Also $\|\boldsymbol{p}_r(U)\|^2 = 1 - \|\bar{\boldsymbol{v}}_r\|^2 \leq 1$.

For (F.4), set $c_r = \sqrt{(n-r)/n}$. Each r-subset of U appears in exactly $n-r$ deleted-label sets, so $n^{-1}\sum_{i\in U}\boldsymbol{v}_r(U\setminus\{i\}) = c_r\boldsymbol{v}_r(U)$. Each summand has inner product c_r with $\boldsymbol{v}_r(U)$. The bracket

in (F.3) therefore has squared norm $1 - c_r^2 = r/n$ and is orthogonal to $\boldsymbol{v}_r(U)$. Its coordinate sum is $\sqrt{\binom{n-1}{r}} - c_r\sqrt{\binom{n}{r}} = 0$, proving orthogonality to $\bar{\boldsymbol{v}}_r$.

Counting shared coordinates proves the first overlap identity. For the second, the inner product before changing sign is

$$\sqrt{\frac{n}{r}}\left[\frac{\binom{\ell-1}{r}}{\sqrt{\binom{n-1}{r}\binom{n}{r}}} - \sqrt{\frac{n-r}{n}}\frac{\binom{\ell}{r}}{\binom{n}{r}}\right].$$

Use $\binom{\ell-1}{r} = (\ell - r)\binom{\ell}{r}/\ell$. Both terms vanish when $\ell < r$. □

Lemma F.2 (Multiscale branch margin)**.** *For every $k \in \mathcal{K} \setminus \{1\}$ and $1 \le \ell < n$,*

$$1 - \rho_k(\ell) \le 7n \sum_{r \in \mathcal{K},\ r<k} \sqrt{r/n}\, m_r(\ell). \tag{F.6}$$

Proof. First $1 - \rho_1(\ell) = (n - \ell)\rho_1(\ell)/\ell$. For consecutive levels $r < k' \le 4r$, interpret ρ_r as the probability that r draws without replacement all fall in a fixed ℓ-subset. Conditional on those successes, the chance of at least one failure in the next $k' - r$ draws is at most the expected number of failures. Thus

$$\rho_r(\ell) - \rho_{k'}(\ell) \le \rho_r(\ell)\frac{(k' - r)(n - \ell)}{n - r} \le 6r\frac{n - \ell}{\ell}\rho_r(\ell).$$

When $\rho_r = 0$, both subsequent probabilities are zero and the inequality still holds. Telescope across the levels and include the initial $1 - \rho_1$ term. This yields $1 - \rho_k \le 7\sum_{r<k} r(n - \ell)\rho_r/\ell$. Finally, $n\sqrt{r/n}\, m_r = r\sqrt{n/(n-r)}(n - \ell)\rho_r/\ell$ majorizes each summand. □

Lemma F.3 (Correlated branch differences)**.** *If $m = n - |S \cap U|$ and $i \in S \cap U$, then*

$$\|\boldsymbol{E}_{r,i}(S) - \boldsymbol{E}_{r,i}(U)\| \le \min\{2, 2\sqrt{rm/n}\}. \tag{F.7}$$

Proof. The coefficients of $\boldsymbol{E}_{r,i}(U)$ on coordinates containing i and not containing i are respectively $-\sqrt{(n-r)/r}/\sqrt{\binom{n}{r}}$ and $\sqrt{r/(n-r)}/\sqrt{\binom{n}{r}}$. Their total squared masses are $(n - r)/n$ and r/n. Common coordinates in the two supports have identical coefficients. A union bound on which deleted element a coordinate contains gives

$$1 - \langle \boldsymbol{E}_{r,i}(S), \boldsymbol{E}_{r,i}(U)\rangle \le \frac{m[(n - r)(r - 1) + r^2]}{n(n - 1)} \le \frac{2rm}{n}.$$

The vectors have norm one, so their squared difference is at most $4rm/n$. The triangle inequality gives the other bound. There is no $r < n$ case when $n = 1$. □

F.2 Simultaneous finite certificates and dimension reduction

Set $\epsilon = 1/(1000n)$. Leave level one unchanged, $\boldsymbol{P}_1 = \boldsymbol{I}$, and for every other level choose a linear map $\boldsymbol{P}_r$ preserving the norms of $\boldsymbol{v}_r(U)$, $\bar{\boldsymbol{v}}_r$, $\boldsymbol{p}_r(U)$ and $\boldsymbol{E}_{r,i}(U)$ within squared relative error ϵ. In addition require

$$|\langle \boldsymbol{P}_r\boldsymbol{a}, \boldsymbol{P}_r\boldsymbol{b}\rangle - \langle \boldsymbol{a}, \boldsymbol{b}\rangle| \le \epsilon\|\boldsymbol{a}\|\|\boldsymbol{b}\| \tag{F.8}$$

for the following finite pairs: (i) $\boldsymbol{a} = \boldsymbol{E}_{r,i}(S) - \boldsymbol{E}_{r,i}(U)$, $\boldsymbol{b} = \boldsymbol{p}_r(S)$, (ii) $\boldsymbol{a} = \boldsymbol{v}_r(S) - \boldsymbol{v}_r(U)$, $\boldsymbol{b} = \boldsymbol{p}_r(S)$, and (iii) at the top level, $\boldsymbol{a} = \boldsymbol{e}_U$, $\boldsymbol{b} = \boldsymbol{p}_n(S)$.

There are at most $C|\mathcal{K}|nN^2$ atoms and pairs. For each nonzero pair, apply a finite-set Johnson–Lindenstrauss embedding to its normalized sum and difference as well as the individual normalized vectors. Polarization gives (F.8). Reducing the JL accuracy by an absolute constant handles the desired coefficient. The standard finite-set bound (Dasgupta & Gupta, 2003) therefore gives each mapped block dimension

$$O(\epsilon^{-2}\log(C|\mathcal{K}|nN^2)) = O(n^3\log(2n)).$$

This can be seen directly from Gaussian squared-norm concentration and a union bound over this finite certificate set. Success has positive probability, so fix one successful collection of maps before any training sample is drawn. No randomness in the maps is averaged over in the lower bound. Including the unchanged $M = 4n^2$ first coordinates and $O(\log(2n))$ mapped blocks gives dimension $O(n^3 \log^2(2n))$. Importantly, $\sum_{i \in U} \boldsymbol{P}_r \boldsymbol{E}_{r,i}(U) = 0$ remains an exact identity, not an approximate one.

Take $\beta = 2^{-10}$ and

$$b_r = \frac{\beta}{n}(r/n)^{1/4}, \qquad a_{kr} = 14\beta(k/n)^{1/2}(r/n)^{1/4}, \quad r < k. \tag{F.9}$$

In the mapped coordinates $\boldsymbol{y} = (\boldsymbol{y}_r)_{r \in \mathcal{K}}$ define

$$\psi_{k,i}(\boldsymbol{y}) = \max_{U \in \mathcal{U}:\, i \in U} \left\{ \sum_{r<k} a_{kr} \langle \boldsymbol{P}_r \boldsymbol{E}_{r,i}(U), \boldsymbol{y}_r \rangle - b_k \langle \boldsymbol{P}_k \boldsymbol{v}_k(U), \boldsymbol{y}_k \rangle \right\}.$$

For the intended direction $\boldsymbol{y}_r = b_r \boldsymbol{P}_r \boldsymbol{p}_r(S)$, the uncompressed correct-minus-wrong margin is $M_k - B_k$, where

$$M_k = \sum_{r<k} a_{kr} b_r m_r(\ell), \qquad B_k = b_k^2 (1 - \rho_k(\ell)).$$

Lemma F.2 gives $M_k \geq 2B_k$. The compression error in this margin is at most

$$\epsilon \sum_{r<k} a_{kr} b_r \|\boldsymbol{E}_{r,i}(S) - \boldsymbol{E}_{r,i}(U)\| + \epsilon b_k^2 \sqrt{2(1 - \rho_k(\ell))}. \tag{F.10}$$

We verify that this is less than $B_k/2$, uniformly over all supports and levels. Let $m = n - \ell$ and $D = 1 - \rho_k(\ell)$. Sampling without replacement gives $\rho_k \leq (1 - m/n)^k \leq e^{-km/n}$, so $D \geq \frac{1}{2} \min\{km/n, 1\}$. Since $a_{kr} b_r = 14 n b_k^2 \sqrt{r/n}$, when $km/n \leq 1$ the first error divided by B_k is at most

$$\frac{14n\epsilon}{D} \sum_{r<k} \sqrt{r/n}\, 2\sqrt{rm/n} \leq \frac{56\epsilon k \sqrt{m}}{D} \leq 112\epsilon n.$$

Here $\sum_{r<k} r \leq 2k$ and $m \geq 1$. If $km/n \geq 1$, use the bound two in Lemma F.3 and (F.1). The ratio is at most $224\epsilon n$. The final error divided by B_k is $\epsilon\sqrt{2/D} \leq 2\epsilon n$. The total is at most $226\epsilon n B_k < B_k/2$. Hence the correct branch remains strictly active on the entire positive intended ray.

F.3 A CALIBRATED READOUT WITH A NONVANISHING MARGIN

Ordinary compression cannot preserve the exponentially small wrong-coordinate margin in a one-hot readout. Add one scalar parameter x and one independent Rademacher bit ξ to each sample. Data are $(i, \xi) \sim \mathrm{Unif}[M] \otimes \mathrm{Rad}$. Let $\kappa_0 = 1/16$, $\boldsymbol{c}_U = \boldsymbol{P}_n \boldsymbol{e}_U$, and

$$h_i(\boldsymbol{y}_n, x) = \max\{0, \max_{U \in \mathcal{U}:\, i \notin U} [\langle \boldsymbol{c}_U, \boldsymbol{y}_n \rangle - \kappa_0 x]\}. \tag{F.11}$$

Define the fixed loss

$$\begin{aligned} f(\boldsymbol{y}, x; (i, \xi)) = \frac{1}{4} h_i(\boldsymbol{y}_n, x) - \frac{\beta}{\sqrt{n}} \xi x - \sqrt{n} b_1 \langle \boldsymbol{e}_i - M^{-1}\mathbf{1}, \boldsymbol{y}_1 \rangle \\ + \sum_{k \in \mathcal{K} \setminus \{1\}} [\psi_{k,i}(\boldsymbol{y}) + b_k \langle \boldsymbol{P}_k \bar{\boldsymbol{v}}_k, \boldsymbol{y}_k \rangle]. \end{aligned} \tag{F.12}$$

Every term is linear or a finite maximum of linear functions. Thus the losses are globally finite, convex, and positively homogeneous.

We check global Lipschitzness on every branch. In block r, incoming gradients have total norm at most

$$\sum_{k>r} 2a_{kr} \leq 112\beta(r/n)^{1/4}.$$

The target branch and its mean correction add at most $4\beta(r/n)^{1/4}$. The first-level linear term is smaller. Across the orthogonal blocks this gives encoder norm at most $116\beta(\sum_r \sqrt{r/n})^{1/2} \le 232\beta$. The auxiliary scalar adds at most β. The readout has Lipschitz constant below two, so its coefficient $1/4$ adds at most $1/2$. Altogether

$$\mathrm{Lip}(f) \le 233\beta + 1/2 < 3/4 < 1. \tag{F.13}$$

This uses only bounds on the displayed atoms, not a global operator-norm bound for the embedding maps.

For $z = (i, \xi)$ and a sample $S = (z_1, \dots, z_n)$, write $p_i = \frac{1}{4}h_i$, $e_z = f(\cdot; z) - p_i$, and $e_S = n^{-1}\sum_{j=1}^n e_{z_j}$. Then

$$\mathrm{Lip}(e_z) < 1/4, \quad \mathrm{Lip}(p_i) \le 1/2, \quad p_i \ge 0, \quad \mathbb{E}_z e_z(\boldsymbol{w}) \ge 0, \quad F^* = F(0) = 0. \tag{F.14}$$

To prove the nontrivial population inequality, lower bound each $\psi_{k,i}$ by the uniform average of its branches over $U \ni i$. Choosing i uniformly and then $U \ni i$ uniformly is the same as choosing $U \in \mathcal{U}$ uniformly and then $i \in U$ uniformly. Equation (F.4) annihilates the memory term. The remaining expected target term is $-b_k\langle \boldsymbol{P}_k \bar{\boldsymbol{v}}_k, \boldsymbol{y}_k\rangle$, canceled by the fixed correction. Both added linear terms have zero population mean. Consequently

$$F(\boldsymbol{w}) \ge P(\boldsymbol{w}) := \mathbb{E}_i p_i(\boldsymbol{w}) \ge 0. \tag{F.15}$$

The auxiliary drift is thus generated by centered sample data, not by a hidden unpenalized population force.

F.4 A COMMON INITIAL POINT, THE EXACT TRAJECTORY, AND EXPECTED RISK

Let $\mathcal{E}$ be the event that all labels are distinct and $\sqrt{n}/2 \le Z := \sum_j \xi_j \le 4\sqrt{n}$. The label event has probability at least $7/8$. Since $\mathbb{E}Z^2 = n$ and $\mathbb{E}Z^4 \le 3n^2$, Paley–Zygmund gives $\Pr(|Z| \ge \sqrt{n}/2) \ge 3/16$. Markov gives $\Pr(|Z| > 4\sqrt{n}) \le 1/16$. Symmetry leaves probability at least $1/16$ in the positive band. The two parts are independent, so $\Pr(\mathcal{E}) \ge 1/32$.

On this event, write $S \in \mathcal{U}$ for the support and set

$$\nu = \frac{\beta Z}{n\sqrt{n}} \in [b_n/2, 4b_n], \qquad \boldsymbol{V}_S = ((b_r \boldsymbol{P}_r \boldsymbol{p}_r(S))_{r\in\mathcal{K}}, \nu),$$
$$a_S = \|\boldsymbol{V}_S\|, \quad \tau_S = 1/a_S. \tag{F.16}$$

The norm-preservation and level sum imply

$$b_n/2 \le a_S \le 5b_n, \qquad \tau_S > n. \tag{F.17}$$

The lower bound uses the scalar coordinate. The upper bound follows by summing at most $4(1+\epsilon)b_n^2$ from the mapped blocks and $16b_n^2$ from ν.

At $\boldsymbol{w} = \alpha \boldsymbol{V}_S$, $\alpha > 0$, the compressed encoder's unique winning branch for every training label and level is $U = S$, by (F.10). For a training label $i \in S$, every readout candidate has $U \ne S$. Its top-layer correlation is at most ϵ, and therefore its score is at most

$$\alpha(b_n\epsilon - \kappa_0\nu) \le \alpha b_n(\epsilon - 1/32) < 0.$$

The training readout is strictly zero. For an unseen label, $U = S$ is admissible, and its score is at least

$$\alpha b_n(1 - 1/N - \epsilon) - \alpha\kappa_0\nu \ge \alpha b_n/4.$$

Thus

$$P(\alpha \boldsymbol{V}_S) \ge \alpha b_n/32, \qquad p_{i_j}(\alpha \boldsymbol{V}_S) = 0 \quad \text{for every training point.} \tag{F.18}$$

The centered features cancel exactly in the empirical gradient. The first layer contributes $-b_1(\boldsymbol{v}_1(S) - \bar{\boldsymbol{v}}_1)$, higher layers contribute $-b_r \boldsymbol{P}_r(\boldsymbol{v}_r(S) - \bar{\boldsymbol{v}}_r)$, and the scalar contributes $-\nu$. Hence

$$\nabla F_S(\alpha \boldsymbol{V}_S) = \nabla e_S(\alpha \boldsymbol{V}_S) = -\boldsymbol{V}_S, \qquad \alpha > 0. \tag{F.19}$$

At zero all affine branches are active. Choosing the S branch gives $-\boldsymbol{V}_S \in \partial F_S(0)$ and $-\boldsymbol{V}_S \in \partial e_S(0)$. Positive homogeneity gives $F_S(\boldsymbol{V}_S) = -\|\boldsymbol{V}_S\|^2$. Every $\boldsymbol{g} \in \partial F_S(0)$ obeys $\langle \boldsymbol{g}, \boldsymbol{V}_S\rangle \le$

$-\|\boldsymbol{V}_S\|^2$, so $\|\boldsymbol{g}\| \geq \|\boldsymbol{V}_S\|$, with equality only at $\boldsymbol{g} = -\boldsymbol{V}_S$. The desired direction is therefore the unique minimum-norm empirical subgradient, not a sample-dependent initialization.

The actual constrained flow and minimum-norm-selector iterates are

$$\boldsymbol{u}_S(t) = \min\{t, \tau_S\}\boldsymbol{V}_S, \qquad \boldsymbol{w}_j = \min\{\eta j, \tau_S\}\boldsymbol{V}_S. \tag{F.20}$$

Before the boundary, (F.19) proves the flow inclusion. At the boundary, $\boldsymbol{V}_S$ is an outward normal and cancels the loss gradient. Uniqueness identifies the flow. The discrete formula follows by induction, since projecting the next positive ray point only truncates its scalar coefficient. It holds for every $\eta > 0$.

The minimum-norm selector is fixed on the entire data–state domain. For each finite dataset the objective is a finite max-affine function. Each of its finitely many Borel active-set regions has a fixed compact convex subdifferential, whose minimum-norm element is unique. Thus this selector is measurable. Every selected vector is bounded by (F.13).

Writing $\boldsymbol{A}_q = \alpha_q \boldsymbol{V}_S$, direct integration gives

$$\alpha_q = \begin{cases} q/2, & q \leq \tau_S, \\ \tau_S - \tau_S^2/(2q), & q > \tau_S, \end{cases} \qquad \alpha_q \geq \tfrac{1}{2}\min\{q, \tau_S\} \geq \tfrac{1}{2}\min\{q, n\}.$$

The discrete post-update coefficient is the right-endpoint Riemann sum for this nondecreasing scalar function, so it is no smaller. Equations (F.15) and (F.18) imply, on $\mathcal{E}$,

$$F(\boldsymbol{X}) - F^* \geq \frac{\beta}{64}\min\{q/n, 1\}.$$

Excess risk is nonnegative outside $\mathcal{E}$. Therefore, for one fixed instance and the stated selector,

$$\mathbb{E}[F(\boldsymbol{A}_q) - F^*], \quad \mathbb{E}[F(\bar{\boldsymbol{w}}_T) - F^*] \geq 2^{-21}\min\{q/n, 1\}, \tag{F.21}$$

simultaneously for all horizons and, in the discrete case, all steps. This proves the required new lower mechanism.

F.5 Upper bounds and the complete nonsmooth formula

We include the standard stability calculation to specify the oracle quantifiers precisely (Bassily et al., 2020). If S, S' differ in one sample and $\boldsymbol{g} \in \partial F_S(\boldsymbol{x})$, $\boldsymbol{g}' \in \partial F_{S'}(\boldsymbol{y})$ are arbitrary bounded ambient selections, two subgradient inequalities give

$$\langle \boldsymbol{g} - \boldsymbol{g}', \boldsymbol{x} - \boldsymbol{y}\rangle \geq F_S(\boldsymbol{x}) - F_S(\boldsymbol{y}) + F_{S'}(\boldsymbol{y}) - F_{S'}(\boldsymbol{x}) \geq -2\|\boldsymbol{x} - \boldsymbol{y}\|/n.$$

For the pathwise distance $D_j = \|\boldsymbol{w}_j(S) - \boldsymbol{w}_j(S')\|$, projection gives the recurrence

$$D_j^2 \leq D_{j-1}^2 + 4\eta D_{j-1}/n + 4\eta^2.$$

Induction yields $D_j \leq 2\eta j/n + 2\eta\sqrt{j}$. One way to verify the induction is to put $a = 2\eta/n$, $b = 2\eta$, and compare to $aj + b\sqrt{j}$: the increase in its square, less $2a(a(j-1) + b\sqrt{j-1})$, is at least b^2. Averaging and replace-one yield $\Gamma(\bar{\boldsymbol{w}}_T) \leq 2q/n + 2\eta\sqrt{T}$. The usual projected subgradient inequality summed at the pre-update iterates gives empirical comparison error at most $1/(2q) + \eta/2$. Moving each point to its updated version adds at most η, so the updated average has error at most $1/(2q) + 3\eta/2$. Since $\eta \leq \eta\sqrt{T}$, this gives the stated upper formula, capped by two. This proof permits arbitrary measurable dataset-dependent bounded selectors. It does not require consistent choices on shared samples.

For continuous flow, adding the two EVIs gives $\frac{1}{2}(\|\boldsymbol{u}_S - \boldsymbol{u}_{S'}\|^2)' \leq 2\|\boldsymbol{u}_S - \boldsymbol{u}_{S'}\|/n$. Regularize the norm by $\sqrt{\|\boldsymbol{u}_S - \boldsymbol{u}_{S'}\|^2 + \varepsilon^2}$, integrate, and let $\varepsilon \downarrow 0$. Thus the distance is at most $2t/n$ and the average distance at most q/n. Let $G_{n,d}(q)$ denote the worst-case expected excess risk of the full time average of constrained flow at time q, over globally convex, 1-Lipschitz losses in dimension at most d. Together with (C.6), this gives $G_{n,d}(q) \leq \min\{2, (2q)^{-1} + q/n\}$.

The affine optimization lower bound is in Appendix C.7. For the nonsmooth $\eta\sqrt{T}$ term, the published lower bound of Livni (2024, Theorem 1 and Corollary 2) applies in dimension $C(n + T^{1/3})$ for

$T \geq 10$, after a universal Lipschitz rescaling. If $T \leq n^2$, this is $O(n)$ and fits our dimension budget. If $T > n^2$, then $\eta\sqrt{T} \leq q/n$, so the new lower mechanism already absorbs it.

For completeness, bounded $T < 10$ needs no external construction. If $\eta < 1$, use $f(w) = |w|$ and the fixed selector -1 at zero. Iterates alternate between η and zero, and the average loss is at least $\eta/2$, enough up to an absolute factor. If $\eta \geq 1$, use $f(w) = |w - 3/4|$ on $[-1, 1]$. The iterates alternate between 1 and $\max\{1 - \eta, -1\}$, and their average is separated from $3/4$ by at least $1/12$. These deterministic losses are globally convex and 1-Lipschitz. Taking the maximum over the three lower mechanisms proves Proposition M.2. The statement does not assert that every selector attains all three mechanisms.

G Uniform smoothing and the curvature–time threshold

This section proves the comparison lemmas used by Appendix H and the exact time-n statement in Theorem M.1. Smoothing is always performed on the individual sample losses, before taking their empirical average.

G.1 Envelope facts and an oracle-uniform comparison

For a finite convex G-Lipschitz function h, write

$$h_\lambda(\boldsymbol{w}) = \inf_{\boldsymbol{z}} \left\{ h(\boldsymbol{z}) + \frac{\|\boldsymbol{w} - \boldsymbol{z}\|^2}{2\lambda} \right\}, \qquad \boldsymbol{z}_\lambda(\boldsymbol{w}) = \operatorname{prox}_{\lambda h}(\boldsymbol{w}).$$

The quadratic makes the minimizer unique. Its optimality condition is $(\boldsymbol{w} - \boldsymbol{z}_\lambda(\boldsymbol{w}))/\lambda \in \partial h(\boldsymbol{z}_\lambda(\boldsymbol{w}))$. Monotonicity of ∂h implies firm nonexpansiveness of the proximal map and its complementary map. Consequently

$$\nabla h_\lambda(\boldsymbol{w}) = \frac{\boldsymbol{w} - \boldsymbol{z}_\lambda(\boldsymbol{w})}{\lambda}, \qquad \operatorname{Lip}(\nabla h_\lambda) \leq 1/\lambda, \qquad \operatorname{Lip}(h_\lambda) \leq G.$$

The last assertion also follows from the optimality condition and the global subgradient bound. Choosing $\boldsymbol{z} = \boldsymbol{w}$ and then minimizing the Lipschitz lower estimate $h(\boldsymbol{z}) \geq h(\boldsymbol{w}) - G\|\boldsymbol{w} - \boldsymbol{z}\|$ gives

$$0 \leq h(\boldsymbol{w}) - h_\lambda(\boldsymbol{w}) \leq \lambda G^2/2. \tag{G.1}$$

If $h \geq 0$, the envelope is nonnegative and has exactly the same zero set. For a support function $h = \sigma_K$ with compact convex K, convex duality gives

$$h_\lambda(\boldsymbol{w}) = \max_{\boldsymbol{g} \in K} \{\langle \boldsymbol{g}, \boldsymbol{w} \rangle - \lambda\|\boldsymbol{g}\|^2/2\}, \qquad \nabla h_\lambda(\boldsymbol{w}) = \Pi_K(\boldsymbol{w}/\lambda). \tag{G.2}$$

These are classical Moreau identities (Parikh & Boyd, 2014; Bauschke & Combettes, 2017). They involve no logarithm of the number of exposed faces. In particular, none of the arguments below interchanges an envelope with a sample average.

Lemma G.1 (Bounded-oracle discretization)**.** *Let Φ be globally convex and 1-Lipschitz, let $\mathbf{u}$ be its unit-ball constrained flow from zero, and let $\boldsymbol{w}_j$ be projected subgradient descent with an arbitrary bounded ambient subgradient at each step. Then*

$$\|\boldsymbol{w}_j - \boldsymbol{u}(j\eta)\| \leq 2\eta\sqrt{j}, \qquad \left\| \frac{1}{T}\sum_{j=1}^{T} \boldsymbol{w}_j - \frac{1}{q}\int_0^q \boldsymbol{u}(t)\, dt \right\| \leq 3\eta\sqrt{T}. \tag{G.3}$$

Proof. The constrained flow has speed at most one. Indeed, at almost every differentiability time its velocity is the projection of the negative ambient subgradient onto the tangent cone, or equivalently the negative minimum-norm element of $\partial(\Phi + I_{\mathbb{B}})$. Its norm is bounded by that of a bounded ambient subgradient. This also follows by taking the limit of the unit-speed proximal interpolants.

Put $a = (j-1)\eta$, $b = j\eta$, and $E_j = \|\boldsymbol{w}_j - \boldsymbol{u}(b)\|^2$. Projection and the subgradient inequality imply

$$\|\boldsymbol{w}_j - \boldsymbol{u}(a)\|^2 \leq E_{j-1} - 2\eta[\Phi(\boldsymbol{w}_{j-1}) - \Phi(\boldsymbol{u}(a))] + \eta^2.$$

The EVI with comparator $\boldsymbol{w}_j$, integrated over $[a, b]$, gives

$$\|\boldsymbol{w}_j - \boldsymbol{u}(b)\|^2 - \|\boldsymbol{w}_j - \boldsymbol{u}(a)\|^2 \le 2\int_a^b [\Phi(\boldsymbol{w}_j) - \Phi(\boldsymbol{u}(t))]dt.$$

Adding the two inequalities yields

$$E_j \le E_{j-1} + \eta^2 + 2\eta[\Phi(\boldsymbol{w}_j) - \Phi(\boldsymbol{w}_{j-1})] + 2\int_a^b [\Phi(\boldsymbol{u}(a)) - \Phi(\boldsymbol{u}(t))]dt.$$

The second increment is at most $2\eta^2$, since one projected step moves at most η. The integral is at most η^2, since both Φ and $\boldsymbol{u}$ are 1-Lipschitz. Thus $E_j \le E_{j-1} + 4\eta^2$, proving the first assertion. Averaging these distances and comparing the flow integral with its right-endpoint Riemann sum gives

$$\frac{2\eta}{T}\sum_{j=1}^{T}\sqrt{j} + \frac{\eta}{2} \le 3\eta\sqrt{T}.$$

No compatibility between the choices of subgradient on different data sets is used. □

G.2 A UNIFORM-IN-TIME RADIAL PERTURBATION BOUND

The geometry of the ball supplies a growth inequality even when the sample losses are not strongly convex. Suppose a reference objective Φ is sublinear and

$$-\boldsymbol{V} \in \partial\Phi(0), \quad \Phi(\alpha\boldsymbol{V}) = -\alpha\|\boldsymbol{V}\|^2 \ (\alpha \ge 0).$$

Let $a = \|\boldsymbol{V}\| > 0$, $\tau = 1/a$, and $\boldsymbol{v} = \boldsymbol{V}/a$. Its constrained flow is $\boldsymbol{u}(t) = \min\{t, \tau\}\boldsymbol{V}$. For every $\boldsymbol{w} \in \mathbb{B}$,

$$\Phi(\boldsymbol{w}) - \Phi(\boldsymbol{v}) \ge a(1 - \langle \boldsymbol{v}, \boldsymbol{w}\rangle) \ge \frac{a}{2}\|\boldsymbol{w} - \boldsymbol{v}\|^2. \tag{G.4}$$

The first inequality is its supporting plane at zero. The second uses $\|\boldsymbol{w}\| \le 1$. Write $A_q(\boldsymbol{u}) = q^{-1}\int_0^q \boldsymbol{u}(t)\,dt$ for the full time average.

Lemma G.2 (Uniform two-sided approximation)**.** *In the preceding setting, assume Φ is 1-Lipschitz and Ψ is convex with $0 \le \Phi - \Psi \le \lambda/2$ on the ball. Its constrained flow $\boldsymbol{x}$ from zero obeys*

$$\|\boldsymbol{x}(t) - \boldsymbol{u}(t)\|^2 \le \lambda \min\{t, \tau\}. \tag{G.5}$$

If Ψ is also 1-Lipschitz and $1/\lambda$-smooth, $\lambda \le \tau$, and projected GD uses $\eta \le \lambda$, then

$$\|\boldsymbol{x}_j - \boldsymbol{u}(j\eta)\|^2 \le 18\lambda \min\{j\eta, \tau\}, \tag{G.6}$$

$$\|\bar{\boldsymbol{x}}_T - A_q(\boldsymbol{u})\| \le \sqrt{18\lambda \min\{q, \tau\}} + a\eta/2. \tag{G.7}$$

Proof. Adding the two EVIs gives $(\|\boldsymbol{x} - \boldsymbol{u}\|^2)' \le \lambda$. Hence the flow bound holds up to τ. After τ, (G.4) and the approximation give

$$(\|\boldsymbol{x} - \boldsymbol{v}\|^2)' \le -a\|\boldsymbol{x} - \boldsymbol{v}\|^2 + \lambda.$$

At time τ its squared distance is at most λ/a. Solving this scalar differential inequality proves (G.5) for all future times.

Before the first grid point past τ, Lemma G.1 applied to Ψ and the flow bound give $\|\boldsymbol{x}_j - \boldsymbol{u}(j\eta)\| \le 2\sqrt{\eta j\eta} + \sqrt{\lambda j\eta} \le 3\sqrt{\lambda j\eta}$. At the crossing index $j_0 = \lceil \tau/\eta \rceil$, $j_0\eta \le 2\tau$, and the sharper use of the saturated flow bound yields

$$\|\boldsymbol{x}_{j_0} - \boldsymbol{v}\|^2 \le (2\sqrt{2} + 1)^2\lambda\tau < 18\lambda\tau.$$

For $j > j_0$, the smooth projected three-point inequality of Appendix C, with comparator $\boldsymbol{v}$, gives

$$(1 + \eta a)\|\boldsymbol{x}_j - \boldsymbol{v}\|^2 \le \|\boldsymbol{x}_{j-1} - \boldsymbol{v}\|^2 + \eta\lambda.$$

This preserves the bound $18\lambda/a$, since that bound exceeds λ/a. Averaging the pointwise distances, and using the a-Lipschitz reference path to bound its right-endpoint quadrature error by $a\eta/2$, proves the second assertion. □

G.3 THE EXACT TIME-n SMOOTHNESS THRESHOLD

Apply a Moreau envelope with parameter $\lambda = c_*/n$ to each of the full sample losses in Appendix F, with a fixed sufficiently small universal $c_* > 0$. They are globally 1-Lipschitz, so their envelopes are globally 1-Lipschitz and n/c_*-smooth. For each data set, the empirical function changes by between zero and $\lambda/2$ everywhere. On the good event of that construction, the original flow is the reference ray above, $\tau > n$, and

$$F(\boldsymbol{A}_n) - F^\star \geq \beta/64, \qquad \beta = 2^{-10}.$$

The population optimum after smoothing satisfies $F^\star - \lambda/2 \leq F_\lambda^\star \leq F^\star$. For the smoothed flow at exactly $q = n$, and also for GD on that time grid with $\eta \leq \lambda$, Lemma G.2 implies

$$F_\lambda(\bar{\boldsymbol{x}}) - F_\lambda^\star \geq \beta/64 - \sqrt{18c_*} - a\lambda/2 - \lambda/2.$$

For example $c_* = \beta^2/2^{24}$ is more than sufficiently small to keep the last three terms below $\beta/128$. The same conclusion for the flow follows from the smaller error in (G.5). Multiplying by the good-event probability at least $1/32$ proves a constant lower bound at the exact horizon n, in dimension $d_n = O(n^3 \log^2(2n))$.

Conversely, (C.12) and Lipschitz readout give in every dimension

$$\mathcal{R}_{n,d,L}(\boldsymbol{A}_n) \leq \frac{1}{2n} + \sqrt{\frac{2}{\sqrt{n}} + \frac{L}{n}}.$$

If this risk exceeds a fixed $r_0 > 0$, then $L \geq c(r_0)n$ for all sufficiently large n. The discrete version follows from (C.13). For a fixed sufficiently small r_0, these two directions prove (M.2). The sufficient direction here is at $q = n$, not merely at an unspecified constant multiple of n.

G.4 NONNEGATIVE PENALTIES THAT VANISH ON A REFERENCE PATH

Uniformly approximating the entire loss is unnecessarily expensive for the general time threshold. The following one-sided comparison keeps a possibly large penalty because it helps, rather than harms, the comparison.

Lemma G.3 (Inactive-penalty comparison)**.** *On the Euclidean unit ball, suppose $r > 0$, $\Phi(0) = 0$, $-r\boldsymbol{v} \in \partial\Phi(0)$, $\|\boldsymbol{v}\| = 1$, and the reference flow is $\boldsymbol{u}(t) = \min\{rt, 1\}\boldsymbol{v}$, with $\Phi(\alpha\boldsymbol{v}) = -r\alpha$ for $0 \leq \alpha \leq 1$. Let*

$$\Psi = \widetilde{\Phi} + P_S, \quad 0 \leq \Phi - \widetilde{\Phi} \leq \varepsilon, \quad P_S \geq 0, \quad P_S(\boldsymbol{u}(t)) = 0.$$

All functions are convex. Then its flow $\boldsymbol{x}$ from zero satisfies

$$\|\boldsymbol{x}(t) - \boldsymbol{u}(t)\|^2 \leq 2\varepsilon \min\{t, 1/r\}. \tag{G.8}$$

If Ψ is ℓ-smooth, projected GD from $\boldsymbol{x}_0 = 0$ with $0 < h \leq 1/\ell$ satisfies the following bounds at every grid time, with $q = hT$.

$$\|\boldsymbol{x}_j - \boldsymbol{u}(jh)\|^2 \leq E_0 := (2\varepsilon + hr^2)(1/r + h), \quad \|\bar{\boldsymbol{x}}_T - A_q(\boldsymbol{u})\| \leq \sqrt{E_0} + hr/2. \tag{G.9}$$

Proof. Write $a(\boldsymbol{w}) = \Phi(\boldsymbol{w}) - \widetilde{\Phi}(\boldsymbol{w}) \in [0, \varepsilon]$. The two EVIs give

$$\tfrac{1}{2}(\|\boldsymbol{x} - \boldsymbol{u}\|^2)' \leq a(\boldsymbol{x}) - a(\boldsymbol{u}) + P_S(\boldsymbol{u}) - P_S(\boldsymbol{x}) \leq \varepsilon.$$

At the boundary point $\boldsymbol{v}$, the support plane gives $\Phi(\boldsymbol{w}) - \Phi(\boldsymbol{v}) \geq r\|\boldsymbol{w} - \boldsymbol{v}\|^2/2$. Hence $\Psi(\boldsymbol{w}) - \Psi(\boldsymbol{v}) \geq r\|\boldsymbol{w} - \boldsymbol{v}\|^2/2 - \varepsilon$. For $t \geq 1/r$, the squared distance $D(t) = \|\boldsymbol{x}(t) - \boldsymbol{v}\|^2$ obeys $D'(t) \leq -rD(t) + 2\varepsilon$, with $D(1/r) \leq 2\varepsilon/r$. This proves the flow bound.

For the discrete statement set $\boldsymbol{u}_j = \min\{rhj, 1\}\boldsymbol{v}$ and $j_0 = \lceil 1/(rh) \rceil$. Before and at j_0, write $\boldsymbol{u}_j - \boldsymbol{u}_{j-1} = rs_j\boldsymbol{v}$, where $0 \leq s_j \leq h$. The smooth three-point inequality and the reference support plane yield

$$\|\boldsymbol{x}_j - \boldsymbol{u}_{j-1}\|^2 \leq \|\boldsymbol{x}_{j-1} - \boldsymbol{u}_{j-1}\|^2 + 2hr\langle \boldsymbol{v}, \boldsymbol{x}_j - \boldsymbol{u}_{j-1}\rangle + 2h\varepsilon.$$

The penalty is zero at $\boldsymbol{u}_{j-1}$ and nonnegative at $\boldsymbol{x}_j$, which is the required direction for this inequality. Subtracting the reference movement gives

$$\begin{aligned}\|\boldsymbol{x}_j - \boldsymbol{u}_j\|^2 \leq \|\boldsymbol{x}_{j-1} - \boldsymbol{u}_{j-1}\|^2 + 2h\varepsilon \\ + 2(h - s_j)r\langle \boldsymbol{v}, \boldsymbol{x}_j - \boldsymbol{u}_j\rangle + (2hs_j - s_j^2)r^2.\end{aligned}$$

If $s_j = h$, the inner-product term is zero. If $s_j < h$, the reference has reached $\boldsymbol{v}$ and that term is nonpositive. Thus the last two terms are at most $h^2 r^2$. Since $j_0 h \le 1/r + h$, summing bounds every pre-boundary distance, including the crossing step, by E_0.

For $j > j_0$, the three-point inequality and radial growth give

$$(1 + hr)\|\boldsymbol{x}_j - \boldsymbol{v}\|^2 \le \|\boldsymbol{x}_{j-1} - \boldsymbol{v}\|^2 + 2h\varepsilon.$$

Because $E_0 \ge 2\varepsilon/r$, this recurrence preserves the bound. The reference is r-Lipschitz, so averaging introduces at most $hr/2$ of additional quadrature error. □

G.5 Completion of the constant-risk threshold

For $L \le \sqrt{n}$, the low-curvature bounds in Appendices C and E give constant late risk at $q \asymp n^{3/2}$ using linear dimension. For $L \ge \sqrt{n}$, Lemma H.1 with $k = n$ gives a persistent constant lower bound at $q \gtrsim \max\{n, n^2/L\}$. The matching upper follows from (H.2) and the low-curvature law, restricting to $q \ge n$. This proves (M.1) for both outputs and flow. On a fixed safe GD grid, rounding a prescribed threshold up adds at most $\eta \le 1$, absorbed in the universal constants.

For the exact time-n assertion above, the nonsmooth reference last point has at least the full-average readout: the readout is nonnegative and positively homogeneous along the ray. Lemma G.2 controls the last-point displacement with no quadrature term, and Appendix C.8 gives the converse upper. Thus the exact-time statement holds for both outputs. The small choice of r_0 is universal. No risk monotonicity is assumed.

H Closing the full curvature–time law

This appendix proves Lemma H.1 and the high-curvature half of Theorem 2.1. The new construction truncates the subset encoder at an integer degree k and replaces its population readout by a polar one. The averaged supporting normal, explicitly included in the compression certificate, has norm $O(\sqrt{k/n})$. This is what preserves subconstant risk after smoothing. The low-curvature half uses Appendix E. After the upper bound, the lower proof has four tasks: truncate the shared geometry, preserve the small polar normal under compression, track the smoothed trajectory through the boundary, and optimize the integer degree. Only the shared geometry and the smoothing comparison are imported from Appendices F–G.

For the formal construction statement, put $r_k = \sqrt{k/n}$.

Lemma H.1 (Persistent tunable lower bound)**.** *For every $n \ge 2$, integer $1 \le k \le n$, and $L \ge \sqrt{n}$, one globally convex, 1-Lipschitz, L-smooth loss family in $O(n^3 \log^2(2n))$ dimensions, fixed before sampling, satisfies*

$$q \ge C \max\{n r_k, n^2 r_k^2 / L\} \quad \Longrightarrow \quad \mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] \ge c r_k. \tag{H.1}$$

The constants are universal. The same instance works from zero on the unit ball for both complete outputs, every $0 < \eta \le 1/L$, every integer T meeting the condition with $q = \eta T$, and constrained flow at every such time.

H.1 The high-curvature upper bound

For all four outputs, Theorem C.2, (C.12), the optimization bounds (C.5)–(C.6), and Lemma C.3 give, by replace-one symmetrization and Lipschitzness,

$$\mathbb{E}[F(\boldsymbol{X}) - F^*] \le (2q)^{-1} + \sqrt{24Lq}/n \qquad (L \ge \sqrt{n}).$$

Taking the minimum with (C.20) and capping by two proves

$$\mathbb{E}[F(\boldsymbol{X}) - F^*] \lesssim \min\{1, q^{-1} + \min(q/n, \sqrt{Lq}/n)\}. \tag{H.2}$$

H.2 A truncated subset geometry

Fix $n \ge 2$ and an integer $1 \le k \le n$. Set

$$M = 4n^2, \qquad \mathcal{U} = \binom{[M]}{n}, \qquad N = |\mathcal{U}|, \qquad \mathcal{K}_k = \{2^j : j \in \mathbb{Z}_{\ge 0},\ 2^j \le k/2\} \cup \{k\}.$$

For $k = 1$ the level set is $\{1\}$. Consecutive levels $r < j$ have $j \le 4r$, and

$$\sum_{r \in \mathcal{K}_k} \sqrt{r/k} < 4. \tag{H.3}$$

Use the normalized subset vectors $\boldsymbol{v}_r$, their centered versions $\boldsymbol{p}_r$, and the deleted-label features $\boldsymbol{E}_{r,i}$ from Lemma F.1. The lemma applies to every $1 \le r \le n$, independently of the chosen level set, and gives the centering and overlap identities (F.2)–(F.5). In particular, $\boldsymbol{E}_{r,i}$ and m_r are used only for $r < n$. The top level $r = n$ never requires them. Use $\rho_r(\ell) = \binom{\ell}{r}/\binom{n}{r}$ also at $\ell = 0$, and set $\rho_0 \equiv 1$. The change here is the truncated level set $\mathcal{K}_k$.

Lemma H.2 (Truncated multiscale certificates)**.** *For $j \in \mathcal{K}_k \setminus \{1\}$ and $1 \le \ell < n$,*

$$1 - \rho_j(\ell) \le 7n \sum_{r \in \mathcal{K}_k :\, r < j} \sqrt{r/n}\, m_r(\ell). \tag{H.4}$$

If $m = n - |S \cap U|$ and $i \in S \cap U$, then

$$\|\boldsymbol{E}_{r,i}(S) - \boldsymbol{E}_{r,i}(U)\| \le \min\{2, 2\sqrt{rm/n}\}. \tag{H.5}$$

Proof. For consecutive levels $r < j' \le 4r$, the probability interpretation of ρ_r as r successes without replacement gives

$$\rho_r - \rho_{j'} \le \rho_r \frac{(j' - r)(n - \ell)}{n - r} \le 6r \frac{n - \ell}{\ell} \rho_r.$$

Here $r \le n/2$. Start with $1 - \rho_1 = (n - \ell)\rho_1/\ell$ and telescope. Since $n\sqrt{r/n}\, m_r = r\sqrt{n/(n-r)}(n - \ell)\rho_r/\ell$, this proves (H.4), also when the top degree is not a power of two.

The chord bound (H.5) is Lemma F.3: its coefficient calculation applies to every $r < n$ and does not depend on the level set. □

H.3 Polar separation and the small averaged norm

Let

$$a_k = \sqrt{\frac{k}{n+1}}, \qquad b_k^{\mathrm{pol}} = \sqrt{\frac{n-k+1}{n+1}}.$$

For $U \in \mathcal{U}$, $i \notin U$, define

$$\boldsymbol{u}_{U,i} = \binom{n}{k-1}^{-1/2} \sum_{A \subseteq U :\, |A| = k-1} \boldsymbol{e}_{A \cup \{i\}}, \qquad \boldsymbol{c}_{U,i} = a_k \boldsymbol{v}_k(U) - b_k^{\mathrm{pol}} \boldsymbol{u}_{U,i}.$$

The two summands have disjoint supports, so $\|\boldsymbol{c}_{U,i}\| = 1$.

Lemma H.3 (Polar separation and cancellation)**.** *Every $\boldsymbol{c}_{U,i}$ has coordinate sum zero and is orthogonal to $\bar{\boldsymbol{v}}_k$. If $i \in S$ and $i \notin U$, then*

$$\langle \boldsymbol{c}_{U,i}, \boldsymbol{p}_k(S) \rangle = a_k[\rho_k(|S \cap U|) - \rho_{k-1}(|S \cap U|)] \le 0. \tag{H.6}$$

If $i \notin S$, then $\langle \boldsymbol{c}_{S,i}, \boldsymbol{p}_k(S) \rangle = a_k$. For $\boldsymbol{C}_S = M^{-1} \sum_{i \notin S} \boldsymbol{c}_{S,i}$ and $\gamma = (M - n)/M$,

$$\|\boldsymbol{C}_S\|^2 = \gamma^2 a_k^2 + (b_k^{\mathrm{pol}})^2 \frac{M - n}{M^2} \le a_k^2 + M^{-1} \le \tfrac{3}{2} a_k^2. \tag{H.7}$$

Proof. The coordinate sum is $a_k\sqrt{\binom{n}{k}} - b_k^{\mathrm{pol}}\sqrt{\binom{n}{k-1}} = 0$, by the ratio $(n-k+1)/k$ of the binomial coefficients. For a training label, with $\ell = |S \cap U|$,

$$\langle \boldsymbol{u}_{U,i}, \boldsymbol{v}_k(S) \rangle = \frac{\binom{\ell}{k-1}}{\sqrt{\binom{n}{k-1}\binom{n}{k}}}.$$

Multiplication by b_k^{pol} turns this into $a_k \rho_{k-1}(\ell)$. An additional draw cannot increase the all-success probability, proving the sign. For an unseen label, $\boldsymbol{u}_{S,i}$ is orthogonal to $\boldsymbol{v}_k(S)$. Finally, $\boldsymbol{v}_k(S)$ and all $\boldsymbol{u}_{S,i}$, $i \notin S$, are mutually orthogonal: each latter vector has a different unique index outside S. Expanding their average proves the exact identity. The last bound uses $M = 4n^2$ and $a_k^2 \ge 1/(n+1)$. □

Individual normals have norm one. Their average has norm $O(a_k)$. Replacing its norm by the average of the individual norms would remove the gain needed below.

H.4 Finite compression, fixed before sampling

Set $\epsilon = (1000n)^{-1}$ and keep $\boldsymbol{P}_1 = \boldsymbol{I}$. For each other level choose a fixed linear map $\boldsymbol{P}_r$ preserving, within squared relative error ϵ, the norms of $\boldsymbol{v}_r(U)$, $\bar{\boldsymbol{v}}_r$, $\boldsymbol{p}_r(U)$ and all required $\boldsymbol{E}_{r,i}(U)$. Require also

$$|\langle \boldsymbol{P}_r \boldsymbol{a}, \boldsymbol{P}_r \boldsymbol{b}\rangle - \langle \boldsymbol{a}, \boldsymbol{b}\rangle| \le \epsilon \|\boldsymbol{a}\|\|\boldsymbol{b}\| \tag{H.8}$$

for the pairs $(\boldsymbol{E}_{r,i}(S) - \boldsymbol{E}_{r,i}(U), \boldsymbol{p}_r(S))$ and $(\boldsymbol{v}_r(S) - \boldsymbol{v}_r(U), \boldsymbol{p}_r(S))$ whenever defined. At the top level include every $\boldsymbol{c}_{U,i}$ and every $\boldsymbol{C}_S$ in the norm list, and every pair $(\boldsymbol{c}_{U,i}, \boldsymbol{p}_k(S))$ in (H.8).

Each mapped level can have $O(n^3 \log(2n))$ coordinates. Indeed, the atom and pair count is at most N^2 times a universal polynomial in $M, n, \log(2n)$. For each nonzero pair add its normalized vectors and their sum and difference to a finite norm list. A Gaussian map with s rows preserves a fixed squared norm to relative error δ with failure probability at most $2\exp(-c\delta^2 s)$, by the chi-square moment generating function. A union bound and polarization give the finite certificates with

$$s = O\big(\epsilon^{-2} \log[N^2 \operatorname{poly}(M, n)]\big) = O(n^3 \log(2n)),$$

since $\log N \le n \log(4en)$. This is the finite-set dimension reduction used in Appendix F. No infinite-set or global operator-norm assertion is used. Fix one successful set of maps before data are drawn. Including the M unchanged first coordinates and one scalar, the total dimension is at most

$$D_n = \lceil C_d n^3 \log^2(2n) \rceil. \tag{H.9}$$

At $k = 1$ all certificates hold exactly. The small averaged norm survives because $\boldsymbol{C}_S$ itself is in the list. Linearity preserves exact cancellations such as $\sum_{i \in U} \boldsymbol{P}_r \boldsymbol{E}_{r,i}(U) = 0$.

H.5 The encoder and readout losses

Fix $\beta = 2^{-10}$. For $r < j$ in $\mathcal{K}_k$, set

$$b_r = \frac{\beta}{\sqrt{nk}} (r/k)^{1/4}, \qquad A_{jr} = 14\beta (j/k)^{1/2} (r/k)^{1/4},$$
$$A_{jr} b_r = 14 n b_j^2 \sqrt{r/n}. \tag{H.10}$$

The speed $b_k = \beta/\sqrt{nk}$ is distinct from b_k^{pol}. For block parameter $\boldsymbol{y} = (\boldsymbol{y}_r)_{r \in \mathcal{K}_k}$, let

$$\psi_{j,i}(\boldsymbol{y}) = \max_{U \in \mathcal{U}:\, i \in U} \left\{ \sum_{r<j} A_{jr} \langle \boldsymbol{P}_r \boldsymbol{E}_{r,i}(U), \boldsymbol{y}_r \rangle - b_j \langle \boldsymbol{P}_j \boldsymbol{v}_j(U), \boldsymbol{y}_j \rangle \right\}.$$

Data are $z = (i, \xi) \sim \mathrm{Unif}[M] \otimes \mathrm{Rad}$, and $x \in \mathbb{R}$ is one additional parameter. Define

$$e_z(\boldsymbol{y}, x) = -\frac{\beta}{\sqrt{k}} \xi x - \sqrt{n}\, b_1 \langle \boldsymbol{e}_i - M^{-1}\mathbf{1}, \boldsymbol{y}_1 \rangle + \sum_{j \in \mathcal{K}_k \setminus \{1\}} [\psi_{j,i}(\boldsymbol{y}) + b_j \langle \boldsymbol{P}_j \bar{\boldsymbol{v}}_j, \boldsymbol{y}_j \rangle], \tag{H.11}$$

$$p_i(\boldsymbol{y}, x) = \frac{1}{4} \max\left\{0, \max_{U \in \mathcal{U}:\, i \notin U} [\langle \boldsymbol{P}_k \boldsymbol{c}_{U,i}, \boldsymbol{y}_k \rangle - \delta a_k x]\right\}, \qquad \delta = 1/16. \tag{H.12}$$

The nonsmooth loss is $f_z = e_z + p_i$. All these functions are finite, convex and positively homogeneous.

Lemma H.4 (Global regularity and calibration)**.** *One has* $\mathrm{Lip}(e_z) < 1/4$, $\mathrm{Lip}(p_i) \le 1/2$, $p_i \ge 0$, *and* $\mathbb{E} e_Z(\boldsymbol{w}) \ge 0$ *throughout the ambient space. Thus* $F(\boldsymbol{w}) \ge \mathcal{P}(\boldsymbol{w}) := \mathbb{E}_i p_i(\boldsymbol{w}) \ge 0$ *and* $F^* = F(0) = 0$ *for the nonsmooth loss.*

Proof. Incoming branch gradients at block r have total norm at most $\sum_{j>r} 2A_{jr} \le 112\beta(r/k)^{1/4}$. The target and centering terms add at most $4\beta(r/k)^{1/4}$. At level one the linear term is at most

$\sqrt{n}b_1 \le \beta k^{-1/4}$, within the same bound. By (H.3), the norm over the orthogonal blocks is at most 232β. The scalar adds at most β. Hence $\mathrm{Lip}(e_z) \le 233\beta < 1/4$. Each readout atom has norm at most $\sqrt{1+\epsilon+\delta^2 a_k^2} < 2$, proving its Lipschitz bound.

Lower bound each $\psi_{j,i}$ by the uniform average over its branches $U \ni i$. Drawing i uniformly and then such a U is equivalent to drawing U uniformly and then $i \in U$ uniformly. Averaging cancels the incoming $\boldsymbol{E}$ features exactly. The target average $-b_j\langle \boldsymbol{P}_j \bar{\boldsymbol{v}}_j, \boldsymbol{y}_j\rangle$ is canceled by its centering correction. Both extra linear terms have mean zero. This proves the pointwise population calibration. □

Lemma H.5 (Uniform compressed branch margin)**.** *For distinct training labels with support $S \in \mathcal{U}$, at $\boldsymbol{y}_r = tb_r \boldsymbol{P}_r \boldsymbol{p}_r(S)$ for every $t > 0$, branch $U = S$ strictly maximizes every training $\psi_{j,i}$.*

Proof. For $U \ne S$ containing i, set $\ell = |S \cap U|$, $m = n - \ell$, and $D = 1 - \rho_j(\ell) > 0$. Before compression, the correct-minus-wrong margin divided by t is $M_j - B_j$, where

$$M_j = \sum_{r<j} A_{jr} b_r m_r(\ell), \qquad B_j = b_j^2 D.$$

Equations (H.4) and (H.10) give $M_j \ge 2B_j$. The compression error is at most

$$\epsilon \sum_{r<j} A_{jr} b_r \min\{2, 2\sqrt{rm/n}\} + \epsilon b_j^2 \sqrt{2D}.$$

If $jm/n \le 1$, use $D \ge jm/(2n)$, $\sum_{r<j} r \le 2j$ and the coefficient identity: the first term is at most $112\epsilon n B_j$. If $jm/n \ge 1$, use $D \ge 1/2$ and $\sum_{r<j}\sqrt{r/n} < 4$ to bound it by $224\epsilon n B_j$. The last term is at most $2\epsilon n B_j$. Here $D \ge \frac{1}{2}\min\{jm/n, 1\}$ follows from $\rho_j \le (1-m/n)^j \le e^{-jm/n}$. Since $226\epsilon n < 1/2$, the margin remains strictly positive. □

H.6 The reference ray and a global population certificate

Let $\mathcal{E}$ be the event that all labels are distinct and $\sqrt{n}/2 \le Z := \sum_{t=1}^n \xi_t \le 4\sqrt{n}$. A birthday union bound gives distinctness probability at least $7/8$. Since $\mathbb{E}Z^2 = n$ and $\mathbb{E}Z^4 \le 3n^2$, Paley–Zygmund gives $\Pr(|Z| \ge \sqrt{n}/2) \ge 3/16$. Markov gives $\Pr(|Z| > 4\sqrt{n}) \le 1/16$. Symmetry therefore gives probability at least $1/16$ for the positive band. Independence implies

$$\Pr(\mathcal{E}) \ge 1/32. \tag{H.13}$$

On this event put

$$\nu = \frac{\beta Z}{n\sqrt{k}} \in [b_k/2, 4b_k], \qquad \boldsymbol{V} = ((b_r \boldsymbol{P}_r \boldsymbol{p}_r(S))_r, \nu), \qquad A = \|\boldsymbol{V}\| \in [b_k/2, 5b_k]. \tag{H.14}$$

The norm bound uses the finite certificates and (H.3).

Proposition H.6 (Reference ray and small-norm support)**.** *On $\mathcal{E}$, the empirical encoder is differentiable at $t\boldsymbol{V}$ for every $t > 0$, with gradient $-\boldsymbol{V}$, and every training $p_i(t\boldsymbol{V}) = 0$. Also $-\boldsymbol{V} \in \partial e_S(0)$ and $e_S(t\boldsymbol{V}) = -t\|\boldsymbol{V}\|^2$. Its constrained reference flow is $\boldsymbol{u}(t) = \min\{t, 1/A\}\boldsymbol{V}$. There is a linear functional ℓ_S such that, globally,*

$$\mathcal{P}(\boldsymbol{w}) \ge \ell_S(\boldsymbol{w}), \qquad \|\ell_S\| \le a_k, \qquad \ell_S(\boldsymbol{V}) \ge a_k b_k/32. \tag{H.15}$$

Proof. Lemma H.5 identifies the actual encoder branches. Averaging their gradients over the distinct training labels cancels incoming $\boldsymbol{E}$ features, and the remaining terms, including block one and the scalar, give $-\boldsymbol{V}$. For training $i \in S$, Lemma H.3 and (H.8) bound every readout score at $t\boldsymbol{V}$ by

$$t(b_k\epsilon - \delta a_k \nu) \le t b_k(\epsilon - a_k/32) < 0.$$

The inequality $\epsilon < a_k/32$ holds for every $n \ge 2, k \ge 1$. Thus the training readout is zero. Selecting branch S in each encoder maximum at zero gives subgradient $-\boldsymbol{V}$. Positive homogeneity gives $e_S(t\boldsymbol{V}) = -t\|\boldsymbol{V}\|^2$. The ray to the boundary and its stationary endpoint solve the normal-cone flow inclusion, whose solution is unique by monotonicity.

For every unseen i, keep branch $U = S$ in the readout. For seen i, keep zero. This gives the global supporting functional

$$\ell_S(\boldsymbol{y}, x) = \tfrac{1}{4}[\langle \boldsymbol{P}_k \boldsymbol{C}_S, \boldsymbol{y}_k \rangle - \delta a_k \gamma x], \qquad \gamma = (M-n)/M. \tag{H.16}$$

The explicit norm certificate for $\boldsymbol{C}_S$ and (H.7) give

$$\|\ell_S\|^2 \le \tfrac{1}{16}[(1+\epsilon)\tfrac{3}{2}a_k^2 + \delta^2 a_k^2] \le a_k^2.$$

For each unseen label, compression preserves pairing at least $a_k - \epsilon$. Since $\nu \le 4b_k$, $\gamma \ge 3/4$ and $\epsilon \le a_k/4$,

$$\ell_S(\boldsymbol{V}) \ge \frac{\gamma b_k}{4}[(1-4\delta)a_k - \epsilon] \ge \frac{3a_k b_k}{32} \ge \frac{a_k b_k}{32}.$$

The functional is an analysis certificate chosen for a sample support, while the loss was fixed before sampling. Its pointwise supporting inequality holds at any data-dependent perturbed output. □

H.7 Samplewise smoothing and all safe discrete steps

We now prove Lemma H.1. Use the samplewise Moreau notation of (G.1), and choose

$$c_\theta = 2^{-32}, \quad c_\mu = 2^{-16}, \quad K = 2^{18}, \qquad \theta = c_\theta \min\{1, L/\sqrt{nk}\}, \qquad \mu = c_\mu a_k.$$

For each data value, define the actual loss

$$\widehat{f}_z(\boldsymbol{w}) = K^{-1}\big[(\theta e_z)_{1/L}(\boldsymbol{w}) + (p_i)_\mu(\boldsymbol{w})\big]. \tag{H.17}$$

Its global Lipschitz constant is at most $(\theta/4 + 1/2)/K < 1$. Its gradient Lipschitz constant is at most $(L + 1/\mu)/K \le L$, because $La_k \ge \sqrt{n/(n+1)} \ge 1/\sqrt{2}$ and $1 + (L\mu)^{-1} \le 1 + \sqrt{2}\,2^{16} < K$.

Fix a data set in $\mathcal{E}$. In raw time $Q = q/K$, with raw step $h = \eta/K$, the reference objective is $\Phi = \theta e_S$, speed $r = \theta A > 0$, and direction $\boldsymbol{v} = \boldsymbol{V}/A$. Averaging samplewise envelope inequalities gives $0 \le \Phi - \widetilde{\Phi} \le \varepsilon := \theta^2/(32L)$. The averaged smoothed readout is nonnegative and vanishes on the entire reference ray, since each training readout does and a nonnegative envelope preserves its zero set. The actual raw objective has smoothness at most $L + 1/\mu$, and $h \le 1/(KL) \le 1/(L + 1/\mu)$. Lemma G.3, for zero-initialized GD on the unit ball, therefore applies to the actual empirical objective.

Its main comparison term satisfies

$$\frac{2\varepsilon}{r} = \frac{\theta}{16LA} \le \frac{\theta\sqrt{nk}}{8\beta L} \le \frac{c_\theta}{8\beta} = 2^{-25}.$$

The other terms in E_0 are $2h\varepsilon + hr + h^2r^2$. Since $h \le 1/K$, $\theta \le c_\theta$ and $r \le 5\beta c_\theta/\sqrt{nk}$, their sum is less than 2^{-25}. Hence $E_0 < 2^{-24}$, and both last-point and complete-average displacement, including quadrature, are less than 2^{-11}. Flow has a smaller bound. The estimate covers the crossing update and every later iterate through the radial recurrence in Lemma G.3.

If $Q \ge 1/r$, either continuous reference output $\boldsymbol{U}$ satisfies

$$\ell_S(\boldsymbol{U}) \ge \frac{\ell_S(\boldsymbol{V})}{2A} \ge a_k/320.$$

For the average, its radial coefficient is $1 - (2rQ)^{-1} \ge 1/2$. The last output is $\boldsymbol{v}$. The supporting functional gives, for the actual output,

$$\mathcal{P}(\widehat{\boldsymbol{w}}) \ge \ell_S(\widehat{\boldsymbol{w}}) \ge a_k/320 - a_k 2^{-11}.$$

Using merely a unit Lipschitz bound here would charge a constant error against a vanishing signal. The small norm in (H.15) avoids this loss.

Let $\widehat{F}$ be the actual population objective of (H.17). Global calibration and the samplewise envelope estimates give, pointwise,

$$K\widehat{F}(\boldsymbol{w}) \ge \mathcal{P}(\boldsymbol{w}) - \varepsilon - \mu/8, \qquad \widehat{F}^* \le \widehat{F}(0) \le 0.$$

No preservation of the population minimizer by smoothing is assumed. Since $\varepsilon/a_k \le \sqrt{2}c_\theta^2/32$ and $\mu/(8a_k) = 2^{-19}$, the total smoothing and displacement loss is smaller than $a_k/1024$. Because $1/320 - 1/1024 > 1/512$, on $\mathcal{E}$,

$$\widehat{F}(\widehat{\boldsymbol{w}}) - \widehat{F}^* \ge a_k/(512K).$$

Actual excess risk is nonnegative outside $\mathcal{E}$, since the output is feasible. Equation (H.13) therefore yields

$$\mathbb{E}[\widehat{F}(\widehat{\boldsymbol{w}}) - \widehat{F}^*] \ge \frac{a_k}{16384K} \ge \frac{1}{16384\sqrt{2}K}\sqrt{k/n}. \tag{H.18}$$

Finally,

$$\frac{K}{r} \le \frac{2K\sqrt{nk}}{\beta\theta} = C_* \max\{\sqrt{nk}, nk/L\}, \qquad C_* = \frac{2K}{\beta c_\theta} = 2^{61}.$$

This data-independent threshold ensures $Q \ge 1/r$ for every good sample, every safe step and all later horizons. All parameters depend only on n, k, L and were fixed before sampling. This proves Lemma H.1, with deliberately conservative universal constants. The theorem describes asymptotic rates. These constants are not a sharp finite-sample training prescription.

H.8 Integer degree optimization and completion

Theorem H.7 (Sharp high-curvature band at every time)**.** *For $L \ge \sqrt{n}$, $d \ge D_n$ and each of the four specified output rules $\widehat{\boldsymbol{w}}$,*

$$\mathcal{R}_{n,d,L}(\widehat{\boldsymbol{w}}) = \Theta\big(\min\{1, q^{-1} + \min(q/n, \sqrt{Lq}/n)\}\big). \tag{H.19}$$

The upper is dimension independent and the constants are independent of the safe constant step. In particular, $L \ge n$ suffices for $\min\{1, q^{-1} + q/n\}$ at every time.

Proof. The upper is (H.2). For $q \ge C_*\sqrt{n}$ choose the integer

$$x = \min\{n, q^2/(C_*^2 n), Lq/(C_* n)\}, \qquad k = \lfloor x \rfloor.$$

Each term in the minimum is at least one, so $1 \le k \le n$ and $k \ge x/2$. Also $C_*\sqrt{nk} \le q$ and $C_* nk/L \le q$. Thus (H.18) is a universal multiple of $H(n, q, L) := \min\{1, q/n, \sqrt{Lq}/n\}$. For $q < C_*\sqrt{n}$, $H \le q/n \le C_*^2/q$. The optimization witness absorbs this term. Explicitly, on $[-1, 1]$ take $f(w) = -cw$, $c = \min\{1, 1/(2q)\}$. Both dynamics and both outputs are at most $1/2$ by time q, so their excess is at least $c/2$. This affine loss is admissible at every curvature. Taking the maximum over the affine and degree-selected witnesses lower-bounds the capped sum. Both fit into D_n. At $L \ge n$, $q \le n$ selects the linear branch, and for $q \ge n$ both growth branches are at least one. The claim follows, including all integer GD horizons. □

For $1 \le L \le \sqrt{n}$, the bounds proved in Appendices C and E give exactly (2). For $L \ge \sqrt{n}$, its extra $n^{-1/2}$ term is redundant: it is at most q^{-1} when $q \le \sqrt{n}$, and at most both growth branches when $q \ge \sqrt{n}$. Theorem H.7 completes Theorem 2.1. Zero-padding embeds every lower construction into any larger allowed dimension and preserves the ball trajectory.

For $\sqrt{n} \le L \le n$, the resulting risk phases are

$$\mathcal{R}_{n,d,L}(\widehat{\boldsymbol{w}}) \asymp \begin{cases} \min\{1, q^{-1}\}, & q \le \sqrt{n}, \\ q/n, & \sqrt{n} \le q \le L, \\ \sqrt{Lq}/n, & L \le q \le n^2/L, \\ 1, & q \ge n^2/L. \end{cases}$$

At $L = n$, $q = n^{3/4}$ the lower bound is $\Omega(n^{-1/4})$, matching the upper bound in the stated polynomial dimension. The result does not assert the same lower in every $d \ge Cn$, an efficient gradient oracle for this encoder, a high-probability risk law, or one loss tracing the whole optimized envelope.

H.9 Proof of Theorem 2.4

Proof. Let C_{low} be the universal constant KC_0 required by the construction in Appendix E, and let $C_{\mathrm{high}} \geq 1$ be a universal time constant for Lemma H.1. Choose $C \geq \max\{C_{\mathrm{low}}, C_{\mathrm{high}}\}$ and fix $n \geq 2$, $L \geq 1$, and $q_0 \geq C\sqrt{n}$.

Low curvature: $1 \leq L \leq \sqrt{n}$. Use the losses in (E.8), with the parameters in (E.7) chosen for the prescribed q_0. These losses are globally convex, 1-Lipschitz, and L-smooth, and their dimension is at most a universal multiple of n. Equation (E.19) gives the full-average bound for every later horizon, and Lemma E.5 gives the same bound for the last iterate on the same instance. Both results hold for every safe constant GD step and for constrained flow. Consequently, for either output $\widehat{\boldsymbol{w}}$ at any such horizon,

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] \geq c_{\mathrm{low}} \min\left\{1, \frac{Lq_0}{n^{3/2}}, \sqrt{\frac{q_0}{n^{3/2}}}\right\} = c_{\mathrm{low}} \min\{1, g_{n,L}(q_0)\}.$$

The instance parameters depend on n, L, q_0, but not on the sample, step size, or later horizon.

High curvature: $L > \sqrt{n}$. Choose the degree before sampling by setting

$$x = \min\left\{n, \frac{q_0^2}{C_{\mathrm{high}}^2 n}, \frac{Lq_0}{C_{\mathrm{high}} n}\right\}, \qquad k = \lfloor x \rfloor.$$

Each quantity in this minimum is at least one, because $q_0 \geq C_{\mathrm{high}}\sqrt{n}$ and $L > \sqrt{n}$. Thus $1 \leq k \leq n$ and $k \geq x/2$. Also,

$$C_{\mathrm{high}}\sqrt{nk} \leq q_0, \qquad C_{\mathrm{high}} nk/L \leq q_0.$$

Fix the instance supplied by Lemma H.1 for this integer k. Its time condition is therefore satisfied at every $q \geq q_0$. The lemma gives, on this one instance, for the last iterate and the complete post-update average, simultaneously for all $0 < \eta \leq 1/L$ and all integer T with $\eta T \geq q_0$, the lower bound

$$\begin{aligned}
\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] &\geq c_{\mathrm{high}}\sqrt{k/n} \\
&\geq \frac{c_{\mathrm{high}}}{\sqrt{2}} \min\left\{1, \frac{q_0}{C_{\mathrm{high}} n}, \frac{\sqrt{Lq_0}}{\sqrt{C_{\mathrm{high}}\, n}}\right\} \\
&\geq \frac{c_{\mathrm{high}}}{\sqrt{2}C_{\mathrm{high}}} \min\{1, q_0/n, \sqrt{Lq_0}/n\} = \frac{c_{\mathrm{high}}}{\sqrt{2}C_{\mathrm{high}}} \min\{1, g_{n,L}(q_0)\}.
\end{aligned}$$

The same lemma gives this bound for the flow last point and complete time average at every $q \geq q_0$. The losses satisfy the required global regularity in at most a universal multiple of $n^3 \log^2(2n)$ dimensions. Their degree, coefficients, and compression maps are fixed before sampling and are independent of the step size and later horizon.

Taking $c = \min\{c_{\mathrm{low}}, c_{\mathrm{high}}/(\sqrt{2}C_{\mathrm{high}})\}$ and a common dimension constant C_d proves (7) in both regimes. □

For the pointwise lower side of Theorem 2.1, choosing $q_0 = q \geq C\sqrt{n}$ supplies the growth term. At smaller horizons, $g_{n,L}(q) \leq q/n \leq C^2/q$ in either curvature regime. The separate affine optimization and statistical witnesses in Appendix C.7 then complete the capped sum. This pointwise choice of q_0 is distinct from the fixed-instance persistence asserted by Theorem 2.4.

I Proof of the risk law under weak negative curvature

Theorem 3.1 is proved in three stages. First, we obtain both response kernels using convexification and the arbitrary-state estimate of Appendix C. Next, we construct a conic readout on the entire learned ray. Finally, we verify actual seed-and-release dynamics using the code/prox construction of Appendix E. Section I.7 adds uniform checkpoint selection and completes the three-output law. Section I.8 proves the common-instance separation in Corollary 3.2. Section I.5.5 extends the implementation in Section E.8, including propagation of gradient and rounding errors.

I.1 Assumptions and exact response kernels

Let $n \geq 2$, $0 \leq \rho \leq 1$, and $S = (Z_1, \ldots, Z_n) \sim P^n$. Sample losses are jointly measurable, finite on $\mathbb{R}^d$, globally 1-Lipschitz, have globally 1-Lipschitz gradients, and are ρ-weakly convex, meaning that

$$f_z(\boldsymbol{w}) + \frac{\rho}{2}\|\boldsymbol{w}\|^2 \quad \text{is convex on } \mathbb{R}^d. \tag{I.1}$$

Normalize $f_z(0) = 0$, and write $F_S = n^{-1}\sum_i f_{Z_i}$, $F = \mathbb{E}f_Z$, and $F^* = \min_{\mathbb{B}^d} F$. Run the actual updates

$$\boldsymbol{w}_0 = 0, \qquad \boldsymbol{w}_{j+1} = \Pi_{\mathbb{B}^d}(\boldsymbol{w}_j - \eta\nabla F_S(\boldsymbol{w}_j)), \qquad 0 < \eta \leq 1, \quad q = \eta T, \tag{I.2}$$

with outputs $\boldsymbol{w}_T$ and $\bar{\boldsymbol{w}}_T = T^{-1}\sum_{j=1}^T \boldsymbol{w}_j$. For either fixed output rule $\widehat{\boldsymbol{w}}$, let $\mathcal{R}^{\mathrm{wc}}_{n,d,\rho}(\widehat{\boldsymbol{w}})$ be the supremum of $\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*]$ over this weakly convex class in dimension at most d. Here $\widehat{\boldsymbol{w}}$ denotes the sample-to-output rule, with learner, step η, and horizon T fixed. There is no infimum over algorithms.

For $\rho > 0$, put $a = 1 + \eta\rho$ and define

$$\begin{aligned} K_j^{(1)} &= \eta\sum_{s=0}^{j-1} a^s = \frac{a^j - 1}{\rho}, \\ K_j^{(2)} &= \eta\sum_{s=0}^{j-1} a^{2s} = \frac{a^{2j} - 1}{\rho(2 + \eta\rho)}. \end{aligned} \tag{I.3}$$

$$d_j = \min\left\{1, \sqrt{\frac{K_j^{(2)}}{n^{3/2}}}, \frac{K_j^{(1)}}{n}\right\}, \qquad M_{\mathrm{last}} = d_T, \qquad M_{\mathrm{av}} = \frac{1}{T}\sum_{j=1}^T d_j. \tag{I.4}$$

At $\rho = 0$, define $K_j^{(1)} = K_j^{(2)} = \eta j$ by continuity. Write $M(\boldsymbol{w}_T) = M_{\mathrm{last}}$ and $M(\bar{\boldsymbol{w}}_T) = M_{\mathrm{av}}$. For the fixed n, T, ρ, set

$$\Psi_\eta(\widehat{\boldsymbol{w}}) = \min\{1, q^{-1} + n^{-1/2} + \rho + M(\widehat{\boldsymbol{w}})^2\}. \tag{I.5}$$

Theorem I.1 (Matching risk envelope for every safe constant step)**.** *There are universal constants $c, C, C_d > 0$ such that, for every $n \geq 2$, $0 \leq \rho \leq 1$, $0 < \eta \leq 1$, $T \geq 1$, and $\widehat{\boldsymbol{w}} \in \{\boldsymbol{w}_T, \bar{\boldsymbol{w}}_T\}$,*

$$c\Psi_\eta(\widehat{\boldsymbol{w}}) \leq \mathcal{R}^{\mathrm{wc}}_{n,d,\rho}(\widehat{\boldsymbol{w}}) \leq C\Psi_\eta(\widehat{\boldsymbol{w}}).$$

The upper holds in every finite dimension. The lower holds for $d \geq C_d n$. All lower losses are fixed before sampling. For each n and sufficiently small $\rho > 0$, one amplifier family, depending on neither η nor T, supplies the $M(\widehat{\boldsymbol{w}})^2$ lower contribution for both outputs whenever $\rho q \geq 20$. Its losses have certified polynomial-accuracy sample-gradient oracles. The remaining lower contributions use the usual separate optimization/statistical witnesses, a deterministic weak-convexity witness, and the contained convex class.

For flow, use the unique constrained evolution $\dot{\boldsymbol{w}} + \nabla F_S(\boldsymbol{w}) + N_{\mathbb{B}^d}(\boldsymbol{w}) \ni 0$, $\boldsymbol{w}(0) = 0$. At fixed time q, let $\widehat{\boldsymbol{w}}$ denote either the endpoint rule $\boldsymbol{w}(q)$ or the full-average rule $q^{-1}\int_0^q \boldsymbol{w}(t)\,dt$, and define its risk by the same supremum over instances. Define

$$K^{(1)}(t) = \frac{e^{\rho t} - 1}{\rho}, \qquad K^{(2)}(t) = \frac{e^{2\rho t} - 1}{2\rho}. \tag{I.6}$$

$$\begin{aligned} d(t) &= \min\left\{1, \sqrt{K^{(2)}(t)/n^{3/2}}, K^{(1)}(t)/n\right\}, \\ M_{\mathrm{last}} &= d(q), \qquad M_{\mathrm{av}} = q^{-1}\int_0^q d(t)\,dt. \end{aligned} \tag{I.7}$$

Again $K^{(1)}(t) = K^{(2)}(t) = t$ at $\rho = 0$. For these flow rules, $M(\widehat{\boldsymbol{w}})$ denotes M_{last} and M_{av}, respectively.

Well-posedness of the flow under weak negative curvature. Projected Euler interpolants stay in the compact ball and have speed at most one. On each finite time interval, a uniformly convergent subsequence and weakly convergent velocities satisfy the normal-cone inclusion by the projection variational inequality and continuity of ∇F_S. Weak convexity and normal-cone monotonicity yield $\frac{d}{dt}\|\boldsymbol{u}-\boldsymbol{v}\|^2 \le 2\rho\|\boldsymbol{u}-\boldsymbol{v}\|^2$ for two solutions. Gronwall gives uniqueness. Hence a fixed Euler sequence converges, and its measurable limit defines the random flow. These arguments also justify the weak EVI used below. Existence here is not inferred from convexity of F_S.

Corollary I.2 (Complete three-parameter flow law)**.** *With the same dimension and instance conventions, both flow risks are*

$$\mathcal{R}^{\mathrm{wc}}_{n,d,\rho}(\widehat{\boldsymbol{w}}) = \Theta\big(\min\{1, q^{-1} + n^{-1/2} + \rho + M(\widehat{\boldsymbol{w}})^2\}\big),$$

with $M(\widehat{\boldsymbol{w}})$ from (I.7).

I.2 Closed forms and the two output laws

The kernel formula is exact as an order characterization. The following elementary closed form makes its phases easier to read. For $\rho > 0$, define

$$\delta_{n,\rho} = \min\left\{\frac{1}{\sqrt{\rho n^{3/2}}}, \frac{1}{\rho n}\right\}, \quad x_\eta = T\log(1+\eta\rho), \quad x_0 = \rho q, \quad B = q^{-1} + n^{-1/2} + \rho. \tag{I.8}$$

For $\delta, x > 0$, set

$$\mathcal{A}(\delta, x) = \frac{1}{x}\int_0^x \min\{1, \delta e^u\}\, du = \begin{cases} 1, & \delta \ge 1, \\ \delta(e^x - 1)/x, & \delta < 1,\ x \le \log(1/\delta), \\ [x - \log(1/\delta) + 1 - \delta]/x, & \delta < 1,\ x > \log(1/\delta). \end{cases} \tag{I.9}$$

In the following comparisons, abbreviate $\delta = \delta_{n,\rho}$.

Corollary I.3 (Explicit envelope)**.** *If $\rho q \le 1$, both GD outputs have the original order*

$$\mathcal{R}^{\mathrm{wc}}_{n,d,\rho}(\widehat{\boldsymbol{w}}) = \Theta\big(\min\{1, q^{-1} + n^{-1/2} + q/n^{3/2}\}\big). \tag{I.10}$$

If $\rho q > 1$, then

$$\mathcal{R}^{\mathrm{wc}}_{n,d,\rho}(\boldsymbol{w}_T) = \Theta\big(\min\{1, B + \delta_{n,\rho}^2 e^{2x_\eta}\}\big), \tag{I.11}$$

$$\mathcal{R}^{\mathrm{wc}}_{n,d,\rho}(\bar{\boldsymbol{w}}_T) = \Theta\big(\min\{1, B + \mathcal{A}(\delta_{n,\rho}, x_\eta)^2\}\big). \tag{I.12}$$

For flow, use the endpoint and full-time-average rules and replace x_η by $x_0 = \rho q$. The dimension conventions are those of Theorem I.1.

Proof. If $\rho q \le 1$, $K_j^{(2)} \asymp \eta j$ and $K_j^{(1)} \asymp \eta j$, with universal constants. On $j \in [\lceil T/2\rceil, T]$, $d_j \ge c d_T$. Hence $M(\widehat{\boldsymbol{w}})^2$ for either output is comparable to $\min\{1, q/n^{3/2}, q^2/n^2\}$. If $q \ge \sqrt{n}$, this equals $\min\{1, q/n^{3/2}\}$. Below $\sqrt{n}$, the missing term is bounded by $n^{-1/2}$. Also $\rho \le 1/q$.

For the second claim, let $\ell = \log(1+\eta\rho)$. We have $x_\eta = \ell T \ge (\log 2)\rho q > \log 2$ since $\eta\rho \le 1$. For every j, $d_j \le \min\{1, \delta_{n,\rho} e^{\ell j}\}$. For $j \ge T/2$, $\ell j \ge (\log 2)/2$, so $a^j - 1 \ge c a^j$ and $a^{2j} - 1 \ge c a^{2j}$. Consequently the reverse inequality holds up to a universal constant on this latter half. Nondecreasing sequences have at least half their total mass on their last half. Thus M_{av} is comparable to the average of $\min\{1, \delta e^{\ell j}\}$, and $M_{\mathrm{last}} \asymp \min\{1, \delta e^{x_\eta}\}$. The geometric right-endpoint sum is between the integral in (I.9) and e^ℓ times that integral, since shifting u by ℓ multiplies the clipped exponential by at most $e^\ell \le 2$. This proves (I.11)–(I.12). The continuous calculation is identical. □

Corollary I.4 (A step-uniform three-parameter law for full averaging)**.** *Uniformly over all safe constant steps, the full-average GD risk has the flow envelope: in (I.12), x_η can be replaced by ρq without changing the universal order. This includes the ρ optimization floor. It is not an assertion that individual trajectories or losses are step-size invariant.*

Proof. For fixed δ, put $A(x) = \mathcal{A}(\delta, x)$. It is nondecreasing and

$$0 \le A'(x) = \frac{\min\{1, \delta e^x\} - A(x)}{x} \le \frac{1}{x}.$$

Let $v = \eta\rho \le 1$. Since $\log(1+v) \ge v/(1+v)$,

$$0 \le A(\rho q) - A(x_\eta) \le \log \frac{\rho q}{x_\eta} = \log \frac{v}{\log(1+v)} \le \log(1+v) \le v \le \rho.$$

Thus $A(\rho q)^2 \le 2A(x_\eta)^2 + 2\rho^2 \le 2A(x_\eta)^2 + 2\rho$. The reverse comparison follows from monotonicity. Add $B \ge \rho$ and cap. The small-budget case was already covered. □

The two seed branches cross at $\rho = n^{-1/2}$. When $\delta < 1$, the logarithmic last-point transition center is

$$h = \log(1/\delta) = \begin{cases} \frac{1}{2}\log(\rho n^{3/2}), & n^{-3/2} < \rho \le n^{-1/2}, \\ \log(\rho n), & \rho \ge n^{-1/2}. \end{cases} \tag{I.13}$$

This is a description of the envelope, not an exact finite-sample threshold for unspecified hidden risk constants. Before saturation and at large x, the last-point growing contribution is $\delta^2 e^{2x}$, whereas the average contribution is of order $\delta^2 e^{2x}/x^2$. After last-point saturation, the averaged contribution is $[(x - h + 1 - \delta)/x]^2$. Constants multiply the risk, not x.

I.3 Upper bound: retain both responses, then average in norm

I.3.1 Squared replacement response

Let $S^{(i)}$ replace one observation independently and put

$$v_j = \frac{1}{n}\sum_i \mathbb{E}\|\boldsymbol{w}_j(S) - \boldsymbol{w}_j(S^{(i)})\|^2.$$

The following uses Lemma C.1. It is valid for arbitrary bounded measurable input states, not only for states of a convex GD trajectory.

Lemma I.5 (Three simultaneous influence bounds)**.** *For every j,*

$$v_j \le 192 K_j^{(2)}/n^{3/2}, \qquad v_j \le \frac{4\big(K_j^{(1)}\big)^2}{n^2}, \qquad v_j \le 4.$$

Consequently $\sqrt{v_j} \le \sqrt{192}\, d_j$. For flow, the same statements hold with $K^{(1)}(t), K^{(2)}(t)$, after a possible universal constant improvement.

Proof. Set $g_z(\boldsymbol{w}) = f_z(\boldsymbol{w}) + \rho\|\boldsymbol{w}\|^2/2$, $g_S = n^{-1}\sum_i g_{Z_i}$, $a = 1 + \eta\rho$, and $h = \eta/a$. The functions g_z are convex and $(1+\rho)$-smooth. At states in the unit ball their gradients have norm at most $G = 1 + \rho$. Also

$$\boldsymbol{w} - \eta\nabla F_S(\boldsymbol{w}) = a[\boldsymbol{w} - h\nabla g_S(\boldsymbol{w})], \qquad h(1+\rho) \le 1.$$

For a replacement i, write $\boldsymbol{\Delta} = \boldsymbol{X} - \boldsymbol{X}^{(i)}$ and split the convexified empirical gradient difference as $\boldsymbol{A}_i + \boldsymbol{D}_i$, with $\boldsymbol{A}_i$ the sum over common observations and $\boldsymbol{D}_i$ the changed-observation term. The proof of Theorem C.2, before projection, gives

$$\|\boldsymbol{\Delta} - h(\boldsymbol{A}_i + \boldsymbol{D}_i)\|^2 \le \|\boldsymbol{\Delta}\|^2 - \frac{h}{2}\langle \boldsymbol{A}_i, \boldsymbol{\Delta}\rangle - 2h\langle \boldsymbol{D}_i, \boldsymbol{\Delta}\rangle + 12h^2G^2/n^2.$$

The original candidate difference is exactly $a[\boldsymbol{\Delta} - h(\boldsymbol{A}_i + \boldsymbol{D}_i)]$. Multiply this inequality by a^2, then apply the nonexpansiveness of the final projection. In particular projection is never commuted with dilation. The arbitrary-state ID square completion, retaining the changed-observation sign and the common-loss negative squares, gives

$$v_{j+1} \le a^2 v_j + a\eta\left(4Gn^{-3/2} + 8(1+\rho)n^{-2}\right) + 12\eta^2G^2n^{-2}.$$

Since $a, G \le 2$, its additive term is at most $96\eta(n^{-3/2}+n^{-2}) \le 192\eta n^{-3/2}$. Sum from $v_0 = 0$. For specificity, the underlying ID algebra expands $\boldsymbol{X}$ and the evaluated gradients into orthogonal sample modes. A mode of order $k \ge 2$ has $k-1$ common observations. Completing its signed interaction against their squares costs at most a factor $k/(k-1) \le 2$. Singletons cost $4G/n^{3/2}$. Nothing in that algebra requires $\boldsymbol{X}$ itself to be produced by a convex objective, and global boundedness of ∇g_z is unnecessary.

Independently, the convexified step $\boldsymbol{w} \mapsto \boldsymbol{w} - h\nabla g_S(\boldsymbol{w})$ is nonexpansive by cocoercivity since $h(1+\rho) \le 1$. Multiplication by a and the final projection make the actual fixed-sample step a-Lipschitz. A replacement at the same point changes one update by at most $2\eta/n$, giving pointwise replacement distance at most $2K_j^{(1)}/n$ by induction. The unit-ball diameter gives the third bound. For flow, put $v(t) = n^{-1}\sum_i \mathbb{E}\|\boldsymbol{w}_S(t) - \boldsymbol{w}_{S^{(i)}}(t)\|^2$. Normal-cone monotonicity and the convexified ID calculation give $v' \le 2\rho v + 12(n^{-3/2}+n^{-2})$. For any fixed replacement pair, the distance $D(t) = \|\boldsymbol{w}_S(t) - \boldsymbol{w}_{S^{(i)}}(t)\|$ satisfies $D' \le \rho D + 2/n$. Integrate both inequalities. □

Lemma I.6 (Averaging before squaring)**.** *For the complete post-update average,*

$$\left(\frac{1}{n}\sum_i \mathbb{E}\|\bar{\boldsymbol{w}}_T(S) - \bar{\boldsymbol{w}}_T(S^{(i)})\|^2\right)^{1/2} \le \frac{1}{T}\sum_{j=1}^{T}\sqrt{v_j} \le \sqrt{192}\, M_{\mathrm{av}}.$$

For the last point use $\sqrt{v_T}$. Flow uses the corresponding integral.

Proof. Apply the triangle inequality in the single Hilbert space $L^2(\{1,\dots,n\}\times\Omega;\mathbb{R}^d)$ to the paired state differences. This requires no independence across times. It is strictly more informative here than replacing the average's squared influence by the average of v_j. □

I.3.2 Optimization and original-loss risk

Lemma I.7 (Empirical comparison and population risk under weak negative curvature)**.** *For a deterministic $\boldsymbol{w}^* \in \arg\min_{\mathbb{B}^d} F$, the actual outputs obey*

$$F_S(\boldsymbol{w}_T) - F_S(\boldsymbol{w}^*) \le (2q)^{-1} + 2\rho, \qquad F_S(\bar{\boldsymbol{w}}_T) - F_S(\boldsymbol{w}^*) \le (2q)^{-1} + \tfrac{5}{2}\rho.$$

If $\boldsymbol{X}$ has expected empirical comparison at most b and squared replacement influence $D_{\boldsymbol{X}}^2$, then

$$\mathbb{E}[F(\boldsymbol{X}) - F^*] \le 2b + 3n^{-1/2} + 2\rho + (1+\rho)D_{\boldsymbol{X}}^2.$$

Both assertions also hold for the actual flow outputs.

Proof. Projection optimality, the lower inequality for ρ-weakly convex functions, and smoothness give

$$F_S(\boldsymbol{x}^+) - F_S(\boldsymbol{w}^*) \le \frac{\|\boldsymbol{x}-\boldsymbol{w}^*\|^2 - \|\boldsymbol{x}^+ - \boldsymbol{w}^*\|^2}{2\eta} + \frac{\rho}{2}\|\boldsymbol{x}-\boldsymbol{w}^*\|^2 - \frac{1-\eta}{2\eta}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2.$$

Sum, use $\|\boldsymbol{x}-\boldsymbol{w}^*\| \le 2$ and $\|\boldsymbol{w}^*\| \le 1$, and use empirical descent for the last point. The complete average additionally pays weak Jensen:

$$F_S(\bar{\boldsymbol{w}}) \le T^{-1}\sum_{j=1}^{T} F_S(\boldsymbol{w}_j) + \frac{\rho}{2}\left(T^{-1}\sum_j \|\boldsymbol{w}_j\|^2 - \|\bar{\boldsymbol{w}}\|^2\right).$$

For the risk bound, define the globally nonnegative, convex, $(1+\rho)$-smooth residual

$$r_u(\boldsymbol{w}) = f_u(\boldsymbol{w}) - f_u(\boldsymbol{w}^*) - \langle \nabla f_u(\boldsymbol{w}^*), \boldsymbol{w}-\boldsymbol{w}^*\rangle + \frac{\rho}{2}\|\boldsymbol{w}-\boldsymbol{w}^*\|^2.$$

It satisfies $\|\nabla r_u(\boldsymbol{w})\|^2 \le 2(1+\rho)r_u(\boldsymbol{w})$. Population first-order optimality on the ball and the anchor variance bound imply $n^{-1}\sum_i \mathbb{E}r_{Z_i}(\boldsymbol{X}) \le b + 2n^{-1/2} + 2\rho$. Use the exact replace-one identity for the signed expected generalization gap. Its affine part is at most $n^{-1/2}$. The deterministic quadratic

introduced above cancels in expectation because $\boldsymbol{X}$ and $\boldsymbol{X}^{(i)}$ have the same marginal law. Taylor expansion at $\boldsymbol{X}$ and joint Cauchy–Schwarz give

$$\mathbb{E}[F(\boldsymbol{X}) - F^*] \le b + n^{-1/2} + \sqrt{2(1+\rho)(b + 2n^{-1/2} + 2\rho)}D_{\boldsymbol{X}} + \tfrac{1}{2}(1+\rho)D_{\boldsymbol{X}}^2.$$

Young's inequality proves the claim. In particular no data-dependent output is conditioned upon and then declared independent of a training gradient. Flow uses the weak EVI, empirical descent, and integral weak Jensen. □

Lemmas I.5–I.7, followed by the Lipschitz diameter cap two, prove the upper half of Theorem I.1 and Corollary I.2.

I.4 A SCALE-FREE READOUT THAT SURVIVES COMPLETE AVERAGING

A fixed-radius reader need not observe a small full average when late iterates are much larger: its training radial mask may activate first. The following lift makes its zero set contain the entire intended ray.

Fix the constructive code of Appendix E with message length $k = 2n$, block length $m \le Ck$, and unit binary codewords $\boldsymbol{c_v} \in \mathbb{R}^m$, $\boldsymbol{v} \in \{-1, 1\}^k$. Their pairwise correlations are at most $\gamma < 1$. Choose a fixed $\kappa \in (0, 1)$ satisfying

$$\kappa > \gamma, \qquad 2\kappa^2 > 1 + \gamma,$$

and the bounded-distance decoding condition in Appendix E. In particular a direction with correlation $> \kappa$ with a codeword is identified by rounding its signs and applying that fixed decoder. Set

$$s = \frac{1+\kappa}{2\kappa} > 1, \qquad \boldsymbol{d_v} = \frac{(\boldsymbol{c_v}, s)}{\sqrt{1+s^2}} \in \mathbb{R}^{m+1}, \qquad \nu = \frac{1 - \kappa s}{\sqrt{2(1+s^2)}} > 0. \tag{I.14}$$

Write the lifted code variable as $\boldsymbol{z} = (\boldsymbol{y}, u)$. Define

$$h_i(\boldsymbol{y}, u) = \frac{1}{\sqrt{2}} \max\left\{0, \|\boldsymbol{y}\| - u, \max_{\boldsymbol{v}: v_i = +1} (\langle \boldsymbol{c_v}, \boldsymbol{y}\rangle - \kappa u)\right\}, \qquad R_i = (h_i)_1. \tag{I.15}$$

The subscript means the Moreau envelope on the entire space.

Lemma I.8 (Conic reading and a global one-code reduction)**.** *Each h_i is nonnegative, convex, globally 1-Lipschitz, and zero at zero. Each R_i is additionally globally 1-smooth. For every message $\boldsymbol{\sigma}$ and $r \ge 0$,*

$$\sigma_i = -1 \implies R_i(r\boldsymbol{d_\sigma}) = 0, \quad \nabla R_i(r\boldsymbol{d_\sigma}) = 0.$$

For $r > 0$ this zero set contains a neighborhood of $r\boldsymbol{d_\sigma}$. If $\sigma_i = +1$ and $0 \le r \le 1/4$, then

$$R_i(r\boldsymbol{d_\sigma}) \ge \tfrac{1}{2}\nu^2 r^2. \tag{I.16}$$

The value and Moreau gradient admit a polynomial-accuracy oracle using at most one bounded-distance decoding call and a scalar convex/concave search. No enumeration of messages is needed.

Proof. The radial term has Lipschitz constant $\sqrt{2}$, and each code affine term has normal norm $\sqrt{1+\kappa^2} \le \sqrt{2}$. On the intended ray the radial score is strictly negative for $r > 0$, because $s > 1$. A wrong code has score proportional to $\gamma - \kappa s < 0$. Thus every training reader is zero on an open neighborhood, and so is the gradient of its nonnegative envelope. For an unseen bit the correct affine margin is νr. The reader dominates the hinge of that affine functional, whose normal has norm at most one. Its parameter-one envelope is at least $(\nu r)^2/2$ for $\nu r \le 1$. This follows directly by minimizing the one-dimensional hinge envelope, including its linear branch. This proves (I.16).

For the oracle, write $H_0(\boldsymbol{y}, u) = \max\{0, \|\boldsymbol{y}\| - u\}$, temporarily omitting the common $1/\sqrt{2}$. Any code branch strictly larger than H_0 has input-direction correlation $> \kappa$. Indeed, if $u \ge \|\boldsymbol{y}\|$, positivity gives $\langle \boldsymbol{c_v}, \boldsymbol{y}\rangle > \kappa u \ge \kappa\|\boldsymbol{y}\|$. If $0 \le u \le \|\boldsymbol{y}\|$, dominance over $\|\boldsymbol{y}\| - u$ gives $\langle \boldsymbol{c_v}, \boldsymbol{y}\rangle > \|\boldsymbol{y}\| - (1-\kappa)u \ge \kappa\|\boldsymbol{y}\|$. If $u < 0$, no code branch dominates the radial term. The angular cap contains at most one codeword by $2\kappa^2 > 1 + \gamma$.

Decode the signs of the input $\boldsymbol{y}$, retaining its affine branch only if the returned message has bit $i = +1$. If none is retained, take the proximal point of H_0. This map preserves the direction of $\boldsymbol{y}$ (or sets it to zero). An omitted dominant branch there would have been decoded from the input direction. Thus it is the full proximal point. If a branch $\ell_{\boldsymbol{c}}$ is retained, take the proximal point of $\max\{H_0, \ell_{\boldsymbol{c}}\}$. If the radial branch is strictly active, this is again the radial proximal point. Otherwise $\ell_{\boldsymbol{c}} \geq H_0 \geq 0$ at the proximal point. At a nonzero $\boldsymbol{y}$ this gives correlation at least κ with $\boldsymbol{c}$, leaving no room for a different omitted branch to exceed H_0. At $\boldsymbol{y} = 0$, the inequality forces $u = 0$ unless the radial branch dominates, and all code values are then zero. Hence this reduced proximal point is also the full proximal point.

The radial proximal map is explicit. For input $(\boldsymbol{y}, u)$ and $r = \|\boldsymbol{y}\|$, put $t = \min\{\lambda, \max(0, (r - u)/2, -u)\}$. Then

$$\operatorname{prox}_{\lambda H_0}(\boldsymbol{y}, u) = ((1 - t/r)_+ \boldsymbol{y}, u + t),$$

with first block zero when $r = 0$. This follows by minimizing over the radius and the scalar coordinate, or maximizing the one-dimensional dual over $0 \leq t \leq \lambda$. Mixing it with one affine function gives the same one-dimensional concave dual and strong-convexity error certificate as Lemma E.7. Restoring the $1/\sqrt{2}$ only changes the proximal parameter. Exact input signs and the decoder rule avoid unstable angular threshold tests. The bit-complexity argument is the same certified finite-precision argument, not unit-cost exact real arithmetic. □

I.5 The seed-and-release lower construction

I.5.1 Loss definition and constants

Keep the lifted unit codewords $\boldsymbol{d}_{\boldsymbol{v}}$. For $a_m, b > 0$, define the original convex code mask, now in the lifted variable,

$$\begin{aligned} H(\boldsymbol{x}, \boldsymbol{z}) &= a_m(\|\boldsymbol{x}\|_1 - \min_i |x_i|) + b\|\boldsymbol{z}\|, \\ P(\boldsymbol{x}, \boldsymbol{z}) &= \max\{H(\boldsymbol{x}, \boldsymbol{z}), \max_{\boldsymbol{v}} [a_m \langle \boldsymbol{v}, \boldsymbol{x}\rangle - b\langle \boldsymbol{d}_{\boldsymbol{v}}, \boldsymbol{z}\rangle]\}. \end{aligned} \tag{I.17}$$

Lemma E.1, Theorem E.2 and Lemma E.3 apply to arbitrary unit code maps, so the lift preserves four conclusions. The memory gradient coordinates of P_λ are bounded by a_m, its code-gradient norm is bounded by b, its prox reduces to the compatible-sign branch and H, and

$$a_m(\min_i |x_i| - \lambda a_m) > 2b(\|\boldsymbol{z}\| + \lambda b) \implies \nabla P_\lambda(\boldsymbol{x}, \boldsymbol{z}) = (a_m \boldsymbol{\sigma}, -b\boldsymbol{d}_{\boldsymbol{\sigma}}). \tag{I.18}$$

For compatible memory signs $\boldsymbol{\sigma}$, mix the affine branch with weight $t \in [0, 1]$ and the mask H with weight $1 - t$. At code input $\boldsymbol{z} = r\boldsymbol{d}_{\boldsymbol{\sigma}}$, $r \geq 0$, the code part of this proximal point is

$$\boldsymbol{z}_{\text{prox}} = [\, r + \lambda b(2t - 1)\,]_+ \boldsymbol{d}_{\boldsymbol{\sigma}}.$$

The formula holds also at $r = 0$ and at a dual tie. The strongly convex primal problem has a unique proximal point. Thus $\nabla_{\boldsymbol{z}} P_\lambda = (\boldsymbol{z} - \boldsymbol{z}_{\text{prox}})/\lambda$ is parallel to $\boldsymbol{d}_{\boldsymbol{\sigma}}$, has norm at most b, and can be written $A_{\boldsymbol{z}}\boldsymbol{z} - B_{\boldsymbol{z}}\boldsymbol{d}_{\boldsymbol{\sigma}}$ with $0 \leq A_{\boldsymbol{z}} \leq \lambda^{-1}$ and $0 \leq B_{\boldsymbol{z}} \leq b$. For example, assign its nonnegative scalar part to $A_{\boldsymbol{z}} r$ and its nonpositive part to $-B_{\boldsymbol{z}}$. At $r = 0$ the former vanishes.

Let $J(\boldsymbol{v})$ be the radial unit Huber function, equal to $\|\boldsymbol{v}\|^2/2$ inside the unit ball and $\|\boldsymbol{v}\| - 1/2$ outside. Put

$$\begin{gathered} A = 64, \quad K = 128A, \quad \lambda = (100A)^{-1}, \quad m_n = A/\sqrt{n}, \quad a_m = (4n)^{-1}, \\ R = 1/8, \quad R_+ = 2R = 1/4, \quad \omega = 1/16, \quad \varepsilon = 2^{-20}. \end{gathered} \tag{I.19}$$

Choose once and for all

$$0 < \rho_0 \leq \min\{10^{-3}, \omega/4, \nu^2/(4K)\}. \tag{I.20}$$

For $0 < \rho \leq \rho_0$, set

$$b = \varepsilon \min\left\{\frac{1}{n}, \sqrt{\frac{K\rho}{n^2 m_n}}, K\rho R\right\}, \qquad \beta = b/K, \qquad z_* = \beta/\rho. \tag{I.21}$$

In particular $z_* \leq \varepsilon R$. Define the globally convex radial brake

$$W(\boldsymbol{z}) = \omega J_1((\|\boldsymbol{z}\| - R)_+), \qquad J_1(t) = \begin{cases} t^2/2, & 0 \leq t \leq 1, \\ t - 1/2, & t \geq 1. \end{cases}$$

It is globally ω-Lipschitz and ω-smooth, zero for $\|\boldsymbol{z}\| \le R$, and has derivative $W'(r) = \omega(r - R)_+$ for $r \le R_+$.

Draw i uniformly from $[2n]$, put $\boldsymbol{\delta}_i = n^{-1}\mathbf{1} - 2\boldsymbol{e}_i$, and define the actual sample loss

$$f_i(\boldsymbol{x}, \boldsymbol{z}) = K^{-1}\big[P_\lambda(\boldsymbol{x}, \boldsymbol{z}) + m_n J(\boldsymbol{x}) + R_i(\boldsymbol{z}) - \langle \boldsymbol{\delta}_i, \boldsymbol{x}\rangle\big] - \rho J(\boldsymbol{z}) + W(\boldsymbol{z}). \tag{I.22}$$

This loss depends only on n, ρ and the fixed code, not on a training realization, horizon, or step. There is no code ridge and no quartic reduction in its inner negative curvature.

Lemma I.9 (Global regularity and population risk)**.** *The losses* (I.22) *are globally unit-Lipschitz, unit-smooth, and ρ-weakly convex, in dimension $2n + m + 1 = O(n)$. They vanish at zero. For a sample with unseen-bit pattern $\boldsymbol{\sigma}$, at every feasible point of form $(\boldsymbol{x}, r\boldsymbol{d}_{\boldsymbol{\sigma}})$, $0 \le r \le R_+$,*

$$F(\boldsymbol{x}, r\boldsymbol{d}_{\boldsymbol{\sigma}}) - F^* \ge \frac{\nu^2}{8K} r^2. \tag{I.23}$$

This statement uses only that at least half of the $2n$ labels are unseen.

Proof. All terms other than $-\rho J(\boldsymbol{z})$ are convex. The function J has Hessian between zero and the identity, in the $C^{1,1}$ sense. This proves global ρ-weak convexity. The mask has Lipschitz constant $\sqrt{2na_m^2 + b^2} < 1$, and its envelope has smoothness λ^{-1}. The total smoothness is at most

$$(100A + A + 1)/K + \rho_0 + \omega < 1,$$

and the Lipschitz bound is at most $(A+4)/K + \rho_0 + \omega < 1$. These are global bounds, not trajectory-only bounds. The population linear term cancels because $\mathbb{E}_i \boldsymbol{\delta}_i = 0$. All other terms except $-\rho J(\boldsymbol{z})$ are nonnegative, and $F^* \le F(0) = 0$. Lemma I.8 and the unseen fraction give

$$F(\boldsymbol{x}, r\boldsymbol{d}_{\boldsymbol{\sigma}}) \ge \frac{\nu^2}{4K} r^2 - \frac{\rho}{2} r^2 \ge \frac{\nu^2}{8K} r^2.$$

The last inequality follows from (I.20). No assertion that zero is an exact population minimizer is used. □

I.5.2 Invariants for the projected-gradient trajectory

Let N_i be the training counts, $\mu_i = (1 - 2N_i)/n$, and $\sigma_i = \operatorname{sign} \mu_i$. Then $|\mu_i| \ge 1/n$, $\sigma_i = +1$ precisely for unseen labels, and

$$\Pr(\mathcal{E}) \ge 3/4, \qquad \mathcal{E} = \{\|\boldsymbol{\mu}\| \le 4/\sqrt{n}\}, \tag{I.24}$$

since $\mathbb{E}\|\boldsymbol{\mu}\|^2 = 4/n - 2/n^2 \le 4/n$.

Lemma I.10 (All-step invariants)**.** *On every sample in $\mathcal{E}$, for every $0 < \eta \le 1$ and every update, the actual iterates satisfy*

$$\sigma_i x_{j,i} \ge \frac{3}{4nm_n}\big[1 - (1 - \eta m_n/K)^j\big], \tag{I.25}$$

$$\|\boldsymbol{x}_j\| \le 5/A, \qquad \boldsymbol{z}_j = r_j \boldsymbol{d}_{\boldsymbol{\sigma}}, \qquad 0 \le r_j \le R_+. \tag{I.26}$$

Every training reader and its gradient are zero. Every unprojected next candidate is strictly in the full unit ball, so the actual projection is inactive. Moreover

$$r_j \le \frac{\beta}{\rho}(a^j - 1) \le \beta(j\eta)e^{\rho j\eta}. \tag{I.27}$$

The analogous continuous statements hold for flow.

Proof. We verify all assertions jointly. Given the bounds at a query, Lemma I.8 makes each training reader zero, including at the origin. Both Huber blocks are quadratic there. With $h = \eta/K$, the signed memory forcing is at least $|\mu_i| - a_m \ge 3/(4n)$. The vector recurrence preserves

$$\|\boldsymbol{x}^+\| \le (1 - hm_n)\|\boldsymbol{x}\| + h(\|\boldsymbol{\mu}\| + a_m\sqrt{2n}) \le 5/A.$$

It also gives (I.25) and the correct signs. These comparisons do not assume a fixed decoder branch. The compatible-sign proximal reduction makes the code candidate parallel to $\boldsymbol{d}_{\boldsymbol{\sigma}}$, with

$$r^+ = (1 - hA_{\boldsymbol{z}})r + hB_{\boldsymbol{z}} + \eta\rho r - \eta W'(r).$$

Because $W'(r) \le \omega r$ and $hA_{\boldsymbol{z}} \le 100A/K = 25/32$, this coefficient is nonnegative. Since the signed scalar decoder derivative has absolute value at most b,

$$r^+ \le U(r) := r + \eta\beta + \eta\rho r - \eta\omega(r - R)_+.$$

The map U is increasing on $[0, R_+]$, and $U(R_+) \le R_+$ because $\beta \le \varepsilon\rho R$ and $(2+\varepsilon)\rho \le \omega$. Hence the interval is invariant. The combined next candidate has squared norm at most $(5/A)^2 + R_+^2 < 1$. Thus projection does not change it, closing the joint induction. At initialization any compatible sign vector may be used because the code state and mask gradient are zero. After the first memory update the signs are $\boldsymbol{\sigma}$. Dropping the nonnegative brake and using the decoder bound gives $r_{j+1} \le ar_j + \eta\beta$, proving (I.27). The same invariant-boundary inequalities prove the flow assertions. □

I.5.3 Plant the seed, then release the branch condition

Lemma I.11 (A seed by time $9/\rho$)**.** *On $\mathcal{E}$ there is an index j_0 with $\eta j_0 \le 9/\rho$ such that $r_{j_0} \ge 4z_*$. Until that first crossing, every query with $j\eta \ge 8K\lambda$ obeys the strict branch certificate (I.18). The brake is inactive there, and the actual scalar update is exactly*

$$r_{j+1} = ar_j + \eta\beta. \tag{I.28}$$

For flow the seed is reached by $9/\rho$ with the analogous exact differential equation.

Proof. Write raw time $t = j\eta/K$. The memory bound implies that, when $t \ge 8\lambda$, the left side of (I.18) is at least

$$\frac{t}{16n^2} \quad (t \le 1/m_n), \qquad \frac{1}{12n^2 m_n} \quad (t \ge 1/m_n).$$

Indeed $1 - e^{-m_n t} \ge m_n t/2$ in the first range, and $\lambda m_n \le 1/100$ supplies the slack in the second. Consider only times $j\eta \le 9/\rho$ before the first seed crossing. In the first range, (I.27) bounds the right side of the certificate by

$$2b(r_j + \lambda b) \le (2e^9 + 1/4)b^2(j\eta)/K < (j\eta)/(16Kn^2),$$

using $b^2 \le \varepsilon^2/n^2$ and $j\eta \ge 8K\lambda$. In the second range, $r_j < 4z_* = 4b/(K\rho)$ gives

$$2b(r_j + \lambda b) \le (8 + 2\lambda K\rho)\frac{b^2}{K\rho} \le \frac{9\varepsilon^2}{n^2 m_n} < \frac{1}{12n^2 m_n}.$$

Here $b^2 \le \varepsilon^2 K\rho/(n^2 m_n)$ is precisely the second branch of (I.21). Also $r_j < 4\varepsilon R < R$, so the brake is zero and (I.28) follows from the true gradient.

Let $J = \lceil 8K\lambda/\eta \rceil$ and $J' = \lceil 8/(\eta\rho) \rceil$. If no seed crossing occurred by J', the exact updates would imply $r_{J'} \ge z_*(a^{J'-J} - 1)$. Since $J\eta < 12$, $J'\eta \ge 8/\rho$, and $\log(1+\eta\rho) \ge \eta\rho/(1+\eta\rho)$, the exponent is at least $(8 - 12\rho)/(1+\rho) > \log 5$ under (I.20). This forces $r_{J'} > 4z_*$, a contradiction. Finally $J'\eta \le 8/\rho + 1 \le 9/\rho$. No permanent-branch assumption was made. □

Lemma I.12 (Exact-rate growth despite branch switching)**.** *On $\mathcal{E}$, for every index j with $j\eta \ge 9/\rho$,*

$$r_j \ge c_* \min\{1, \delta_{n,\rho} a^j\} \ge c_* d_j, \tag{I.29}$$

where $c_* > 0$ is universal. Flow obeys $r(t) \ge c_* \min\{1, \delta_{n,\rho} e^{\rho t}\}$ for $t \ge 9/\rho$.

Proof. Below R the worst possible signed decoder force is $+\beta$, so the actual updates satisfy

$$r^+ \ge ar - \eta\beta, \qquad r^+ - z_* \ge a(r - z_*).$$

This is an exact shift, not $r^+ \ge (1 + c\eta\rho)r$ for a reduced constant c. At and above R, the lower map $L(r) = ar - \eta\beta - \eta\omega(r - R)$ is increasing and $L(R) \ge R$ since $z_* < R$. Thus $[R, R_+]$ is forward invariant once entered. Lemma I.11 now gives, for $j \ge j_0$,

$$r_j \ge \min\{R, 3z_* a^{j-j_0}\} \ge \min\{R, 3e^{-9} z_* a^j\}.$$

The exact constants in (I.21) imply

$$z_* = \varepsilon \min\left\{\frac{1}{K\rho n}, \frac{1}{\sqrt{KA\rho n^{3/2}}}, R\right\} \geq \frac{\varepsilon}{K}\min\{R, \delta_{n,\rho}\}.$$

Consequently the preceding minimum is at least a universal positive multiple of $\min\{1, \delta_{n,\rho}a^j\}$. Finally (I.3) directly gives $d_j \leq \min\{1, \delta_{n,\rho}a^j\}$. The continuous shifted inequality $(r - z_*)' \geq \rho(r - z_*)$ gives the identical flow conclusion. □

I.5.4 Risk and the entire prefix

Proposition I.13 (One instance matches the amplified output contribution)**.** *For each $n \geq 2$ and $0 < \rho \leq \rho_0$, the family (I.22), fixed before sampling and independently of the final step/horizon, satisfies*

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] \geq cM(\widehat{\boldsymbol{w}})^2$$

for both actual GD outputs whenever $\rho q \geq 20$, uniformly over $0 < \eta \leq 1$. The same statement holds for flow.

Proof. For the last point, Lemmas I.9 and I.12 give the result on $\mathcal{E}$. The true excess is nonnegative off $\mathcal{E}$, whose probability is at least $3/4$. For the complete average, every code iterate has the same ray direction and nonnegative coefficient, so $\bar{\boldsymbol{z}}_T = \bar{r}_T \boldsymbol{d}_{\boldsymbol{\sigma}}$ exactly. Put $J_* = \lceil 9/(\eta\rho)\rceil$. Since $\rho q \geq 20$, $J_*/T \leq (9+\rho)/20 < 1/2$. The nondecreasing sequence d_j has at least half its mass on the suffix $j \geq J_*$. Therefore

$$\bar{r}_T = \frac{1}{T}\sum_{j=1}^{T} r_j \geq \frac{c_*}{T}\sum_{j \geq J_*} d_j \geq \frac{c_*}{2}M_{\mathrm{av}}.$$

The denominator stays T. No output is replaced by a tail average. Apply the conic population readout directly at $\bar{r}_T \boldsymbol{d}_{\boldsymbol{\sigma}}$, even when $\bar{r}_T$ is much smaller than the late iterate radii. The same suffix argument applies to the full time integral. □

Completion of the lower half of Theorem I.1. For $\rho q \leq 20$, the upper expression (I.5) is at most a universal multiple of the convex risk law: $K_T^{(2)} \leq e^{40}q$, $M(\widehat{\boldsymbol{w}})^2 \leq d_T^2 \leq e^{40}q/n^{3/2}$, and $\rho \leq 20/q$. The contained convex class and Theorem 2.1 at $L = 1$ supply the matching lower in $O(n)$ dimensions. The constant e^{40} is fixed. This argument is not used when ρq grows without bound. For $\rho q > 20$ and $\rho \leq \rho_0$, use Proposition I.13 and take the maximum with the separate affine optimization/statistical witnesses from Appendix C.7 and the deterministic loss $f(\boldsymbol{w}) = -\rho J(\boldsymbol{w})$. The last loss keeps all iterates at zero and has excess $\rho/2$ on the ball, for every output and time. A maximum of four nonnegative contributions dominates their capped sum up to a constant. For $\rho \geq \rho_0$, that same deterministic witness has excess at least $\rho_0/2$, while $\Psi_\eta(\widehat{\boldsymbol{w}}) \leq 1$. This completes the nontrivial small-ρ result to all $\rho \in [0, 1]$ with universal constants. Flow follows identically. □

I.5.5 Pathwise negative curvature and numerical oracles

Set $J = \lceil 8K\lambda/\eta\rceil$. The identified update adds $\eta\beta > 0$, so $r_{J+1} > 0$. Its physical time is below 13 and

$$r_{J+1} \leq (\beta/\rho)(e^{13\rho_0} - 1) < 4\beta/\rho < R.$$

This query remains in the strict identified region. At this positive radius the decoder is locally affine, every training reader is locally zero, and the brake is inactive. The empirical code-block Hessian is exactly $-\rho\mathrm{Id}$. Thus the actual path queries negative curvature. No Hessian is asserted at the cone vertex.

The memory-code proximal reduction requires only one evaluation of $\boldsymbol{d}_{\mathrm{sign}(\boldsymbol{x})}$, which is constructive. Lemma I.8 supplies the new reader's global numerical reduction. The Huber and brake terms have explicit radial formulas. For rational inputs and rational $\rho > 0$ of bit length at most B_0, values and sample gradients to accuracy 2^{-p} therefore take $\mathrm{poly}(n, B_0, p)$ bit operations, using certified scalar searches and interval evaluation of the finitely many square roots. This claim is per sample query. It does not bound the number of updates for arbitrarily small η. For a finite implemented horizon, the exact projected update is a-Lipschitz, where $a = 1 + \eta\rho$, by the convexification argument in

Lemma I.5. Keep implemented iterates feasible. If the gradient error at step j is at most ε_j and the additional update or rounding error is at most ζ_j, the distance e_j to the exact iterate, starting from $e_0 = 0$, satisfies

$$e_{j+1} \le a e_j + \eta\varepsilon_j + \zeta_j, \qquad e_T \le \sum_{j=0}^{T-1} a^{T-1-j}(\eta\varepsilon_j + \zeta_j). \tag{I.30}$$

For uniform bounds $\varepsilon_j \le \varepsilon_g$ and $\zeta_j \le \zeta$, this is $K_T^{(1)}\varepsilon_g + (K_T^{(1)}/\eta)\zeta$, with $K_T^{(1)} = q$ at $\rho = 0$. Since $K_T^{(1)} \le qe^{\rho q}$, gradient error alone requires $O(\rho q + \log(2+q) + \log(1/\epsilon))$ additional bits for population risk tolerance $0 < \epsilon < 1$. The same bound suffices when each update error is at most $\eta\varepsilon_g$. A fixed rounding tolerance instead requires an additional $\log(1/\eta)$ bits. The prefix version of (I.30) also controls the full parameter average and an independent uniform checkpoint, by 1-Lipschitzness of F. When $\rho q = O(\log n)$ and q and $1/\epsilon$ are polynomial in n, the gradient precision is logarithmic in n. The oracle is defined off the good event and off the intended trajectory as well.

I.6 The discrete clock

Corollary I.14 (No step-uniform last-iterate law using only q)**.** *There is no positive function $\Phi(n, q, \rho)$ and universal constants c, C such that $c\Phi(n, \eta T, \rho) \le \mathcal{R}^{\mathrm{wc}}_{n,d,\rho}(\boldsymbol{w}_T) \le C\Phi(n, \eta T, \rho)$ simultaneously for all safe constant steps in sufficient linear dimension.*

Proof. Let $\ell_n = \log n$, $\rho_n = \ell_n^{-1/2}$, and choose the integer

$$q_n = \left\lceil \frac{\log(n\rho_n)}{\rho_n} \right\rceil.$$

Eventually $\rho_n \le \rho_0$ and $\rho_n\sqrt{n} > 1$, so $\delta = 1/(n\rho_n)$. Compare $\eta_1 = 1, T_1 = q_n$ with $\eta_2 = 1/q_n, T_2 = q_n^2$, writing $\boldsymbol{w}_{T_1}^{(1)}$ and $\boldsymbol{w}_{T_2}^{(2)}$ for their endpoint rules. Both have exactly the same cumulative time q_n. For the second step,

$$x_2 = q_n^2 \log(1 + \rho_n/q_n) = \rho_n q_n - \rho_n^2/2 + o(\rho_n^2),$$

so $\delta^2 e^{2x_2} = \Theta(1)$. For the first,

$$x_1 = q_n \log(1 + \rho_n) = \rho_n q_n - \tfrac{1}{2}\rho_n^2 q_n + O(\rho_n^3 q_n),$$

and $\rho_n^2 q_n = \sqrt{\ell_n} + o(\sqrt{\ell_n})$, while $\rho_n^3 q_n = O(1)$. Consequently $\delta^2 e^{2x_1} = \exp[-\sqrt{\ell_n} + O(1)]$. The baseline is $\Theta(\rho_n)$. Equation (I.11), suppressing the common n, d, ρ_n, therefore gives

$$\mathcal{R}^{\mathrm{wc}}(\boldsymbol{w}_{T_2}^{(2)}) = \Theta(1), \qquad \mathcal{R}^{\mathrm{wc}}(\boldsymbol{w}_{T_1}^{(1)}) = \Theta((\log n)^{-1/2}).$$

Their ratio diverges, contradicting a common Φ. This is a separation of actual worst-instance population risks, not just two stability certificates. □

A sufficient condition for replacing the last-point discrete exponential by its flow value is $\eta\rho^2 q = O(1)$, since $0 \le \rho q - x_\eta \le \eta\rho^2 q/2$. This condition is sufficient, not claimed necessary whenever baseline or saturation masks the difference. The averaged law needs no such restriction, by Corollary I.4.

I.7 Uniform checkpoint selection and completion of the main theorem

Let J be uniform on $\{1, \dots, T\}$ and independent of the sample, and use the same J in the original and replaced worlds. Set $V_{\mathrm{pick}} = V(\boldsymbol{w}_J) = T^{-1}\sum_{j=1}^{T} d_j^2$. For every admissible parameter choice,

$$\mathcal{R}^{\mathrm{wc}}_{n,d,\rho}(\boldsymbol{w}_J) \asymp \min\{1, q^{-1} + n^{-1/2} + \rho + V_{\mathrm{pick}}\}. \tag{I.31}$$

The upper is dimension independent and the lower holds for $d \ge Cn$.

Proof. Keep J fixed under replacement and integrate over both the sample and this independent seed. The original-loss risk bound extends to this joint space, giving $D_{\mathrm{pick}}^2 = T^{-1}\sum_j v_j \le 192 V_{\mathrm{pick}}$. The empirical comparison is the average of the objective values, and summing the actual three-point inequality gives $\mathbb{E}_J[F_S(\boldsymbol{w}_J) - F_S(\boldsymbol{w}^*)] \le (2q)^{-1} + 2\rho$. It is important to use this sum directly, rather than summing the looser individual $1/(\eta j)$ bounds. Lemma I.7 allows this expected comparison and the shared independent seed, and proves the upper.

For $0 < \rho \le \rho_0$ and $\rho q \ge 20$, the same amplifier supplies, conditionally on its good sample event,

$$\mathbb{E}_J[F(\boldsymbol{w}_J) - F^*] \ge \frac{c}{T} \sum_{j=\lceil 9/(\eta\rho)\rceil}^{T} d_j^2 \ge \frac{c}{2T}\sum_{j=1}^{T} d_j^2.$$

The final inequality uses monotonicity of d_j^2 and the fact that the suffix occupies at least half the indices. Off the good event true excess is nonnegative. If $\rho q \le 20$, then $V_{\mathrm{pick}} \le C\min\{1, q/n^{3/2}\}$ and $\rho \le 20/q$. The convex complete-average lower from Theorem 2.1 applies: for each of its convex instances, $\mathbb{E}_J F(\boldsymbol{w}_J) \ge F(T^{-1}\sum_j \boldsymbol{w}_j)$ by convexity. Thus the same convex lower bounds uniform checkpoint selection. No new temporal constraint is needed. The separate affine optimization/statistical and deterministic $-\rho J$ witnesses complete the bound as in Theorem I.1. The case $\rho \ge \rho_0$ follows from the last witness and the cap. Flow uses the uniform draw from $[0, q]$, with identical integral arguments. □

Together with Theorem I.1, this proves Theorem 3.1. In the amplified regime $\rho q > 1$, use δ and x_η from (I.8). The same geometric-sum comparison as Corollary I.3 gives

$$V_{\mathrm{pick}} \asymp \mathcal{A}(\delta^2, 2x_\eta), \qquad M_{\mathrm{av}}^2 \asymp \mathcal{A}(\delta, x_\eta)^2.$$

Replacing x_η by ρq changes the first integrated quantity by at most $\log(\eta\rho/\log(1+\eta\rho)) \le \eta\rho \le \rho$. The ρ floor absorbs this difference. Corollary I.4 supplies the corresponding statement for the parameter average. This explains why both uniform readouts have step-uniform q-only orders, whereas Corollary I.14 excludes such a last-point law.

I.8 Proof of Corollary 3.2

Proof. For sufficiently large n, set $\rho = 1/n \le \rho_0$ and fix the loss family (I.22) with its uniform distribution on $[2n]$. Lemma I.9 verifies its global regularity and dimension $O(n)$. This instance is chosen before sampling and independently of the step size and both horizons. Let $q_n = (n/4)\log n$. For every $0 < \eta \le 1$, put $T = \lceil q_n/\eta \rceil$, $q_\eta = \eta T$, and $x_\eta = T\log(1+\eta/n)$. Then, uniformly in η,

$$0 \le q_\eta - q_n < \eta, \qquad x_\eta = h + O((\log n)/n), \qquad h = \tfrac{1}{4}\log n.$$

Here $\delta = \delta_{n,\rho} = n^{-1/4}$ and $h = \log(1/\delta)$, so

$$\delta^2 e^{2h} = 1, \qquad \mathcal{A}(\delta^2, 2h) = \frac{1-\delta^2}{2h}, \qquad \mathcal{A}(\delta, h)^2 = \frac{(1-\delta)^2}{h^2}.$$

The response-kernel comparisons in Corollary I.3 and Section I.7 therefore give

$$M_{\mathrm{last}}^2 \asymp 1, \qquad V_{\mathrm{pick}} \asymp h^{-1}, \qquad M_{\mathrm{av}}^2 \asymp h^{-2}.$$

Since $q_\eta^{-1} + n^{-1/2} + \rho = o(h^{-2})$, the upper bounds in Theorem 3.1 give the three claimed upper rates on this instance.

For the matching lower bounds, $\rho q_\eta \ge 20$ for all sufficiently large n. Proposition I.13 supplies the last-iterate and full-average bounds for the fixed family above. For the independent uniform checkpoint, the same trajectory and good event give, as in Section I.7,

$$\mathbb{E}[F(\boldsymbol{w}_J) - F^*] \ge \frac{c}{T} \sum_{j=\lceil 9/(\eta\rho)\rceil}^{T} d_j^2 \ge c' V_{\mathrm{pick}}.$$

The good event has probability at least $3/4$, and the suffix contains at least half of the nondecreasing sequence's total squared mass. These bounds prove (10) for the three outputs of the same trajectory, with constants uniform over $0 < \eta \le 1$.

At the earlier horizon $T_0 = \lceil \sqrt{n}/\eta \rceil$, let $q_0 = \eta T_0 \in [\sqrt{n}, \sqrt{n} + \eta)$. Then $\rho q_0 = O(n^{-1/2})$ and $K_j^{(2)} \le C\eta j$ for $j \le T_0$. Hence every output contribution satisfies $V(\widehat{\boldsymbol{w}}) \le d_{T_0}^2 \le C q_0 / n^{3/2} = O(n^{-1})$. The same class-uniform upper bound gives $O(n^{-1/2})$ risk for all three outputs of this fixed instance at T_0.

For constrained flow, use the same instance at times q_n and $\sqrt{n}$. The clock is exactly $h = \rho q_n$. The preceding kernel comparisons and the amplifier bounds hold with integrals in place of sums. The uniform checkpoint is sampled independently from the corresponding time interval. This proves the flow statement as well. □

J SCALAR-COORDINATE LOWER BOUNDS FOR PAIRWISE LEARNING

We prove the lower bound for uniform scalar-coordinate updates on the complete U-statistic of the original observations. After the deterministic invariants and gradient identification, the essential stochastic step is (J.28): binomial marginal tail bounds and a union bound, without independence between coordinate visit counts. Equations (J.30)–(J.31) charge the entire preheating prefix to the complete post-update average. The clock is $q = \eta T/D$. For $D = 66n$, $\eta = 1$, the comparison horizon $q \asymp n^{4/3}$ uses $T \asymp n^{7/3}$ scalar updates. The estimation example below first explains the first-projection parameter.

J.1 A LEARNING INTERPRETATION OF THE FIRST PROJECTION

Proposition J.1 (Estimating a squared mean)**.** *Let $Z \in [-1,1]$, $\mu = \mathbb{E}Z$, and $\sigma^2 = \mathrm{Var}(Z)$. On $w \in [-1,1]$, consider the symmetric loss*

$$f(w; z, z') = \tfrac{1}{2}\mathsf{H}_2(w - zz'), \qquad \mathsf{H}_2(t) = \begin{cases} t^2/2, & |t| \le 2, \\ 2|t| - 2, & |t| > 2. \end{cases}$$

The loss is globally nonnegative, convex, 1-Lipschitz and $1/2$-smooth. Its population minimizer is $w^* = \mu^2$, and the first and second Hoeffding projections of its gradient at w^* are

$$h_1(z) = \tfrac{1}{2}\mu(\mu - z), \qquad h_2(z, z') = -\tfrac{1}{2}(z - \mu)(z' - \mu), \qquad \|h_1\|_{L^2} = \tfrac{1}{2}|\mu|\sigma. \tag{J.1}$$

When $\mu = 0$ and $\sigma > 0$, the first projection vanishes, whereas $\mathbb{E}\|\nabla f(w^*; Z, Z')\|^2 = \sigma^4/4$ and $F^* = \sigma^4/4$ are strictly positive.

Proof. For $w, z, z' \in [-1,1]$, the residual lies in the quadratic region. Hence

$$F(w) = \tfrac{1}{4}(w - \mu^2)^2 + \tfrac{1}{4}\big((\mathbb{E}Z^2)^2 - \mu^4\big), \qquad a(z, z') = \tfrac{1}{2}(\mu^2 - zz').$$

Taking a conditional expectation over Z' proves the first formula in (J.1). Subtracting its two first projections proves the second. The stated moments follow by independence. The global regularity follows directly from the Huber derivative, whose absolute value is at most 2 and whose Lipschitz constant is 1. □

Thus a small first projection can express cancellation in a meaningful estimation target, without a small variance for a single pair gradient or samplewise interpolation. Additive shifts of the loss change neither projection. This example explains the parameter, but it is not a witness for late overfitting. Indeed, its empirical minimizer is the U-statistic $w_{\mathrm{emp}} = \binom{n}{2}^{-1} \sum_{i<j} Z_i Z_j \in [-1,1]$, and

$$\mathbb{E}[F(w_{\mathrm{emp}}) - F^*] = \frac{\mu^2\sigma^2}{n} + \frac{\sigma^4}{2n(n-1)}.$$

Its strong convexity allows faster estimation than the worst-case convex class studied below.

J.2 ALGORITHM, CLASS, AND QUANTIFIERS

Fix $n \ge 64$, and put $k = 2n$, $r = 64n$, and $D = k + r = 66n$. The feasible set is the fixed coordinate box

$$\mathcal{K}_n = [-1/(2\sqrt{k}), 1/(2\sqrt{k})]^k \times [-1/(2\sqrt{r}), 1/(2\sqrt{r})]^r \subset \mathbb{B}^D. \tag{J.2}$$

Write $\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{y})$ for its memory and code blocks. For i.i.d. observations $Z_1, \dots, Z_n$, let

$$G_S(\boldsymbol{w}) = \binom{n}{2}^{-1} \sum_{i<j} f(\boldsymbol{w}; Z_i, Z_j), \qquad F(\boldsymbol{w}) = \mathbb{E} f(\boldsymbol{w}; Z, Z').$$

Initialize $\boldsymbol{X}_0 = 0$. Independently of the data, sample I_t i.i.d. uniformly from $[D]$, and perform the scalar update

$$X_{t+1,I_t} = \Pi_{(\mathcal{K}_n)_{I_t}}(X_{t,I_t} - \eta \partial_{I_t} G_S(\boldsymbol{X}_t)), \qquad X_{t+1,j} = X_{t,j} \quad (j \neq I_t). \tag{J.3}$$

Here $0 < \eta \leq 1$, and $T \geq 1$ and η are chosen before sampling. The two output rules are $\boldsymbol{X}_T$ and $\bar{\boldsymbol{X}}_T = T^{-1} \sum_{t=1}^{T} \boldsymbol{X}_t$, and $q = \eta T / D$.

For $0 \leq \nu \leq 1$, let $\mathcal{P}_\nu(\mathcal{K}_n)$ be the class of i.i.d. data laws and symmetric pair kernels that are globally nonnegative, convex, 1-Lipschitz and 1-smooth on $\mathbb{R}^D$, and that have a deterministic minimizer $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w} \in \mathcal{K}_n} F(\boldsymbol{w})$ at which

$$\left(\mathbb{E} \left\| \mathbb{E}[\nabla f(\boldsymbol{w}^*; Z, Z') \mid Z] - \mathbb{E} \nabla f(\boldsymbol{w}^*; Z, Z') \right\|^2 \right)^{1/2} \leq \nu.$$

Set $s_\nu = n^{-1} + \nu n^{-1/2}$.

Theorem J.2 (Scalar-coordinate lower bound and persistence)**.** *There are universal $c, C > 0$ such that, for either output rule separately, every $n \geq 64$, $0 \leq \nu \leq 1$, $0 < \eta \leq 1$, and integer $T \geq 1$ obey*

$$\sup_{(P,f) \in \mathcal{P}_\nu(\mathcal{K}_n)} \mathbb{E}[F(\widehat{\boldsymbol{w}}) - \min_{\mathcal{K}_n} F] \geq c \min \left\{ 1, \frac{1}{q} + s_\nu + \frac{q s_\nu}{n} \right\}. \tag{J.4}$$

More strongly, for every prescribed $q_0 \geq C / s_\nu$, one may choose one kernel and one data distribution before sampling, independently of the eventual step size, such that for both outputs and every $0 < \eta \leq 1$, $\eta T / D \geq q_0$,

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - \min_{\mathcal{K}_n} F] \geq c \min \left\{ 1, \frac{q_0 s_\nu}{n} \right\}. \tag{J.5}$$

The dimension is $66n$.

The supremum is over instances of the specified learner. In (J.5), one instance is fixed once n, ν, q_0 are fixed. Expectation is over the i.i.d. data and actual coordinate choices.

J.3 A CANONICAL DRIVE THAT RECORDS UNSEEN LABELS

The observation space for the late-time example is $[k]$, with its uniform distribution. Fix balanced signs $\gamma_j \in \{-1, 1\}$, with n signs of either kind, and define

$$\boldsymbol{\delta}_j = n^{-1} \mathbf{1} - 2 \boldsymbol{e}_j, \qquad \boldsymbol{d}_{ij} = \tfrac{1}{2}(\gamma_i \boldsymbol{\delta}_j + \gamma_j \boldsymbol{\delta}_i), \qquad \boldsymbol{b}^\epsilon_{ij} = \boldsymbol{d}_{ij} + \tfrac{\epsilon}{2}(\boldsymbol{\delta}_i + \boldsymbol{\delta}_j), \quad 0 \leq \epsilon \leq 1. \tag{J.6}$$

Lemma J.3 (Exact U-statistic drive)**.** *For independent uniform $I, J \in [k]$,*

$$\mathbb{E} \boldsymbol{\delta}_I = 0, \quad \mathbb{E}_J \boldsymbol{d}_{iJ} = 0, \quad \mathbb{E} \|\boldsymbol{d}_{IJ}\|^2 = 2, \quad \|\boldsymbol{b}^\epsilon_{ij}\| \leq 4.$$

Given the original n observations, write N_j for the count of label j, $R = \sum_{t=1}^{n} \gamma_{Z_t}$, $\boldsymbol{\mu}_S = n^{-1} \mathbf{1} - 2\boldsymbol{N}/n$, and $\boldsymbol{m}_S = \binom{n}{2}^{-1} \sum_{s<t} \boldsymbol{b}^\epsilon_{Z_s Z_t}$. Then

$$(\boldsymbol{m}_S)_j = \epsilon \frac{1 - 2N_j}{n} + \frac{R}{n^2} - \frac{2(R - \gamma_j) N_j}{n(n-1)}. \tag{J.7}$$

There is a data event $\mathcal{E}$ of probability at least $1/16$ on which, with $\alpha = \epsilon + n^{-1/2}$,

$$\operatorname{sign}(\boldsymbol{m}_S)_j = +1 \iff N_j = 0, \qquad \min_j |(\boldsymbol{m}_S)_j| \geq \frac{\alpha}{4n}, \qquad \|\boldsymbol{m}_S\| \leq \frac{88\alpha}{\sqrt{n}}. \tag{J.8}$$

Proof. We have $\|\boldsymbol{\delta}_i\|^2 = 4(1 - 1/k)$ and $\|\mathbb{E}[\gamma_I \boldsymbol{\delta}_I]\|^2 = 4/k$. The conditional mean of $\boldsymbol{d}$ is zero, because both $\mathbb{E}\gamma_I$ and $\mathbb{E}\boldsymbol{\delta}_I$ vanish. Expanding the square and using independence gives

$$\mathbb{E}\|\boldsymbol{d}_{IJ}\|^2 = \tfrac{1}{2}\mathbb{E}\|\boldsymbol{\delta}_I\|^2 + \tfrac{1}{2}\|\mathbb{E}[\gamma_I \boldsymbol{\delta}_I]\|^2 = 2.$$

Also $\|\boldsymbol{\delta}_i\| \le 2$ and $\|\boldsymbol{d}_{ij}\| \le 2$, proving the pointwise bound. Summing the unordered pairs, with their original indices, gives

$$\binom{n}{2}^{-1} \sum_{s<t} \boldsymbol{d}_{Z_s Z_t} = \frac{R}{n-1}\boldsymbol{\mu}_S - \frac{\sum_t \gamma_{Z_t}\boldsymbol{\delta}_{Z_t}}{n(n-1)}.$$

Adding $\epsilon\boldsymbol{\mu}_S$ and taking coordinate j proves (J.7), including the repeated-label and off-diagonal corrections.

Take

$$\mathcal{E} = \{\sqrt{n}/2 \le R \le 4\sqrt{n},\ \|\boldsymbol{\mu}_S\| \le 16/\sqrt{n}\}.$$

The summands defining R are independent Rademacher variables, so $\mathbb{E}R^2 = n$ and $\mathbb{E}R^4 \le 3n^2$. Paley–Zygmund applied to R^2 and Markov applied to R^4 yield

$$\Pr\{\sqrt{n}/2 \le |R| \le 4\sqrt{n}\} \ge 3/16 - 3/256 = 45/256.$$

Moreover $\mathbb{E}\|\boldsymbol{\mu}_S\|^2 = 4/n - 2/n^2 \le 4/n$, whence $\Pr\{\|\boldsymbol{\mu}_S\| > 16/\sqrt{n}\} \le 1/64$. Swapping the two label halves changes R to $-R$ and preserves $\|\boldsymbol{\mu}_S\|$, so $\Pr(\mathcal{E}) \ge 41/512 > 1/16$.

On this event $R \ge 4$. An unseen coordinate in (J.7) is at least $\epsilon/n + 1/(2n^{3/2})$. For a seen coordinate, the canonical part is at most

$$\frac{R}{n^2} - \frac{2(R-1)}{n(n-1)} \le -\frac{R}{2n^2}.$$

The last inequality is equivalent to $R(n+3) \ge 4n$. Its additive part is at most $-\epsilon/n$. These estimates give the sign and minimum-magnitude claims. Finally, the preceding vector identity and $\|\boldsymbol{\delta}_i\| \le 2$ give

$$\left\| \binom{n}{2}^{-1} \sum_{s<t} \boldsymbol{d}_{Z_s Z_t} \right\| \le \frac{64+2}{n-1} \le \frac{88}{n}.$$

Adding $\epsilon\|\boldsymbol{\mu}_S\| \le 16\epsilon/\sqrt{n}$ proves the norm bound in (J.8). □

J.4 A COORDINATE-INVARIANT DECODER

Choose and fix unit binary codewords

$$\boldsymbol{c}_{\boldsymbol{v}} \in \{-1/\sqrt{r}, 1/\sqrt{r}\}^r, \quad \boldsymbol{v} \in \{-1,1\}^k, \qquad \langle \boldsymbol{c}_{\boldsymbol{u}}, \boldsymbol{c}_{\boldsymbol{v}} \rangle \le 1/2 \quad (\boldsymbol{u} \ne \boldsymbol{v}). \tag{J.9}$$

Their existence at $r = 32k$ is elementary. Independently sampled binary words violate the inequality for a specified pair with probability at most $e^{-r/8} = e^{-4k}$, by the Rademacher tail bound. There are fewer than 2^{2k} pairs, so the union bound is strictly less than one. We fix a successful family. Code selection is not algorithmic randomness. This existence argument alone makes no efficient-oracle claim.

For $a, b > 0$, use the mask already analyzed in Lemma E.1:

$$\begin{aligned} H(\boldsymbol{x}, \boldsymbol{y}) &= a(\|\boldsymbol{x}\|_1 - \min_j |x_j|) + b\|\boldsymbol{y}\|, \\ L_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{y}) &= a\langle \boldsymbol{v}, \boldsymbol{x} \rangle - b\langle \boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{y} \rangle, \\ P(\boldsymbol{x}, \boldsymbol{y}) &= \max\{H(\boldsymbol{x}, \boldsymbol{y}), \max_{\boldsymbol{v}} L_{\boldsymbol{v}}(\boldsymbol{x}, \boldsymbol{y})\}. \end{aligned} \tag{J.10}$$

Let P_ℓ denote its Moreau envelope with parameter ℓ. The precise inputs we use from the earlier mask results are: $P \ge 0$, $P(0) = 0$, its Lipschitz constant is $\sqrt{ka^2 + b^2}$, every envelope-gradient memory coordinate has magnitude at most a, and the code-gradient norm is at most b. Theorem E.2 gives the stronger proximal identity

$$\operatorname{prox}_{\ell P}(\boldsymbol{x}, \boldsymbol{y}) = \operatorname{prox}_{\ell \max\{H, L_{\boldsymbol{\sigma}}\}}(\boldsymbol{x}, \boldsymbol{y}) \quad \text{if } \boldsymbol{\sigma} \text{ agrees with all nonzero signs of } \boldsymbol{x}. \tag{J.11}$$

The proof of Theorem E.2 shows that zero input memory coordinates remain zero under the prox. Thus (J.11) also applies before every memory coordinate has been selected.

Lemma J.4 (Coordinate-box invariance). *For a compatible $\boldsymbol{\sigma}$ and any input $(\boldsymbol{x}, \boldsymbol{y})$, there are scalars $A_y \in [0, \ell^{-1}]$ and $B_y \in [0, b]$ such that*

$$\nabla_{\boldsymbol{y}} P_\ell(\boldsymbol{x}, \boldsymbol{y}) = A_y \boldsymbol{y} - B_y \boldsymbol{c}_{\boldsymbol{\sigma}}. \tag{J.12}$$

Let $\lambda_c > 0$ and $\Gamma = b/\lambda_c$. If $y_j = (\boldsymbol{c}_{\boldsymbol{\sigma}})_j u_j$ with $0 \le u_j \le \Gamma$, a gradient step in any single code coordinate for $P_\ell(\boldsymbol{x}, \boldsymbol{y}) + \lambda_c \|\boldsymbol{y}\|^2/2$ preserves these intervals whenever $h \le (\ell^{-1} + \lambda_c)^{-1}$. On this entire box,

$$\|\boldsymbol{y}\| \le \Gamma, \qquad \langle \boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{y} \rangle \le 3\Gamma/4 \quad (\boldsymbol{v} \ne \boldsymbol{\sigma}). \tag{J.13}$$

Proof. Use (J.11), and let $\boldsymbol{v}$ be the code block of the proximal minimizer. Optimality for the maximum of its two convex branches supplies $t \in [0, 1]$ and $\boldsymbol{g} \in \partial \|\boldsymbol{v}\|$ such that

$$\boldsymbol{y} - \boldsymbol{v} = \ell b \{ t\boldsymbol{g} - (1 - t)\boldsymbol{c}_{\boldsymbol{\sigma}} \}.$$

This is part of the joint proximal condition. The memory variable has not been held at its input. If $\boldsymbol{v} \ne 0$, then

$$\boldsymbol{v} = \rho \{ \boldsymbol{y} + \ell b (1 - t) \boldsymbol{c}_{\boldsymbol{\sigma}} \}, \qquad \rho = (1 + \ell b t / \|\boldsymbol{v}\|)^{-1} \in (0, 1].$$

Thus $A_y = (1 - \rho)/\ell$ and $B_y = \rho b (1 - t)$ work. If $\boldsymbol{v} = 0$, take $A_y = \ell^{-1}$ and $B_y = 0$.

In a selected coordinate the exact step becomes

$$u_j^+ = (1 - h(\lambda_c + A_y)) u_j + h B_y.$$

It is nonnegative under the step condition, and is at most $(1 - h\lambda_c)\Gamma + hb = \Gamma$. The other coordinates do not change. Since the binary codewords have entries of equal magnitude, $\|\boldsymbol{y}\|^2 = r^{-1} \sum_j u_j^2 \le \Gamma^2$. For a wrong code, the number of matching signs is at most $3r/4$ by (J.9). Dropping the nonpositive terms at the other coordinates gives the second inequality in (J.13). In particular, no assumption that $\boldsymbol{y}$ is parallel to $\boldsymbol{c}_{\boldsymbol{\sigma}}$ is needed. □

Fix $\kappa = 9/10$, $\zeta = 19/20$, and $\Delta = \zeta - \kappa = 1/20$. For a label i, define

$$h_{i,\theta}(\boldsymbol{y}) = \max \left\{ (\|\boldsymbol{y}\| - \theta)_+, \max_{\boldsymbol{v}: v_i = +1} (\langle \boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{y} \rangle - \kappa\theta) \right\}, \qquad R_{i,\theta} = (h_{i,\theta})_1. \tag{J.14}$$

The readout is globally nonnegative, convex, 1-Lipschitz and 1-smooth. It and its gradient vanish wherever $h_{i,\theta} = 0$. For the risk lower bound we only need the following elementary fact, which does not require a decoding oracle or an exact proximal formula:

$$v_i = +1, \quad s = \langle \boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{y} \rangle - \kappa\theta \in [0, 1] \quad \Longrightarrow \quad R_{i,\theta}(\boldsymbol{y}) \ge s^2/2. \tag{J.15}$$

Indeed $h_{i,\theta} \ge (\langle \boldsymbol{c}_{\boldsymbol{v}}, \cdot \rangle - \kappa\theta)_+$. Moreau envelopes preserve pointwise order, and the parameter-one envelope of this unit-normal hinge has value $s^2/2$ for $0 \le s \le 1$.

J.5 One fixed globally admissible pair kernel

Use the universal constants and parameters

$$\begin{aligned} &A = 2048, \quad K_0 = 128A, \quad \ell = (100A)^{-1}, \quad L_0 = \log(4\zeta/\Delta), \\ &C_0 \ge 256\zeta(1 + L_0)/\Delta, \qquad 0 < \theta_0^2 \le \min\{1/64, (1600AC_0)^{-1}\}. \end{aligned} \tag{J.16}$$

For a prescribed raw horizon $Q_0 \ge C_0\sqrt{n}/\alpha$, set

$$a = \frac{\alpha}{16n}, \quad \lambda_m = \frac{A\alpha}{\sqrt{n}}, \quad \theta = \theta_0 \min\{\sqrt{Q_0\alpha/n^{3/2}}, 1\}, \quad \lambda_c = \frac{C_0}{Q_0}, \quad b = \zeta\theta\lambda_c. \tag{J.17}$$

For $t \ge 0$, let $H_1(t) = t^2/2$ if $t \le 1$ and $t - 1/2$ otherwise. Define the radial Huber function $J_0(\boldsymbol{u}) = H_1(\|\boldsymbol{u}\|)$ and the exterior function $\chi(\boldsymbol{w}) = H_1((\|\boldsymbol{w}\| - 1)_+)$. Both are convex, 1-Lipschitz and 1-smooth. The function χ and its gradient vanish on the unit ball. The kernel is

$$\begin{aligned} f_\epsilon(\boldsymbol{x}, \boldsymbol{y}; i, j) = K_0^{-1} \big[&P_\ell(\boldsymbol{x}, \boldsymbol{y}) + \lambda_m J_0(\boldsymbol{x}) + \lambda_c J_0(\boldsymbol{y}) + \tfrac{1}{2}(R_{i,\theta}(\boldsymbol{y}) + R_{j,\theta}(\boldsymbol{y})) \\ &- \langle \boldsymbol{b}_{ij}^\epsilon, \boldsymbol{x} \rangle + 5\chi(\boldsymbol{x}, \boldsymbol{y}) + 6 \big]. \end{aligned} \tag{J.18}$$

These are terms of the loss, not algorithmic regularizers. The loss depends only on n, ϵ, Q_0, fixed codes, and its two labels.

Lemma J.5 (Global properties and the true empirical objective)**.** *Every kernel in* (J.18) *is symmetric, globally nonnegative, convex,* 1*-Lipschitz and* 1*-smooth. Its population minimum on* $\mathbb{R}^D$*, hence on* $\mathcal{K}_n$*, is attained at zero and equals* $6/K_0$*. For every point, deterministically,*

$$F(\boldsymbol{x}, \boldsymbol{y}) - F(0) \geq K_0^{-1} \mathbb{E}_I R_{I,\theta}(\boldsymbol{y}). \tag{J.19}$$

At zero the gradient projections satisfy

$$\boldsymbol{h}_1(i) = -\frac{\epsilon}{2K_0}(\boldsymbol{\delta}_i, 0), \qquad \|\boldsymbol{h}_1\|_{L^2} = \frac{\epsilon}{K_0}\sqrt{1 - \frac{1}{k}}, \qquad \|\boldsymbol{h}_2\|_{L^2}^2 = \frac{2}{K_0^2}. \tag{J.20}$$

For the original complete-pair training objective, exactly,

$$\begin{aligned} K_0 G_S(\boldsymbol{x}, \boldsymbol{y}) =& P_\ell(\boldsymbol{x}, \boldsymbol{y}) + \lambda_m J_0(\boldsymbol{x}) + \lambda_c J_0(\boldsymbol{y}) \\ &+ \frac{1}{n}\sum_{t=1}^{n} R_{Z_t,\theta}(\boldsymbol{y}) - \langle \boldsymbol{m}_S, \boldsymbol{x} \rangle + 5\chi(\boldsymbol{x}, \boldsymbol{y}) + 6. \end{aligned} \tag{J.21}$$

Proof. Convexity is preserved by the positive coefficients and the linear term. Since $\|\boldsymbol{b}_{ij}^\epsilon\| \leq 4$, writing $r_* = \|(\boldsymbol{x}, \boldsymbol{y})\|$ gives a lower bound of $5H_1((r_* - 1)_+) - 4r_* + 6$ for the raw loss. For $r_* \leq 1$ this is at least 2. For $1 \leq r_* \leq 2$ its minimum is $2/5$ at $r_* = 9/5$. For $r_* \geq 2$ it equals $r_* - 3/2$. Thus the global loss is nonnegative.

The parameters and $n \geq 64$ imply $\lambda_m \leq A$, $\lambda_c \leq 1$, and $\sqrt{ka^2 + b^2} < 1$. The raw Lipschitz constant is consequently at most $A + 12 < K_0$, and its smoothness constant is at most $\ell^{-1} + A + 7 = 101A + 7 < K_0$. For completeness, the radial formulas for J_0 and χ have radial and tangential Hessian eigenvalues in $[0, 1]$ wherever the Hessians exist, and their gradients are continuous across the joining spheres. This verifies their asserted global smoothness.

Every nonlinear term is nonnegative and zero at the origin, and $\mathbb{E}\boldsymbol{b}_{IJ}^\epsilon = 0$. This proves the population assertions and (J.19). Each smooth nonlinear term has gradient zero at its global minimum. Thus the gradient at zero is just $-(\boldsymbol{b}_{ij}^\epsilon, 0)/K_0$. The seed identities prove (J.20). Finally, every training index occurs in exactly $n - 1$ unordered pairs. This gives the readout coefficient $1/n$ in (J.21), with repeated labels retained. □

J.6 The entire scalar-coordinate trajectory

Write $h = \eta/K_0$ for a raw step. The constants guarantee

$$h\lambda_m < 1, \qquad h(\ell^{-1} + \lambda_c) < 1. \tag{J.22}$$

Fix a dataset in $\mathcal{E}$ and put $\boldsymbol{\sigma} = \operatorname{sign}(\boldsymbol{m}_S)$. Let $M_j(t)$ be the number of selections of memory coordinate j among the first t updates.

Lemma J.6 (Actual path on the product box)**.** *For every coordinate schedule and all* $t \geq 0$*, the actual algorithm* (J.3) *on this dataset satisfies*

$$\sigma_j x_{t,j} \geq \frac{3\alpha}{16n\lambda_m}\{1 - (1 - h\lambda_m)^{M_j(t)}\}, \tag{J.23}$$

$$y_{t,j} = (\boldsymbol{c}_{\boldsymbol{\sigma}})_j u_{t,j}, \qquad 0 \leq u_{t,j} \leq \zeta\theta. \tag{J.24}$$

Every selected memory coordinate has sign σ_j*, while an unselected coordinate is zero. All training readout values and gradients remain zero. Code projections are inactive. Memory projections may be active, and the lower bound accounts for them.*

Proof. We prove all assertions together by induction, starting from zero. The entire domain $\mathcal{K}_n$ is contained in the unit ball, and its two blocks have norm at most $1/2$. Thus χ is inactive and both J_0 terms are quadratic at every iterate. Suppose the bounds hold at the current iterate. A seen label i has $\sigma_i = -1$ by Lemma J.3. Every code admissible for that training readout is therefore different from $\boldsymbol{c}_{\boldsymbol{\sigma}}$. By (J.13) its inner product is at most $3\zeta\theta/4 < \kappa\theta$, whereas $\|\boldsymbol{y}_t\| \leq \zeta\theta < \theta$. Hence $h_{i,\theta}(\boldsymbol{y}_t) = 0$ and $\nabla R_{i,\theta}(\boldsymbol{y}_t) = 0$.

For a memory update, the bound on the mask gradient and (J.8) show that the signed forcing is at least $|(\boldsymbol{m}_S)_j| - a \geq 3\alpha/(16n)$. The preprojection signed coordinate is thus at least

$$(1 - h\lambda_m)\sigma_j x_{t,j} + h\frac{3\alpha}{16n} > 0.$$

The comparison recursion has limiting value $3/(16A\sqrt{n}) < 1/(2\sqrt{k})$, strictly inside the interval. Clipping the actual signed coordinate to this interval preserves its lower bound by the comparison recursion, because clipping is monotone and the comparison value is already feasible. This proves (J.23) and the sign assertion for the next iterate. The other memory coordinates are unchanged.

For a code update, $\boldsymbol{\sigma}$ is compatible with every nonzero memory coordinate, even if other coordinates have not yet been selected. Lemma J.4 and (J.22) therefore preserve (J.24), with $\Gamma = b/\lambda_c = \zeta\theta$. The resulting coordinate has magnitude at most $\zeta\theta/\sqrt{r} < 1/(2\sqrt{r})$, so its projection changes nothing. The induction closes without deleting any term of the actual empirical objective in advance. □

Lemma J.7 (Permanent identification after enough memory visits)**.** *Once every memory coordinate has been selected at least $\lceil 1/(h\lambda_m)\rceil$ times, and at every subsequent iterate,*

$$\nabla P_\ell(\boldsymbol{x}_t, \boldsymbol{y}_t) = (a\boldsymbol{\sigma}, -b\boldsymbol{c}_{\boldsymbol{\sigma}}). \tag{J.25}$$

Every subsequent selected code coordinate therefore follows exactly

$$u_j^+ = (1 - h\lambda_c)u_j + hb. \tag{J.26}$$

Its coefficient never decreases.

Proof. By (J.23) and $(1 - h\lambda_m)^M \leq e^{-1}$,

$$\min_j |x_{t,j}| - \ell a \geq \frac{3(1 - e^{-1})}{16A\sqrt{n}} - \frac{\alpha}{1600An} \geq \frac{1}{12A\sqrt{n}}.$$

This persists because the comparison lower bound is nondecreasing in the visit count. Hence the left side of the exact branch certificate in Lemma E.3 is at least $\alpha/(192An^{3/2})$. Meanwhile

$$\begin{aligned} 2b(\|\boldsymbol{y}_t\| + \ell b) &\leq 2\zeta^2\theta^2\lambda_c(1 + \ell\lambda_c) \\ &\leq 3C_0\theta_0^2\alpha/n^{3/2} < \alpha/(192An^{3/2}). \end{aligned}$$

Here $\theta^2/Q_0 \leq \theta_0^2\alpha/n^{3/2}$ and the strict slack in (J.16) were used. The exact certificate therefore applies forever. The real code update is (J.26), by Lemma J.6. Since $u_j \leq b/\lambda_c$, it is nondecreasing. □

J.7 Random visits, every safe step, and the full average

Define the deterministic integers, depending on the chosen step,

$$m_0 = \left\lceil \frac{1}{h\lambda_m} \right\rceil, \qquad r_0 = \left\lceil \frac{L_0}{h\lambda_c} \right\rceil, \qquad J_1 = 4Dm_0, \quad J_2 = 4Dr_0, \quad J = J_1 + J_2. \tag{J.27}$$

Let $\mathcal{A}$ be the event that each memory coordinate is visited at least m_0 times in the first J_1 updates, and each code coordinate is visited at least r_0 times in the next J_2 updates. Each marginal count is binomial with mean $4m_0$ or $4r_0$. Chernoff and a union bound, without independence between coordinates, give

$$\Pr(\mathcal{A}) \geq 1 - ke^{-m_0/4} - re^{-r_0/4} \geq 1/2. \tag{J.28}$$

For the last inequality, $m_0 \geq 128\sqrt{n}/\alpha$, $\alpha \leq 9/8$, and $\lambda_m/\lambda_c \geq A$ imply $r_0 \geq m_0$. Thus the subtracted quantity is at most $66n\exp(-256\sqrt{n}/9)$, which is less than $1/2$ for $n \geq 64$ and decreases thereafter. The event $\mathcal{A}$ is independent of $\mathcal{E}$.

On $\mathcal{E} \cap \mathcal{A}$, identification holds from time J_1 onward. The exact recursion and the second set of visits imply for every code coordinate

$$u_{J,j} \geq \zeta\theta(1 - e^{-h\lambda_c r_0}) \geq (\zeta - \Delta/4)\theta, \qquad u_{t,j} \geq u_{J,j} \quad (t \geq J). \tag{J.29}$$

Set $q_0 = K_0 Q_0$. If $q = \eta T/D \geq q_0$, then $T \geq DQ_0/h$, and the integer roundings in (J.27) cost only

$$\frac{J}{T} \leq \frac{4}{\lambda_m Q_0} + \frac{4L_0}{C_0} + \frac{8h}{Q_0} \leq \frac{4}{AC_0} + \frac{4L_0}{C_0} + \frac{8}{K_0 C_0} \leq \frac{\Delta}{8\zeta}. \tag{J.30}$$

All preheating code coefficients are nonnegative. Therefore both the last iterate and the full updated average satisfy

$$\langle \boldsymbol{c}_{\boldsymbol{\sigma}}, \widehat{\boldsymbol{w}}_y \rangle \geq (1 - J/T)(\zeta - \Delta/4)\theta \geq (\kappa + \Delta/2)\theta, \qquad \|\widehat{\boldsymbol{w}}_y\| \leq \zeta\theta. \tag{J.31}$$

For the last iterate, the factor $1 - J/T$ is simply an unnecessary weakening. For the average, nonnegativity of every preheating term allows us to retain the last $T - J$ terms of the sum while keeping the original denominator T. Thus the factor $1 - J/T$ pays for the entire prefix. The output is still the complete post-update average.

At least half of the $2n$ labels are unseen. For each such label i, the code $\boldsymbol{c}_{\boldsymbol{\sigma}}$ is admissible, and (J.15) and (J.31) give $R_{i,\theta}(\widehat{\boldsymbol{w}}_y) \geq \Delta^2\theta^2/8$. Using the pointwise population inequality, independence of the two events, and nonnegative excess risk off the event, we obtain

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] \geq \frac{\Delta^2\theta^2}{512K_0} = \frac{\Delta^2\theta_0^2}{512K_0} \min\{1, Q_0\alpha/n^{3/2}\}. \tag{J.32}$$

This is a lower bound on actual population risk, not on a stability certificate or a parameter distance.

To obtain any requested $\nu \in [0, 1]$, choose $\epsilon = \min\{1, K_0\nu\}$. Then (J.20) puts the instance in $\mathcal{P}_\nu(\mathcal{K}_n)$, and

$$s_\nu \leq \alpha/\sqrt{n} \leq K_0 s_\nu. \tag{J.33}$$

Every $q_0 \geq K_0 C_0/s_\nu$ is consequently an admissible prescribed horizon. Substituting $Q_0 = q_0/K_0$ into (J.32) proves (J.5), with universal constants. The kernel uses q_0 but not η or T. For each fixed safe step, the same finite event $\mathcal{A}$ works for every later integer horizon simultaneously. No common event across all real step sizes is required for the stated expectation quantifiers.

Proposition J.8 (The same lower trajectory without projection)**.** *On $\mathcal{E}$, running unprojected uniform scalar-coordinate gradient descent on (J.21), with the same step range, preserves (J.23), (J.24), and all subsequent identification and readout conclusions. Its entire trajectory satisfies*

$$\|\boldsymbol{x}_t\| \leq \frac{88 + \sqrt{2}/16}{A}, \qquad \|\boldsymbol{y}_t\| \leq \zeta\theta, \qquad \|(\boldsymbol{x}_t, \boldsymbol{y}_t)\| < 1. \tag{J.34}$$

Proof. In addition to the lower comparison, the unprojected memory recursion gives, coordinatewise, $|x_{t,j}| \leq (|(\boldsymbol{m}_S)_j| + a)/\lambda_m$. Summing its squared coordinate bounds yields

$$\|\boldsymbol{x}_t\| \leq (\|\boldsymbol{m}_S\| + a\sqrt{k})/\lambda_m \leq (88 + \sqrt{2}/16)/A.$$

The cone argument still gives the code bound. Each next candidate obeys these same inequalities and is strictly inside the unit ball, where the Huber terms are quadratic and χ is inactive. The induction in Lemma J.6 thus closes without projection. The memory lower bound and the subsequent proofs do not otherwise use clipping. This proposition is a good-data-event statement. A global sublevel-set bound needed for a particular external comparison must be checked separately. □

J.8 Optimization and statistical terms on the same box

It remains to cover the other terms in (J.4). A one-coordinate witness on an interval of fixed radius cannot simply be inserted into (J.2), whose scalar intervals shrink with dimension. Instead take the dense unit direction $\boldsymbol{u} = k^{-1/2}\mathbf{1}$ in the memory block and let $w = \langle \boldsymbol{u}, \boldsymbol{x} \rangle$. The range of w on $\mathcal{K}_n$ is exactly $[-1/2, 1/2]$. Composing a scalar loss with this unit linear functional preserves its Lipschitz and smoothness bounds.

Optimization. Use the deterministic, hence symmetric, pair kernel

$$f(\boldsymbol{x}, \boldsymbol{y}; z, z') = -a_q\langle \boldsymbol{u}, \boldsymbol{x} \rangle + \tfrac{1}{4}\chi(\boldsymbol{x}, \boldsymbol{y}) + 1, \qquad a_q = \min\{1/8, 1/(16q)\}.$$

It is globally convex and nonnegative, 1-Lipschitz and 1-smooth. For example, its only possibly negative term is bounded below by $-a_q\|(\boldsymbol{x}, \boldsymbol{y})\|$, whereas the radial term eventually has slope

$1/4 > a_q$. Its minimum is positive by the displayed formula for χ. All random gradient projections are zero. On $\mathcal{K}_n$ the objective is just the affine term and a constant. Each selected memory coordinate moves in the positive $\boldsymbol{u}$ direction by at most $\eta a_q/\sqrt{k}$. Conditional on the past, its contribution to the expected increase in w is at most $\eta a_q/D$. Thus

$$\mathbb{E}w_T \le a_q q \le 1/16, \qquad \mathbb{E}\langle \boldsymbol{u}, \bar{\boldsymbol{x}}_T\rangle \le 1/16.$$

The optimum has $w = 1/2$. Either output has excess at least $a_q(1/2 - 1/16) \ge c\min\{1, q^{-1}\}$.

First-order statistical error. Let $Z \in \{-1, 1\}$ have mean τm, where $\tau \in \{-1, 1\}$ is an unknown parameter and $m = 1/(4\sqrt{n})$. Use the same kernel under both data laws,

$$f(\boldsymbol{x}, \boldsymbol{y}; z, z') = -\frac{\nu}{16}(z + z')\langle \boldsymbol{u}, \boldsymbol{x}\rangle + \tfrac{1}{4}\chi(\boldsymbol{x}, \boldsymbol{y}) + 1.$$

The absolute affine slope is at most $1/8$, so the same global checks apply. At any population optimum in the box, the first gradient projection has norm at most $\nu/16 \le \nu$. The population objective in w has slope $-\tau\nu m/8$, and its optimum is $w = \tau/2$.

For the two n-sample laws, the KL divergence is

$$nm \log \frac{1+m}{1-m} \le \frac{2nm^2}{1-m} \le 1/4.$$

Pinsker consequently bounds their total variation by $1/2$. For any possibly randomized output $w \in [-1/2, 1/2]$, adjoining the same independent algorithm seed cannot increase this distance, and $|\mathbb{E}_+ w - \mathbb{E}_- w| \le \mathrm{TV}(P_+^n, P_-^n)$. The average of the two expected excess risks is therefore at least

$$\frac{\nu m}{16}(1 - \mathrm{TV}(P_+^n, P_-^n)) \ge c\nu/\sqrt{n}.$$

This bound holds for every learner, hence for both specified RCD outputs.

Canonical statistical error. Let $\vartheta \in \{-1/4, 1/4\}$ be the unknown parameter, let $B \sim \mathrm{Bernoulli}(1/n)$ and $Y = B\vartheta$, and let ξ be an independent Rademacher sign. The observation is (B, Y, ξ). Define $\psi(w) = H_1((|w| - 1/2)_+)$. For two observations use the fixed scalar kernel

$$\begin{aligned} f(w; (B, Y, \xi), (B', Y', \xi')) = {} & \frac{B\mathsf{H}_2(w - Y) + B'\mathsf{H}_2(w - Y')}{16} \\ & + \frac{1}{4}w\xi\xi' + \frac{1}{2}\psi(w) + \frac{3}{16}. \end{aligned} \tag{J.35}$$

Compose it with $w = \langle \boldsymbol{u}, \boldsymbol{x}\rangle$ on the same $\mathcal{K}_n$. The absolute derivative is at most $1/4 + 1/4 + 1/2 = 1$, and its derivative has Lipschitz constant at most $1/8 + 1/2 < 1$. Convexity follows term by term. Global nonnegativity follows by dropping the first two nonnegative Huber terms and putting $r_* = |w|$. The remaining lower bound is

$$-r_*/4 + \tfrac{1}{2}H_1((r_* - 1/2)_+) + 3/16.$$

It is positive for $r_* \le 1/2$, equals $(r_* - 1)^2/4$ for $1/2 \le r_* \le 3/2$, and equals $r_*/4 - 5/16$ for $r_* \ge 3/2$.

On $[-1/2, 1/2]$, the population excess is exactly $(w - \vartheta)^2/(16n)$, minimized at $w^* = \vartheta$. At this optimum the gradient is $\xi\xi'\boldsymbol{u}/4$: every informative Huber derivative and the derivative of ψ vanish. Its first projection is therefore identically zero. With probability $(1 - 1/n)^n \ge 1/4$, all B values in the original sample vanish. Conditional on this event, the data and all algorithmic outputs have the same distribution under the two choices of ϑ. For every scalar output,

$$\tfrac{1}{2}\{(w - 1/4)^2 + (w + 1/4)^2\} = w^2 + 1/16 \ge 1/16.$$

Averaging over the two parameters and then the event proves a worst expected excess of at least $1/(1024n)$, for every learner. The points with $w = \pm 1/4$ are feasible in $\mathcal{K}_n$, since $\boldsymbol{x} = \pm\boldsymbol{u}/4$ is feasible.

Completion of Theorem J.2. The three preceding witnesses give universal lower bounds for $\min\{1, q^{-1}\}$, $\nu/\sqrt{n}$, and $1/n$ on exactly the same domain and loss class. If $q \geq K_0 C_0/s_\nu$, choose the persistent instance with $q_0 = q$. Its lower bound supplies the capped late term $\min\{1, qs_\nu/n\}$. The maximum of these four lower bounds dominates a universal constant times the capped sum in (J.4). If $q < K_0 C_0/s_\nu$, the omitted late term is at most $K_0 C_0/n$ and is absorbed, with a universal constant, by the canonical statistical witness. This completes the lower bound for all horizons and all $0 \leq \nu \leq 1$. The earlier construction already proved persistence for both outputs on one fixed instance. □

J.9 Full-gradient PGD on the Euclidean unit ball

The scalar-coordinate result does not by itself prove a statement about full-gradient PGD or a different projection geometry. We now verify this additional algorithm and domain directly. Write $\mathcal{P}_\nu(\mathbb{B}^D)$ for the same global loss and first-projection class as above, with the deterministic optimum taken over $\mathbb{B}^D$.

Corollary J.9 (The full-gradient lower law on a unit ball)**.** *Let $n \geq 64$, $D = 66n$, and $0 \leq \nu \leq 1$. Initialize at zero and run the actual full-gradient update*

$$\boldsymbol{X}_{t+1} = \Pi_{\mathbb{B}^D}(\boldsymbol{X}_t - \eta \nabla G_S(\boldsymbol{X}_t)), \qquad 0 < \eta \leq 1, \qquad q = \eta T, \tag{J.36}$$

with the original complete-pair empirical objective. For both $\widehat{\boldsymbol{w}} = \boldsymbol{X}_T$ and $\widehat{\boldsymbol{w}} = T^{-1}\sum_{t=1}^{T} \boldsymbol{X}_t$, universal constants satisfy

$$\sup_{(P,f)\in\mathcal{P}_\nu(\mathbb{B}^D)} \mathbb{E}[F(\widehat{\boldsymbol{w}}) - \min_{\mathbb{B}^D} F] \geq c \min\left\{1, q^{-1} + s_\nu + \frac{qs_\nu}{n}\right\}. \tag{J.37}$$

For each prescribed $q_0 \geq C/s_\nu$, there is one kernel and data law, fixed before sampling and independently of the step size, whose expected excess for both outputs is at least $c\min\{1, q_0 s_\nu/n\}$ for every $0 < \eta \leq 1$ and every integer T with $\eta T \geq q_0$.

Proof. The actual full-gradient trajectory. Use exactly the global kernel (J.18), code family (J.9), and parameters (J.16)–(J.17). These definitions do not depend on the projection domain. Its global population minimum is zero's value, by Lemma J.5, so the same anchor and first-projection computations apply on $\mathbb{B}^D$. Fix data in $\mathcal{E}$, put $\boldsymbol{\sigma} = \operatorname{sign}(\boldsymbol{m}_S)$, and let $h = \eta/K_0$. We establish, for every full update,

$$\sigma_j x_{t,j} \geq \frac{3\alpha}{16n\lambda_m}\{1 - (1 - h\lambda_m)^t\}, \tag{J.38}$$

$$\|\boldsymbol{x}_t\| \leq B_x := \frac{88 + \sqrt{2}/16}{A}, \qquad y_{t,j} = (\boldsymbol{c}_{\boldsymbol{\sigma}})_j u_{t,j}, \quad 0 \leq u_{t,j} \leq \zeta\theta. \tag{J.39}$$

In particular every unprojected next candidate has squared norm at most $B_x^2 + \zeta^2\theta_0^2 < 1$, so the ball projection in (J.36) is inactive on this event.

Here is an induction proving these statements together, including the assertion about each next candidate. They hold at zero. At an iterate satisfying them, the wrong-code bound (J.13) makes every training readout and its gradient zero, just as in Lemma J.6. This includes the initial point. The iterate lies inside the unit ball, so χ is inactive and both radial Huber terms are quadratic. With $\boldsymbol{p}_t = \nabla_{\boldsymbol{x}} P_\ell(\boldsymbol{x}_t, \boldsymbol{y}_t)$, the next memory candidate is

$$\widetilde{\boldsymbol{x}}_{t+1} = (1 - h\lambda_m)\boldsymbol{x}_t + h(\boldsymbol{m}_S - \boldsymbol{p}_t), \qquad |p_{t,j}| \leq a.$$

The coordinatewise signed forcing is at least $3\alpha/(16n)$, which proves (J.38) for the candidate. Its norm obeys the simultaneous vector bound

$$\begin{aligned}\|\widetilde{\boldsymbol{x}}_{t+1}\| &\leq (1 - h\lambda_m)\|\boldsymbol{x}_t\| + h(\|\boldsymbol{m}_S\| + a\sqrt{k}) \\ &\leq (1 - h\lambda_m)B_x + h\lambda_m B_x = B_x.\end{aligned}$$

There is no coordinate projection in this calculation. The code-gradient expression (J.12) holds with the current compatible sign vector, even when input memory coordinates are zero. The simultaneous code candidates therefore satisfy, for every j,

$$\widetilde{u}_{t+1,j} = (1 - h(\lambda_c + A_y))u_{t,j} + hB_y \in [0, \zeta\theta],$$

by (J.22). The same coefficients A_y, B_y are evaluated at the current full iterate. This is a synchronous full-gradient step, not a sequence of scalar steps. These memory and code bounds put the entire next candidate strictly inside the ball, so its projection does not change it. This closes the induction and also justifies using the true empirical objective throughout.

Deterministic preheating and the full average. Let m_0, r_0 have their definitions in (J.27), and now put $J = m_0 + r_0$. After the first m_0 full steps, (J.38) gives the identical strict identification certificate proved in Lemma J.7. It remains valid at every later full iterate. All code coordinates consequently follow (J.26) at every subsequent step. After r_0 such steps,

$$u_{J,j} \geq (\zeta - \Delta/4)\theta, \qquad u_{t,j} \geq u_{J,j} \quad (t \geq J).$$

If $q = \eta T \geq q_0 := K_0 Q_0$, then $T \geq Q_0/h$. The complete rounding calculation is

$$\frac{J}{T} \leq \frac{1}{\lambda_m Q_0} + \frac{L_0}{C_0} + \frac{2h}{Q_0} \leq \frac{1}{AC_0} + \frac{L_0}{C_0} + \frac{2}{K_0 C_0} \leq \frac{\Delta}{8\zeta}. \tag{J.40}$$

Thus every earlier code coefficient is nonnegative, and both the last point and the average of all T updated iterates satisfy (J.31). At least half the labels are unseen, so the same halfspace-envelope bound gives

$$F(\widehat{\boldsymbol{w}}) - F^* \geq \frac{\Delta^2\theta^2}{16K_0} \quad \text{on } \mathcal{E}.$$

Since $\Pr(\mathcal{E}) \geq 1/16$, the expected lower bound is at least $\Delta^2\theta^2/(256K_0)$. No coordinate-count event or random-clock approximation is involved. Choosing $\epsilon = \min\{1, K_0\nu\}$ and using (J.33) proves the persistent part of the corollary with the stated quantifiers. For a fixed step, the same data event proves the conclusion for every later horizon.

The remaining terms on the ball. The early witnesses must also be checked on the new domain. Keep the unit vector $\boldsymbol{u}$ in the memory block and set $w = \langle \boldsymbol{u}, \boldsymbol{x} \rangle$. Its range on $\mathbb{B}^D$ is now $[-1, 1]$. All global regularity proofs for the three preceding kernels are independent of the feasible domain. For the deterministic optimization kernel,

$$f(\boldsymbol{x}, \boldsymbol{y}) = -a_q \langle \boldsymbol{u}, \boldsymbol{x} \rangle + \chi(\boldsymbol{x}, \boldsymbol{y})/4 + 1, \qquad a_q = \min\{1/8, 1/(16q)\},$$

the full-gradient iterates are exactly $\boldsymbol{x}_t = \eta t a_q \boldsymbol{u}$, $\boldsymbol{y}_t = 0$ for $0 \leq t \leq T$. Indeed their norms are at most $a_q q \leq 1/16$, so both the barrier and the ball projection are inactive at every step. The optimum over the ball has $w = 1$, while each output has $w \leq 1/16$. The resulting risk is at least $(15/16)a_q \geq c \min\{1, q^{-1}\}$, and its first projection vanishes.

For the first-order statistical kernel with $\mathbb{E}Z = \tau m$, $m = 1/(4\sqrt{n})$, the population optimum on the ball is now $w = \tau$, and the excess is $(\nu m/8)(1 - \tau w)$. Its first projection remains at most $\nu/16$, including at these boundary optima, where $\nabla\chi = 0$. The same two original-sample laws have total variation at most $1/2$. Since an output lies in $[-1, 1]$, $|\mathbb{E}_+ w - \mathbb{E}_- w| \leq 2\,\mathrm{TV}(P_+^n, P_-^n)$, and averaging the two excess risks gives at least

$$\frac{\nu m}{16}(2 - 2\,\mathrm{TV}(P_+^n, P_-^n)) \geq c\nu/\sqrt{n}.$$

This statistical argument covers either algorithmic output directly.

Finally, for the canonical statistical kernel (J.35), the larger scalar interval stays within the quadratic region of every informative Huber term: $|w - \vartheta| \leq 5/4 < 2$. On the entire ball the population excess is therefore exactly

$$F(\boldsymbol{x}, \boldsymbol{y}) - F^* = \frac{(w - \vartheta)^2}{16n} + \frac{1}{2}\psi(w) \geq \frac{(w - \vartheta)^2}{16n}.$$

It is minimized at the feasible point $\boldsymbol{x} = \vartheta \boldsymbol{u}$, $\boldsymbol{y} = 0$. The first projection there is still zero. The same event that all original informative indicators vanish has probability at least $1/4$. Conditional indistinguishability and the squared-distance identity used above apply to any real w, hence also to the larger range $[-1, 1]$. They give the same $1/(1024n)$ lower bound for both outputs. The possible additional nonnegative ψ term cannot weaken it.

For $q \geq K_0 C_0/s_\nu$, set $q_0 = q$ in the late witness and take the maximum with these three early witnesses. For smaller q, $qs_\nu/n \leq K_0 C_0/n$ is absorbed by the canonical statistical witness. This proves (J.37) over all horizons, and completes the unit-ball full-gradient argument. □

K FIXED-SOURCE RISK LAWS

This appendix gives a matching worst-loss law for a countable source held fixed across sample sizes and horizons. To verify the upper bound, keep the actual conditional mass $\tau = N/n$ in (K.8), apply only the residual reader in (K.13), and use (K.15) before (K.16). The anchor $\boldsymbol{w}^*$ minimizes the original population objective. It need not minimize a conditional objective. Optimization is used only after reassembly. The algorithms use the full sample throughout.

K.1 STATEMENT AND QUANTIFIERS

Let $p = (p_z)_{z\in\mathcal{Z}}$ be a probability distribution on a countable set. Let $f_z : \mathbb{R}^d \to \mathbb{R}$ be finite, convex, 1-smooth and 1-Lipschitz. Write $B = \{\boldsymbol{w} : \|\boldsymbol{w}\| \le 1\}$, $G_S = n^{-1}\sum_{i=1}^n f_{Z_i}$, and $F(\boldsymbol{w}) = \sum_z p_z[f_z(\boldsymbol{w}) - f_z(0)]$ for $Z_i \overset{\text{i.i.d.}}{\sim} p$. This normalization makes population differences well defined even when the original intercepts are not integrable. It changes neither gradients nor excess risks. The series defining F converges uniformly on B. In particular, F is continuous and has a deterministic minimizer $\boldsymbol{w}^* \in B$.

The first rule is PGD, initialized at $\boldsymbol{X}_0 = 0$, with $0 < \eta \le 1$ and $q = \eta T$:

$$\boldsymbol{X}_{j+1} = \Pi_B(\boldsymbol{X}_j - \eta\nabla G_S(\boldsymbol{X}_j)). \tag{K.1}$$

The second rule is ridge minimization with $\lambda > 0$:

$$\boldsymbol{X}_\lambda = \operatorname*{arg\,min}_{\boldsymbol{w}\in B}\{G_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2/2\}. \tag{K.2}$$

For PGD, $\widehat{\boldsymbol{w}}$ is either $\boldsymbol{X}_T$ or $T^{-1}\sum_{j=1}^T \boldsymbol{X}_j$. For ridge, $\widehat{\boldsymbol{w}} = \boldsymbol{X}_\lambda$. We use $t = q$ in the former case and $t = \lambda^{-1}$ in the latter, after proving their respective update and optimality bounds. All losses, parameters and the source are fixed before the sample is drawn.

For $n \ge 4$, set $a = \lfloor n/2\rfloor$, $b = n - a$, and define

$$\beta_a(p) = \sum_z p_z \Pr\{\mathrm{Bin}(a, p_z) < ap_z/2\}. \tag{K.3}$$

This is a deterministic occupancy quantity, rather than a property of an unknown output map.

Theorem K.1 (A source-dependent bound for the original output)**.** *For every countable source, every dimension, and each rule and output above, a universal constant C satisfies*

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F(\boldsymbol{w}^*)] \le C\min\left\{1,\ t^{-1} + n^{-1/2} + \min\left\{\beta_a(p), \frac{t\sqrt{\beta_a(p)}}{n^{3/2}}\right\}\right\}. \tag{K.4}$$

The algorithm uses all n observations and requires no knowledge of p.

Define $\mathcal{N}_p(x) = \big|\{z : p_z \ge x\}\big|$ and $K_n = \mathcal{N}_p(1/n)$. Assume that for a slowly varying function ℓ and $0 < \alpha < 1$,

$$\mathcal{N}_p(1/u) \sim u^\alpha \ell(u) \qquad (u \to \infty). \tag{K.5}$$

Here slow variation means $\ell(cu)/\ell(u) \to 1$ for every fixed $c > 0$. Let $\mathcal{R}^{\mathsf{A}}_{p,d}(n,t)$ denote the supremum of the excess risk over the preceding losses on $\mathbb{R}^d$, for one of the specified rules and outputs A. For PGD its remaining parameters (η, T) are held fixed and satisfy $\eta T = t$.

Theorem K.2 (Sharp law at a fixed regularly varying source)**.** *For every fixed source satisfying (K.5), there are constants $c_p, C_p, n_0(p) > 0$ such that, for all $n \ge n_0(p)$ and all admissible algorithm parameters, the upper bound*

$$\mathcal{R}^{\mathsf{A}}_{p,d}(n,t) \le C_p\Phi_p(n,t), \qquad \Phi_p(n,t) = \min\left\{1, t^{-1} + n^{-1/2} + \min\left\{\frac{K_n}{n}, \frac{t\sqrt{K_n}}{n^2}\right\}\right\} \tag{K.6}$$

holds for every d. If $d \ge C_pK_n$, then also $\mathcal{R}^{\mathsf{A}}_{p,d}(n,t) \ge c_p\Phi_p(n,t)$. The lower bound can be realized by globally nonnegative losses.

Moreover, for every fixed $n \geq n_0(p)$ and $t_0 \geq C_p n/\sqrt{K_n}$, there is one loss family, fixed before sampling, for which simultaneously

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] \geq c_p \min\left\{\frac{K_n}{n}, \frac{t_0\sqrt{K_n}}{n^2}\right\} \tag{K.7}$$

for all PGD parameters $0 < \eta \leq 1$, $\eta T \geq t_0$, for both PGD outputs, and for every ridge parameter $0 < \lambda \leq 1/t_0$.

The constants may depend on the fixed source. In the envelope (K.6), losses may depend on the predetermined n, t. In (K.7), they depend on n, t_0 and remain fixed thereafter. Neither assertion says that one loss family attains the entire envelope for every sample size and time. The growing dimension in the lower bound is also explicit.

K.2 The conditional response keeps the coefficient $1/n$

We first record the input supplied by interaction–dissipation. Fix a convex background C_0 and $N \geq 1$ remaining i.i.d. labels $W_1, \ldots, W_N \sim r$. The actual empirical objective is

$$C_0(\boldsymbol{w}) + \frac{1}{n}\sum_{j=1}^{N} f_{W_j}(\boldsymbol{w}). \tag{K.8}$$

For PGD suppose that C_0 is L_C-smooth and $L_C + N/n \leq 1$. For ridge it is enough that C_0 be differentiable and convex. The feasible set and initialization are unchanged under all replacements. Let $\widehat{\boldsymbol{w}}^{(j)}$ be the original output rule with only W_j independently replaced, and put $v_N = N^{-1}\sum_j \mathbb{E}\|\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{w}}^{(j)}\|^2$. The unanchored interaction–dissipation estimate gives

$$v_N \leq q\left\{\frac{4}{n\sqrt{N}} + \frac{8}{nN} + \frac{12\eta}{n^2}\right\} \leq \frac{24q}{n\sqrt{N}}, \qquad \text{PGD}, \tag{K.9}$$

$$v_N \leq \frac{1}{\lambda n}\left\{\frac{2}{\sqrt{N}} + \frac{1}{N}\right\} \leq \frac{3}{\lambda n\sqrt{N}}, \qquad \text{ridge}. \tag{K.10}$$

For clarity, the substitution into the general estimate is as follows. The factors are singletons, their normalized weights are $1/N$, and the objective multiplier is $\tau = N/n$. Thus

$$H_0 = N^{-3}, \qquad H_1 \leq 2N^{-2}, \qquad \sum_i p_i^3 = N^{-2}.$$

All interaction quantities use the product law r^N with C_0 held fixed. The unanchored bound is $\mathcal{C}(\boldsymbol{X}) - c\mathcal{D}(\boldsymbol{X}) \leq \sqrt{H_0} + H_1/(4c)$ on the unit ball. Summing the actual PGD supply $\mathcal{I}(\boldsymbol{X}_{j+1}) - \mathcal{I}(\boldsymbol{X}_j) \leq 4\eta\tau(\mathcal{C}(\boldsymbol{X}_j) - \mathcal{D}(\boldsymbol{X}_j)/4) + 12\eta^2\tau^2\sum_i p_i^3$ gives (K.9). Every prefix obeys the same bound. Jensen therefore also gives it for the full average. For the actual ridge solution, the two KKT conditions instead give $\lambda\mathcal{I}(\boldsymbol{X}_\lambda) \leq 2\tau(\mathcal{C}(\boldsymbol{X}_\lambda) - \mathcal{D}(\boldsymbol{X}_\lambda))$, yielding (K.10). These are the unanchored ID and algorithm-supply results of Corollary D.3 and Lemma D.4, specialized to the stated finite product space. No conditional population minimizer is used. In particular, the factor N/n has not been discarded.

K.3 From stability to population risk

Fix the original deterministic optimizer $\boldsymbol{w}^*$, and define

$$\boldsymbol{a}_z = \nabla f_z(\boldsymbol{w}^*), \quad \boldsymbol{\mu} = \mathbb{E}\boldsymbol{a}_Z, \quad \sigma^2 = \mathbb{E}\|\boldsymbol{a}_Z - \boldsymbol{\mu}\|^2 \leq 1,$$
$$r_z(\boldsymbol{w}) = f_z(\boldsymbol{w}) - f_z(\boldsymbol{w}^*) - \langle \boldsymbol{a}_z, \boldsymbol{w} - \boldsymbol{w}^*\rangle.$$

Convexity and smoothness imply, globally, $r_z \geq 0$, $\|\nabla r_z\|^2 \leq 2r_z$, and on B,

$$0 \leq r_z(\boldsymbol{w}) \leq \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}^*\|^2 \leq 2, \qquad r_z(\boldsymbol{y}) \leq 2r_z(\boldsymbol{x}) + \|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{K.11}$$

The second inequality follows from the smooth upper bound followed by $\sqrt{2r_z(\boldsymbol{x})}\|\boldsymbol{y} - \boldsymbol{x}\| \leq r_z(\boldsymbol{x}) + \|\boldsymbol{y} - \boldsymbol{x}\|^2/2$. The first-order condition at a boundary optimizer is $\langle \boldsymbol{\mu}, \boldsymbol{w} - \boldsymbol{w}^*\rangle \geq 0$. It does not require $\boldsymbol{\mu} = 0$.

Split the sample indices in advance into a pilot $P = (Z_1, \dots, Z_a)$ and b further observations. Let N_z^P be a pilot count and define

$$A(P) = \{z : N_z^P < ap_z/2\}, \qquad p_A = \sum_{z \in A(P)} p_z, \qquad Q = \frac{2n}{\min(a,b)} \le 6.$$

For every $\boldsymbol{w} \in B$, including a data-dependent output,

$$\sum_{z \notin A(P)} p_z r_z(\boldsymbol{w}) \le \frac{2}{a} \sum_{i \le a} r_{Z_i}(\boldsymbol{w}). \tag{K.12}$$

This is a pointwise inequality and needs no independence from P.

Condition first on P, then on the membership indicators $\mathbf{1}_{\{Z_i \in A(P)\}}$ for $i > a$, and then on the values of all outside observations. Denote this sigma-field by $\mathcal{H}$ and let $N = \sum_{i>a} \mathbf{1}_{\{Z_i \in A(P)\}}$. If $N \ge 1$, the remaining labels are independent with law $r = p(\cdot \mid A(P))$ conditional on $\mathcal{H}$. The fixed background in (K.8) is exactly

$$C_{\mathcal{H}}(\boldsymbol{w}) = \frac{1}{n}\left\{ \sum_{i \le a} f_{Z_i}(\boldsymbol{w}) + \sum_{i > a : Z_i \notin A(P)} f_{Z_i}(\boldsymbol{w}) \right\}.$$

It includes all pilot losses, even pilot labels in $A(P)$. Its smoothness is at most $(n-N)/n$, so the PGD condition is satisfied. Conditional replacements preserve $\mathcal{H}$ and this background. All conditional statements below hold almost surely. Countability also permits direct conditioning on the positive-probability configurations.

Let $r_A(\boldsymbol{w}) = \sum_{z \in A} p_z r_z(\boldsymbol{w})/p_A$ when $p_A > 0$, and put it equal to zero otherwise. The replacement identity and (K.11) give, conditional on $\mathcal{H}$,

$$\mathbb{E}[r_A(\widehat{\boldsymbol{w}}) \mid \mathcal{H}] = \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}[r_{W_j}(\widehat{\boldsymbol{w}}^{(j)}) \mid \mathcal{H}] \le \frac{2}{N}\sum_{j=1}^{N} \mathbb{E}[r_{W_j}(\widehat{\boldsymbol{w}}) \mid \mathcal{H}] + v_N. \tag{K.13}$$

Here v_N is the replacement influence evaluated conditionally on $\mathcal{H}$. The original anchor $\boldsymbol{w}^*$ need not minimize the conditional objective: only its globally nonnegative residuals were used here.

Multiply (K.13) by the actual N/b before averaging. If $N = 0$ this weighted expression is defined to be zero, without defining v_0 or dividing by zero. The probability correction is necessary because the output can depend on N. Since $0 \le r_A \le 2$ and $N \mid P \sim \mathrm{Bin}(b, p_A)$,

$$\mathbb{E}[p_A r_A(\widehat{\boldsymbol{w}}) \mid P] \le \mathbb{E}[(N/b) r_A(\widehat{\boldsymbol{w}}) \mid P] + 2\mathbb{E}[(p_A - N/b)_+ \mid P] \tag{K.14}$$

$$\le \mathbb{E}[(N/b) r_A(\widehat{\boldsymbol{w}}) \mid P] + \sqrt{p_A(1-p_A)/b}. \tag{K.15}$$

Indeed, the zero mean of $p_A - N/b$ gives $2\mathbb{E}[(p_A - N/b)_+ \mid P] = \mathbb{E}[|p_A - N/b| \mid P]$. Cauchy–Schwarz then gives the displayed standard deviation. Only $0 \le r_A(\widehat{\boldsymbol{w}}) \le 2$ is used about the fitted output, with no independence assumption about $\widehat{\boldsymbol{w}}$ and N.

Let $r_S = n^{-1}\sum_i r_{Z_i}$ and $r_p = \sum_z p_z r_z$. Combining (K.12)–(K.15), and using nonnegativity to reassemble the two disjoint index sets, proves

$$\mathbb{E} r_p(\widehat{\boldsymbol{w}}) \le Q\mathbb{E} r_S(\widehat{\boldsymbol{w}}) + \min\left\{ 2\beta_a(p), \mathbb{E}\sqrt{p_A(1-p_A)/b} + \frac{1}{b}\mathbb{E}[N v_N \mathbf{1}_{\{N>0\}}] \right\}. \tag{K.16}$$

The first branch simply uses $p_A r_A \le 2p_A$ in place of the conditional response. Both branches have the same coefficient Q. All countable sums above are justified by nonnegativity and Tonelli.

Suppose the actual output satisfies $\mathbb{E}[G_S(\widehat{\boldsymbol{w}}) - G_S(\boldsymbol{w}^*)] \le \delta$. With $\bar{\boldsymbol{a}} = n^{-1}\sum_i \boldsymbol{a}_{Z_i}$ and $\boldsymbol{h} = \widehat{\boldsymbol{w}} - \boldsymbol{w}^*$, the population first-order condition now applies to the reassembled original risk:

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F(\boldsymbol{w}^*)] \le Q\delta + Q\mathbb{E}\langle \boldsymbol{\mu} - \bar{\boldsymbol{a}}, \boldsymbol{h}\rangle - (Q-1)\mathbb{E}\langle \boldsymbol{\mu}, \boldsymbol{h}\rangle + \text{the error in (K.16)}$$

$$\le Q\delta + 2Q\sigma/\sqrt{n} + \text{the same error}.$$

Here $\mathbb{E}\|\bar{\boldsymbol{a}} - \boldsymbol{\mu}\|^2 = \sigma^2/n$ and $\|\boldsymbol{h}\| \le 2$.

For PGD, the optimization estimates (C.5) and (C.17), applied to the original unit-smooth G_S, give $\delta \le 1/(2q)$ for both outputs. For ridge, comparison with $\boldsymbol{w}^*$ gives $\delta \le \lambda/2$ directly. Inserting (K.9) or (K.10) into (K.16), and using $\mathbb{E}\sqrt{N} \le \sqrt{b\beta_a(p)}$ and $\mathbb{E}\sqrt{p_A(1-p_A)/b} \le b^{-1/2}$, proves Theorem K.1. The outer cap follows from the diameter and the Lipschitz assumption. The split was entirely an analysis device.

K.4 From occupancy to a fixed-source profile

The regular-variation occupancy theory is classical (Ben-Hamou et al., 2017). We supply the elementary order estimates needed here, so no concentration statement about a random occupancy count is implicit in the risk theorem.

Lemma K.3. *Under (K.5), $K_n \to \infty$, $K_n = o(n)$, and, for $a = \lfloor n/2 \rfloor$,*

$$\beta_a(p) = \Theta_p(K_n/n), \qquad \left|\{z : (16n)^{-1} \le p_z < (8n)^{-1}\}\right| = \Theta_p(K_n). \tag{K.17}$$

Proof. Writing $M(u) = \mathcal{N}_p(1/u)$, regular variation gives $M(cn)/M(n) \to c^\alpha$. The rare-band cardinality divided by K_n therefore converges to $16^\alpha - 8^\alpha > 0$. Choose $\epsilon > 0$ with $\alpha + \epsilon < 1$. For all sufficiently large u, $M(2u)/M(u) \le 2^{\alpha+\epsilon}$, and hence $M(2^j n) \le 2^{j(\alpha+\epsilon)} K_n$ for all $j \ge 0$. Likewise an eventual lower bound greater than one on this ratio shows $K_n \to \infty$. The upper ratio and monotonicity between successive powers of two show $K_n = o(n)$.

For every fixed $c > 0$, dyadic bins below $1/n$ give

$$\sum_{p_z \le 1/n} p_z \le \frac{K_n}{n} \sum_{j \ge 0} 2^{-j} 2^{(j+1)(\alpha+\epsilon)} \lesssim_p K_n/n.$$

For the remaining bins,

$$\sum_{p_z > 1/n} p_z e^{-cnp_z} \le \frac{K_n}{n} \sum_{j \ge 0} 2^{j+1} e^{-c2^j} \lesssim_c K_n/n.$$

The binomial Chernoff bound $\Pr\{\mathrm{Bin}(a,u) < au/2\} \le e^{-au/8}$ and $a \ge n/3$ now give $\beta_a \lesssim_p K_n/n$. For every label in the displayed rare band, $au > 0$ and the zero-count event belongs to the undercoverage event. Its probability $(1-u)^a$ is bounded below by a positive universal constant. There are $\asymp_p K_n$ such labels, each of mass at least $1/(16n)$. This proves the reverse inequality. □

The upper half of Theorem K.2 follows immediately from Theorem K.1 and Lemma K.3.

K.5 A fixed rare layer with unequal probabilities

We now prove the lower bound and its persistence statement. Choose an even number $2m$ of labels in the rare band of Lemma K.3, with $m \asymp_p K_n$. Increasing $n_0(p)$ allows us to assume $m \ge 10^7$ and $n \ge 4m$. Relabel these finitely many selected labels as $1, \dots, 2m$. This notation does not change the source probabilities. Set

$$\omega_i = \frac{1}{16np_i} \in [1/2, 1], \qquad \boldsymbol{d}_i = m^{-1}\mathbf{1} - 2\boldsymbol{e}_i \in \mathbb{R}^{2m}, \qquad \sum_i p_i \omega_i \boldsymbol{d}_i = 0. \tag{K.18}$$

The weights belong to the predetermined loss definitions. The algorithm still takes an ordinary unweighted empirical average.

Let

$$W = \sum_{j : Z_j \text{ selected}} \omega_{Z_j}, \qquad \boldsymbol{\xi} = \frac{1}{n} \sum_{j : Z_j \text{ selected}} \omega_{Z_j} \boldsymbol{d}_{Z_j}, \qquad \boldsymbol{\mu}_S = \frac{n\boldsymbol{\xi}}{W} \quad (W > 0).$$

Direct independence of the original n observations gives

$$\mathbb{E}W = m/8, \qquad \mathrm{Var}(W) \le m/8, \qquad \mathbb{E}\boldsymbol{\xi} = 0, \qquad \mathbb{E}\|\boldsymbol{\xi}\|^2 \le \frac{m}{2n^2}. \tag{K.19}$$

Indeed, $\omega_i \le 1$, $\|\boldsymbol{d}_i\|^2 = 4 - 2/m \le 4$, and $\sum_i p_i \omega_i = m/(8n)$. Chebyshev and Markov imply

$$\Pr(\mathcal{E}) \ge \frac{1}{2}, \qquad \mathcal{E} = \left\{ \frac{.99m}{8} \le W \le \frac{1.01m}{8}, \quad \|\boldsymbol{\mu}_S\| \le \frac{32}{\sqrt{m}} \right\}. \tag{K.20}$$

For example, the two excluded probabilities are at most $80000/m$ and $32/(.99^2 \cdot 32^2)$. No conditional uniformity of the selected labels has been assumed.

On $\mathcal{E}$, an unseen selected label has $(\boldsymbol{\mu}_S)_i = 1/m$. If it was seen, its total weight is at least $1/2$, so $(\boldsymbol{\mu}_S)_i \le 1/m - 1/W < -6/m$. Thus

$$\min_i |\mu_{S,i}| \ge 1/m, \qquad \sigma_i := \operatorname{sign}(\mu_{S,i}) = +1 \quad \Longleftrightarrow \quad i \text{ is unseen.} \tag{K.21}$$

The selected observation count is at most $2W < .26m$, and hence at least m selected labels are unseen. Repeated observations are already included in these formulas.

K.6 Global regularity and population risk

Choose unit binary codewords $\boldsymbol{c}_{\boldsymbol{v}} \in \{-1/\sqrt{64m}, 1/\sqrt{64m}\}^{64m}$, $\boldsymbol{v} \in \{-1,1\}^{2m}$, with $\langle \boldsymbol{c}_{\boldsymbol{u}}, \boldsymbol{c}_{\boldsymbol{v}} \rangle \le 1/2$ for distinct messages. Existence follows from the construction in (J.9) with $k = 2m$. Fix this codebook before sampling. No efficient decoding is required.

Set

$$\begin{aligned} &A_0 = 2048, \quad K_0 = 128A_0, \quad \ell = (100A_0)^{-1}, \\ &\kappa = .9, \quad \zeta = .95, \quad \Delta = .05, \quad L_0 = \log(4\zeta/\Delta). \end{aligned}$$

Fix $C_0 \ge 256\zeta(1+L_0)/\Delta$ and $0 < \theta_0^2 \le \min\{1/64, 1/(1600A_0C_0)\}$. Put $\alpha_- = .99m/(8n)$ and suppose first that $t_0 \ge K_0C_0\sqrt{m}/\alpha_-$. Choose the following deterministic parameters:

$$\begin{aligned} &Q_0 = \alpha_- t_0/K_0, \quad a_0 = (4m)^{-1}, \quad \lambda_m = A_0/\sqrt{m}, \\ &\theta^2 = \theta_0^2 \min\{Q_0/m^{3/2}, 1\}, \quad \lambda_c = C_0/Q_0, \quad b_0 = \zeta\theta\lambda_c. \end{aligned} \tag{K.22}$$

Here a_0, b_0 are mask parameters, unrelated to the sample split. On $\mathbb{R}^{2m} \times \mathbb{R}^{64m}$ define

$$\begin{aligned} H(\boldsymbol{x}, \boldsymbol{y}) &= a_0(\|\boldsymbol{x}\|_1 - \min_i |x_i|) + b_0\|\boldsymbol{y}\|, \\ P(\boldsymbol{x}, \boldsymbol{y}) &= \max\{H(\boldsymbol{x}, \boldsymbol{y}), \max_{\boldsymbol{v}}[a_0\langle \boldsymbol{v}, \boldsymbol{x}\rangle - b_0\langle \boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{y}\rangle]\}, \\ h_{i,\theta}(\boldsymbol{y}) &= \max\{(\|\boldsymbol{y}\| - \theta)_+, \max_{\boldsymbol{v}: v_i = +1}(\langle \boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{y}\rangle - \kappa\theta)\}, \\ P_\ell(\boldsymbol{w}) &= \inf_{\boldsymbol{u}}\{P(\boldsymbol{u}) + \|\boldsymbol{w} - \boldsymbol{u}\|^2/(2\ell)\}, \qquad R_i(\boldsymbol{y}) = \inf_{\boldsymbol{u}}\{h_{i,\theta}(\boldsymbol{u}) + \|\boldsymbol{y} - \boldsymbol{u}\|^2/2\}. \end{aligned}$$

The mask properties in Lemma E.1, Theorem E.2, and Lemma E.3 apply to these parameters and codewords. In particular, memory gradient coordinates of P_ℓ have magnitude at most a_0, its code gradient has norm at most b_0, and its smoothness is ℓ^{-1}. The functions R_i are convex, nonnegative, 1-smooth and 1-Lipschitz. If all its code margins are nonpositive and $\|\boldsymbol{y}\| \le \theta$, then $R_i(\boldsymbol{y}) = 0$ and $\nabla R_i(\boldsymbol{y}) = 0$. For any admissible positive margin $s \le 1$, monotonicity of Moreau envelopes and the scalar hinge envelope give

$$R_i(\boldsymbol{y}) \ge s^2/2, \quad s = \langle \boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{y}\rangle - \kappa\theta > 0, \quad v_i = +1. \tag{K.23}$$

This inequality does not require efficient decoding or an exact multi-branch envelope formula.

Let $J(\boldsymbol{u})$ be the radial Huber function: $J(\boldsymbol{u}) = \|\boldsymbol{u}\|^2/2$ for $\|\boldsymbol{u}\| \le 1$ and $J(\boldsymbol{u}) = \|\boldsymbol{u}\| - 1/2$ otherwise. Define $\chi(\boldsymbol{w}) = \varphi((\|\boldsymbol{w}\| - 1)_+)$, where $\varphi(s) = s^2/2$ for $0 \le s \le 1$ and $\varphi(s) = s - 1/2$ for $s \ge 1$. Both J and χ are convex, 1-Lipschitz and 1-smooth. Both χ and its gradient vanish on B. For a selected label put

$$g_i(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{K_0}[P_\ell(\boldsymbol{x}, \boldsymbol{y}) + \lambda_m J(\boldsymbol{x}) + \lambda_c J(\boldsymbol{y}) + R_i(\boldsymbol{y}) - \langle \boldsymbol{d}_i, \boldsymbol{x}\rangle], \tag{K.24}$$

$$f_i(\boldsymbol{w}) = 1 + \chi(\boldsymbol{w})/8 + \omega_i g_i(\boldsymbol{w}). \tag{K.25}$$

For every other label, set $f_z(\boldsymbol{w}) = 1 + \chi(\boldsymbol{w})/8$. The losses are convex. Their Lipschitz constants are at most $1/8 + (A_0+5)/K_0 < 1$, and their smoothness constants are at most $1/8 + (101A_0+2)/K_0 < 1$. They are globally nonnegative because $g_i(\boldsymbol{w}) \ge -2\|\boldsymbol{w}\|/K_0$. With $c = 2/K_0 < 1/8$, minimization of $1 + \chi(\boldsymbol{w})/8 - c\|\boldsymbol{w}\|$ gives the positive lower bound $1 - c - 4c^2$.

The population linear terms cancel exactly by (K.18). All remaining terms are nonnegative and vanish at zero, so the unnormalized population minimum is $F_{\mathrm{raw}}(0) = 1$, and

$$F(\boldsymbol{w}) - F(0) \ge \frac{1}{16nK_0}\sum_{i=1}^{2m} R_i(\boldsymbol{y}). \tag{K.26}$$

Writing $\alpha_S = W/n$, the actual empirical objective on B is

$$G_S(\boldsymbol{w}) = 1 + \frac{\alpha_S}{K_0}\left[P_\ell + \lambda_m J(\boldsymbol{x}) + \lambda_c J(\boldsymbol{y}) + \frac{1}{W}\sum_{j:Z_j \text{ selected}} \omega_{Z_j} R_{Z_j}(\boldsymbol{y}) - \langle \boldsymbol{\mu}_S, \boldsymbol{x}\rangle\right]. \quad \text{(K.27)}$$

This identity will only be used on $\mathcal{E}$, where $W > 0$. If $W = 0$, the actual PGD trajectory remains at zero and the actual ridge solution is zero. These outputs have not been freely assigned.

K.7 PGD DYNAMICS: IDENTIFICATION, AVERAGING, AND PERSISTENCE

Fix a sample in $\mathcal{E}$. Its raw step and raw horizon are $h = \alpha_S\eta/K_0$ and $Q = \alpha_S q/K_0$. They remain random sample quantities. For $q \geq t_0$, $Q \geq Q_0$.

For memory input in orthant $\boldsymbol{\sigma}$, Lemma J.4, with $a = a_0$ and $b = b_0$, gives

$$\nabla_{\boldsymbol{y}} P_\ell(\boldsymbol{x}, \boldsymbol{y}) = A_{\boldsymbol{y}}\boldsymbol{y} - B_{\boldsymbol{y}}\boldsymbol{c}_{\boldsymbol{\sigma}}, \qquad 0 \leq A_{\boldsymbol{y}} \leq \ell^{-1}, \quad 0 \leq B_{\boldsymbol{y}} \leq b_0. \quad \text{(K.28)}$$

Thus, as $h(\ell^{-1} + \lambda_c) \leq 1$, the actual raw code step preserves

$$\boldsymbol{y} = r\boldsymbol{c}_{\boldsymbol{\sigma}}, \qquad 0 \leq r \leq \Gamma := b_0/\lambda_c = \zeta\theta. \quad \text{(K.29)}$$

To see this directly, the updated coefficient is $(1 - h[A_{\boldsymbol{y}} + \lambda_c])r + hB_{\boldsymbol{y}}$, which is nonnegative and at most $(1 - h\lambda_c)\Gamma + hb_0 = \Gamma$.

The first memory update is $h\boldsymbol{\mu}_S$, since P_ℓ has zero gradient at zero. Subsequently, the memory forcing in its signed coordinate is at least $|\mu_{S,i}| - a_0 \geq 3/(4m)$, and $h\lambda_m < 1$. As long as the terms are quadratic and the training readers are zero, the actual candidates satisfy

$$\sigma_i x_{j,i} \geq \frac{3}{4m\lambda_m}[1 - (1 - h\lambda_m)^j], \quad \text{(K.30)}$$

$$\|\boldsymbol{x}_j\| \leq \frac{\|\boldsymbol{\mu}_S\| + a_0\sqrt{2m}}{\lambda_m} \leq \frac{32 + \sqrt{2}/4}{A_0} < \frac{1}{2}. \quad \text{(K.31)}$$

Every observed selected label has $\sigma_i = -1$. Its admissible code is therefore different from $\boldsymbol{\sigma}$, and every such margin is bounded by $\Gamma/2 - \kappa\theta < 0$. By (K.29), its radial margin is also negative. All training readouts and their gradients are zero. Furthermore $\|\boldsymbol{y}\| \leq \zeta\theta < 1/8$ and (K.31) place every candidate strictly inside B. Consequently the projection is inactive, χ is flat, and J is quadratic along the entire trajectory. This proves the stated invariants by induction for the original update, with none of its training terms removed in advance.

Once $jh \geq 1/\lambda_m$, (K.30) gives

$$\min_i |x_{j,i}| - \ell a_0 \geq \frac{1}{4A_0\sqrt{m}}, \qquad a_0(\min_i |x_{j,i}| - \ell a_0) \geq \frac{1}{16A_0 m^{3/2}}.$$

On the other hand, $2b_0(\|\boldsymbol{y}_j\| + \ell b_0) \leq 3C_0\theta_0^2/m^{3/2}$, strictly less than this quantity. The exact branch certificate, Lemma E.3, applies permanently. Every subsequent code coefficient follows the exact recursion

$$r_{j+1} = (1 - h\lambda_c)r_j + hb_0. \quad \text{(K.32)}$$

Let $m_0 = \lceil 1/(h\lambda_m)\rceil$, $r_0 = \lceil L_0/(h\lambda_c)\rceil$, and $J_0 = m_0 + r_0$. For $T \geq Q_0/h$, the chosen constants give

$$\frac{J_0}{T} \leq \frac{1}{\lambda_m Q_0} + \frac{L_0}{C_0} + \frac{2h}{Q_0} \leq \frac{\Delta}{8\zeta}. \quad \text{(K.33)}$$

At and after J_0, (K.32) ensures $r_j \geq (\zeta - \Delta/4)\theta$, and r_j never decreases thereafter. Before J_0 it is nonnegative. Thus both the last point and the complete updated average satisfy

$$\langle \boldsymbol{c}_{\boldsymbol{\sigma}}, \widehat{\boldsymbol{w}}_y\rangle \geq (\kappa + \Delta/2)\theta, \qquad \|\widehat{\boldsymbol{w}}_y\| \leq \zeta\theta.$$

For each of at least m unseen labels, (K.23) now gives $R_i(\widehat{\boldsymbol{w}}_y) \geq \Delta^2\theta^2/8$. Equations (K.20) and (K.26) therefore imply

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F^*] \geq c\frac{m}{n}\theta^2 \geq c' \min\left\{\frac{m}{n}, \frac{t_0\sqrt{m}}{n^2}\right\}. \quad \text{(K.34)}$$

The constants are universal. The same loss and the same sample event work for every safe step and every later horizon, which proves the PGD part of the persistence claim.

K.8 THE SAME LOSSES: A SEPARATE RIDGE VERIFICATION

Continue to fix a sample in $\mathcal{E}$ and suppose $\lambda \le 1/t_0$. The raw ridge coefficient in (K.27) is $\gamma_S = K_0\lambda/\alpha_S \le 1/Q_0$. Consider the candidate

$$\boldsymbol{x}_* = \frac{\boldsymbol{\mu}_S - a_0\boldsymbol{\sigma}}{\lambda_m + \gamma_S}, \qquad \boldsymbol{y}_* = \frac{b_0}{\lambda_c + \gamma_S}\boldsymbol{c}_{\boldsymbol{\sigma}}. \tag{K.35}$$

Its memory norm is less than $1/2$, and $\min_i |x_{*,i}| \ge 3/[4(A_0+1)\sqrt{m}]$. Its code norm is at most $\zeta\theta$. The same strict branch certificate holds, all training readers have zero gradient, and the candidate is in the interior of B. Substituting (K.35) into the gradient of the complete original empirical objective plus its ridge penalty gives zero in both blocks. Convexity and the positive ridge coefficient make this candidate the unique actual optimizer. Furthermore,

$$\|\boldsymbol{y}_*\| \ge \frac{\zeta}{1 + 1/C_0}\theta \ge (\kappa + \Delta/2)\theta.$$

The population readout therefore gives (K.34) for every $\lambda \le 1/t_0$. This verifies the static rule directly through its optimality conditions, rather than by replacing a time symbol in a trajectory proof.

The same certificate also verifies the unregularized empirical stationary point on $\mathcal{E}$:

$$\boldsymbol{x}_\infty = (\boldsymbol{\mu}_S - a_0\boldsymbol{\sigma})/\lambda_m, \qquad \boldsymbol{y}_\infty = (b_0/\lambda_c)\boldsymbol{c}_{\boldsymbol{\sigma}}, \qquad \nabla G_S(\boldsymbol{x}_\infty, \boldsymbol{y}_\infty) = 0. \tag{K.36}$$

This point lies strictly inside B and is a global minimizer of the unconstrained convex empirical objective. Uniqueness is not asserted when the external ridge coefficient is zero.

K.9 THE REMAINING TERMS AND THE FIXED-SOURCE COMPLETION

The optimization term uses losses independent of the label. On the unit ball take $f(\boldsymbol{w}) = 1 + cw_1 + \chi(\boldsymbol{w})/2$. For PGD choose $c = \min\{1/4, 1/(4q)\}$. The total movement from zero is at most $cq \le 1/4$, whereas the optimizer is $-\boldsymbol{e}_1$. Both outputs incur at least $3c/4$. For ridge choose $c = \min\{\lambda/2, 1/4\}$. Its optimizer is $-(c/\lambda)\boldsymbol{e}_1$, of norm at most $1/2$, and its excess is at least $c/2$. These globally nonnegative convex losses obey the unit Lipschitz and smoothness constraints. They give $c' \min\{1, t^{-1}\}$ without changing p.

For the statistical term choose once and for all an event $A \subset \mathcal{Z}$ of probability $r \in (0,1)$, which exists for every source in (K.5). Put $V(z) = \mathbf{1}_A(z) - r$, $v = r(1-r)$, $u = \sqrt{v}/4$, and take

$$f_z(\boldsymbol{w}) = 1 + \chi(\boldsymbol{w})/4 + \tfrac{1}{8}[V(z) + u/\sqrt{n}]w_1. \tag{K.37}$$

These functions too are globally nonnegative, convex, 1-Lipschitz and 1-smooth. Their population slope on the ball is $u/(8\sqrt{n}) > 0$ and their population optimizer is $-\boldsymbol{e}_1$. The central limit theorem for this fixed Bernoulli variable gives, for all sufficiently large n,

$$\Pr\left\{ n^{-1}\sum_{j=1}^{n} V(Z_j) \le -u/\sqrt{n} \right\} \ge c_p > 0.$$

On this event the empirical slope is nonpositive. Every PGD iterate from zero has nonnegative first coordinate, and the actual ridge solution also has nonnegative first coordinate. Their population excess is at least $u/(8\sqrt{n})$. This proves the $c_p n^{-1/2}$ lower bound on the same fixed source, for every time.

The construction (K.25) uses $66m = O_p(K_n)$ dimensions. The other witnesses can be embedded in that space. Since $m \asymp_p K_n$, the condition on t_0 in (K.22) is implied by $t_0 \ge C_p n/\sqrt{K_n}$, and (K.34) proves (K.7). For $t \ge C_p n/\sqrt{K_n}$ use $t_0 = t$. Below this threshold, $t\sqrt{K_n}/n^2 \le C_p/n$, which is absorbed by the statistical term. The maximum of the three witnesses controls their capped sum up to a constant. This completes Theorem K.2.

K.10 ZIPF SOURCES AND INTERPRETATION

For the fixed Zipf source $p_j = j^{-s}/\zeta(s)$ with $s > 1$, $K_n \asymp_s n^{1/s}$. For all sufficiently large n and $d \geq C_s K_n$, each of the specified rules therefore satisfies

$$\mathcal{R}^{\mathsf{A}}_{p,d}(n,t) \asymp_s \min\{1, t^{-1} + n^{-1/2} + \min(n^{1/s-1}, tn^{-2+1/(2s)})\}. \tag{K.38}$$

If $1 < s < 2$, after the optimization term reaches the statistical scale, the growing term first matches $n^{-1/2}$ at $t \asymp n^{3/2-1/(2s)}$, and reaches its saturation value $n^{1/s-1}$ at $t \asymp n^{1+1/(2s)}$. For $s = 3/2$ these scales are respectively $n^{7/6}$, $n^{4/3}$, and $n^{-1/3}$. They describe the worst-loss envelope. Equation (K.7) separately supplies a fixed witness that maintains a specified late-risk level. For $s \geq 2$, the tail contribution never exceeds $n^{-1/2}$, so the envelope is $\Theta_s(n^{-1/2})$ for all $t \geq \sqrt{n}$. That last upper-bound conclusion already follows from residual coverage and is not an exclusive consequence of ID. For a general regularly varying source with $\alpha = 1/2$, the slowly varying factor remains relevant. Equation (K.6) retains it. Finally, the saturation level tends to zero at a fixed source: the result describes a loss relative to optimal statistical accuracy, not a nonvanishing asymptotic risk at a fixed Zipf distribution.

K.11 A SEPARATION WITH NONVANISHING ANCHOR VARIANCE

The rare-layer construction itself has a small gradient variance at its population optimum. Thus that instance alone does not distinguish the source bound from a suitable application of the first-projection bound. The following augmentation removes this overlap at the level of the global scalar variance input. It adds an ordinary one-dimensional noisy quadratic coordinate while retaining the original fixed source.

Proposition K.4 (A fixed-source witness with constant anchor variance)**.** *Fix a source satisfying* (K.5)*. For all sufficiently large n, and each predetermined $q_0 \geq C_p n/\sqrt{K_n}$, there is a globally nonnegative convex, 1-Lipschitz, 1-smooth loss family $\widetilde{f}_z$ in $O_p(K_n)$ dimensions whose population optimum is zero and whose optimum gradient variance obeys*

$$c_p \leq \mathbb{E}\|\nabla \widetilde{f}_Z(0) - \mathbb{E}\nabla \widetilde{f}_Z(0)\|^2 \leq 1. \tag{K.39}$$

For the original unit-ball PGD, both its last point and full updated average, at every safe step and every $q \geq q_0$, satisfy

$$\mathbb{E}[\widetilde{F}(\widehat{\boldsymbol{w}}) - \widetilde{F}(0)] \geq c_p \min\left\{K_n/n, q_0\sqrt{K_n}/n^2\right\}. \tag{K.40}$$

The same loss family satisfies this lower bound for all actual ridge minimizers with $0 < \lambda \leq 1/q_0$.

In particular, fix $p_j = j^{-3/2}/\zeta(3/2)$ and $q_0 = n^{7/6}$. For all sufficiently large n, at $q = q_0$ for PGD and at $\lambda = 1/q_0$ for ridge, the construction has

$$\mathbb{E}[\widetilde{F}(\widehat{\boldsymbol{w}}) - \widetilde{F}(0)] = \Theta_p(n^{-1/2}). \tag{K.41}$$

Its additive pair embedding has first Hoeffding gradient projection norm $\nu = \Theta_p(1)$. The global scalar-variance certificates (D.25) *and* (D.33)*, applied to that embedding, have order $n^{-1/3}$ at these same parameters.*

Proof. Fix a label z_0 of mass $r = p_{z_0} \in (0,1)$, and let $u_z = \mathbf{1}_{\{z=z_0\}} - r$. Both z_0 and r are fixed independently of n. They can be chosen for every source under consideration. Take the hard loss $f_z(\boldsymbol{w})$ in (K.25) with its predetermined persistence threshold $t_0 = q_0/2$. Increasing the constant in the assumption on q_0 ensures that all its construction conditions hold. With one additional coordinate $v \in \mathbb{R}$, define

$$\widetilde{f}_z(\boldsymbol{w}, v) = \tfrac{1}{2} f_z(\boldsymbol{w}) + \tfrac{1}{2} J(v - u_z/4), \tag{K.42}$$

where J is the scalar Huber function used above. The labels driving the two blocks are the same observed labels. No additional randomness or observations have been introduced. The summands are nonnegative and convex. Block separability bounds the smoothness by $1/2$ and the gradient norm by $1/\sqrt{2}$, which imply the required unit regularity conditions on the entire space.

The original population objective is minimized at $\boldsymbol{w} = 0$. Since $|u_z| \leq 1$, the Huber term is quadratic near $v = 0$ and its expected derivative there is $-\mathbb{E}u_Z/4 = 0$. Convexity makes $v = 0$ a

global minimizer of its population objective. Thus $(0, 0)$ is a population minimizer of the augmented objective, including over the unit ball. Its last gradient coordinate at zero equals $-u_z/8$. Consequently the total gradient variance is at least $r(1-r)/64 > 0$, proving (K.39). The upper bound follows from global Lipschitz continuity. Orthogonality of coordinates suffices for this variance statement, even though their observed labels are shared.

Fix an original sample in $\mathcal{E}$ from (K.20), and write $\bar{u} = n^{-1}\sum_{j=1}^{n} u_{Z_j}$. As long as the full-ball projection is inactive, the $\boldsymbol{w}$ block follows exactly the original hard-loss PGD with step $\eta/2$ and cumulative time $q/2$. The additional coordinate follows the exact update

$$v_{j+1} = (1-\eta/2)v_j + \eta\bar{u}/8, \qquad v_0 = 0, \tag{K.43}$$

and remains in the interval between zero and $\bar{u}/4$, hence in $[-1/4, 1/4]$. In this interval every residual $|v_j - u_z/4| \le 1/2$, so the use of the quadratic Huber branch is valid for every label, not only the observed ones. The original candidate bound (K.31) and the code bound $\|\boldsymbol{y}\| \le \zeta\theta < 1/8$ place the original $\boldsymbol{w}$ candidate at norm less than $1/2$. Together with $|v| \le 1/4$, they place the joint candidate strictly inside the unit ball. Thus induction proves both the recursions and inactivity of the actual full projection. No independence between $\bar{u}$ and $\mathcal{E}$ is needed.

For $q \ge q_0$, the old cumulative time satisfies $q/2 \ge t_0$. The same statement holds for the $\boldsymbol{w}$ component of the complete updated average, with exactly the same number of iterates. The expected population Huber term is minimized at zero, so its excess is always nonnegative, including for the averaged output. Hence (K.26) and the original good-event lower bound give half the original rare-block excess, proving (K.40) for both PGD outputs.

For ridge, multiplying the augmented empirical objective plus $\lambda(\|\boldsymbol{w}\|^2 + v^2)/2$ by two shows that the $\boldsymbol{w}$ block's unconstrained candidate is exactly the original hard-loss ridge candidate with external coefficient 2λ. Its code and memory bounds apply whenever $2\lambda \le 1/t_0$, which holds for $\lambda \le 1/q_0$. The scalar candidate is

$$v_* = \frac{\bar{u}}{4(1+2\lambda)}. \tag{K.44}$$

It lies in $[-1/4, 1/4]$, so all its training Huber losses are quadratic. The combined candidate is again strictly inside the unit ball and has zero gradient for the complete original regularized objective. Strong convexity identifies it as the actual unique ridge solution. The same population readout proves its lower bound.

Theorem K.1 applies unchanged to the augmented loss family on the same fixed source. For the specified Zipf source, $K_n \asymp n^{2/3}$ and $q_0 = n^{7/6}$, its upper bound is $O_p(n^{-1/2})$. The persistent lower at $q = q_0$ or $\lambda = 1/q_0$ has this same order, proving (K.41).

Finally, define the additive pair kernel $\widetilde{f}^{\mathrm{pair}}(\boldsymbol{a}; z, z') = [\widetilde{f}_z(\boldsymbol{a}) + \widetilde{f}_{z'}(\boldsymbol{a})]/2$. Its complete-pair empirical objective equals the original pointwise empirical objective exactly, so both algorithms and their outputs are unchanged. If $\boldsymbol{g}_z = \nabla\widetilde{f}_z(0)$, its first Hoeffding projection is $(\boldsymbol{g}_z - \mathbb{E}\boldsymbol{g}_Z)/2$. Thus $\nu^2 \ge r(1-r)/256$, and $\nu = \Theta_p(1)$. Writing $s_\nu = n^{-1} + \nu n^{-1/2}$, the displayed global first-projection certificate has size

$$\min\{1, q_0^{-1} + s_\nu + q_0 s_\nu/n\} = \Theta_p(n^{-1/3}). \tag{K.45}$$

The ridge certificate has the same order with $\lambda = 1/q_0$. This is a statement about these displayed bounds, not a lower bound on the risk of another algorithm. □

Viewed on its own, the added coordinate is a regular one-dimensional mean-estimation problem: its unconstrained recursion is $v_j = [1-(1-\eta/2)^j]\bar{u}/4$, with expected population excess at most $r(1-r)/(64n)$ under the original i.i.d. sample law. This scalar observation is not used as a conditional expectation bound on $\mathcal{E}$. The coordinate supplies nonvanishing gradient variance without supplying late overfitting. Proposition K.4 shows a strict distinction between the source guarantee and a single global first-projection variance input, even on an explicit matched instance. A coordinate-sensitive or otherwise localized analysis may separate this benign coordinate and recover the stronger rate. No impossibility claim about such analyses, or about every possible source-unaware proof, is implied.

K.12 Empirical frequencies and qualitative consistency

Qualitative consistency at a fixed countable source needs neither ID nor regular variation. Let $\widehat{p}$ be the empirical frequencies and suppose $\mathbb{E}[G_S(\widehat{\boldsymbol{w}}) - G_S(\boldsymbol{w}^*)] \le \delta_n$. Since $|f_j(\widehat{\boldsymbol{w}}) - f_j(\boldsymbol{w}^*)| \le 2$ on the unit ball, for any finite set A of m labels,

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F(\boldsymbol{w}^*)] \le \delta_n + 2\mathbb{E}\|\widehat{p} - p\|_1 \le \delta_n + 2\sqrt{m/n} + 4p(A^c).$$

The last step uses $\sum_{j\in A}\mathbb{E}|\widehat{p}_j - p_j| \le \sum_{j\in A}\sqrt{p_j(1-p_j)/n} \le \sqrt{m/n}$ and $\mathbb{E}\sum_{j\notin A}|\widehat{p}_j - p_j| \le 2p(A^c)$. Let $n\to\infty$ and then exhaust the source by finite sets. Thus $\delta_n \to 0$ implies consistency uniformly over the Lipschitz loss family and output dimension. GD and ridge have $\delta_n = O(1/t_n)$ by their actual optimization budgets. The additional contribution of Theorem K.2 is the sharp pre-saturation time term and its attainable growth and saturation.

L Three comparisons with published bounds for the same learners

This section compares explicitly identified right-hand sides with our risk guarantees, fixing the loss, sampling law and regularization or training budget, with the RERM minimizer or last GD/RCD output. A published upper-bound expression can be large even when the risk is small. The propositions below establish this separation of certificates. They are not lower bounds on the competing algorithms' risk, nor impossibility results for every stability argument. In particular, the pairwise RCD proof of Wu et al. (2025) already uses cocoercivity. The RCD/GD formula numbering follows the full versions arXiv:2503.01530v1 and arXiv:2204.12446v5, respectively.

For each comparison, the verification order is to keep the learner and output fixed, check applicability to the actual paths, evaluate the actual loss or individual-gradient inputs, optimize only admissible bound parameters, and finally apply a constant risk cap. For unprojected GD/RCD, Lemma L.2 supplies boundedness on every original and replacement sample world. The good event used by a matching lower bound alone would not justify the upper comparison.

L.1 Weak regularization in pairwise RERM

Proposition L.1 (A sharp RERM comparison)**.** *On the Euclidean unit ball, let* $\boldsymbol{A}_\lambda = \arg\min\{F_S(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2/2\}$ *for the globally nonnegative, convex,* 1*-smooth,* 1*-Lipschitz symmetric pair class, with integrable reference values. For each fixed* $c > 0$ *and every* $\lambda \ge c/n$*, its worst expected excess risk is* $\Theta_c(\min\{1, \lambda + n^{-1/2}\})$*. The upper is dimension independent and the lower needs one dimension. At* $\lambda = c/n$ *with* $c \ge 8$*, one matching family makes the expression of Lei et al. (2020, Theorem 6, Eq. (4.8)), converted to unregularized excess risk, constant-order after a constant cap. This remains true with the actual uniform smoothness and strong-convexity parameters, and under integrable pair-dependent shifts preserving global nonnegativity.*

Proof. For the upper, take $\nu = 1$ in Corollary D.7. When $\lambda \ge c/n$, its term $s_\nu/(\lambda n)$ is $O_c(n^{-1/2})$. This is a KKT application of the interaction theorem to the specified minimizer, with no small-risk assumption.

For sharpness and the comparison, let $H(s) = s^2/2$ for $0 \le s \le 1$ and $H(s) = s - 1/2$ for $s \ge 1$. Set

$$r = \tfrac{1}{2}, \quad a = \tfrac{1}{4}, \quad c_0 = \tfrac{5}{4}, \quad b = ar + a^2/2 = \tfrac{5}{32},$$
$$h(w) = H((|w| - r)_+), \qquad f_z(w) = \frac{h(w) - azw + b}{c_0}, \tag{L.1}$$
$$f(w; z, z') = \tfrac{1}{2}(f_z(w) + f_{z'}(w)).$$

The derivative $h'(w) = \operatorname{sign}(w)\min\{(|w| - r)_+, 1\}$ is monotone and 1-Lipschitz. Since $\min_w(h(w) - azw) = -b$ for $z = \pm 1$, all losses are globally nonnegative, convex and 1-Lipschitz. Their actual common smoothness is $\alpha = 4/5$, attained on the ball. Every empirical data objective is affine on $(-r, r)$, so the largest strong-convexity parameter valid uniformly over samples for the regularized objective is exactly λ.

Draw i.i.d. signs of mean $\mu = 1/(4\sqrt{n})$ and write $\bar{Z} = n^{-1}\sum_i Z_i$. The complete-pair empirical objective is exactly the pointwise average of f_{Z_i}. Its population minimizer is $w^* = r + a\mu$. On

$\{\bar{Z} \le 0\}$ the regularized minimizer is nonpositive: for $w > 0$ its objective is no smaller than at $-w$, with uniqueness supplied by the penalty. Thus

$$F(A_\lambda) - F(w^*) \ge F(0) - F(w^*) = \frac{ar\mu + a^2\mu^2/2}{c_0} \ge \frac{1}{40\sqrt{n}}. \tag{L.2}$$

Relative to the unbiased sign product law Q, the likelihood ratio L satisfies $\mathbb{E}_Q L^2 = (1+\mu^2)^n \le e^{1/16}$. Cauchy–Schwarz bounds total variation by $\frac{1}{2}\sqrt{\mathbb{E}_Q(L-1)^2}$, and hence $\Pr\{\bar{Z} \le 0\} \ge \frac{1}{2} - \frac{1}{2}\sqrt{e^{1/16}-1} > 0$. This proves the statistical lower for every $\lambda > 0$. The loss is fixed and the label law depends only on the prescribed n, which is allowed in this class.

For the regularization term, use the same H to define the deterministic pair loss $f(w; z, z') = aH(|w-1|)$, where $a = \min\{\lambda, 1\}$. This loss is globally nonnegative, convex, 1-Lipschitz and 1-smooth. On $[-1, 1]$, its population minimizer is 1, while its regularized minimizer is $A_\lambda = a/(a+\lambda)$. Consequently,

$$F(A_\lambda) - F^* = \frac{a}{2}\left(\frac{\lambda}{a+\lambda}\right)^2 \ge \frac{1}{8}\min\{1, \lambda\}.$$

Taking the larger of the two lower bounds proves the claimed sharp law.

For the external comparison, return to the family (L.1). Let σ denote its uniform strong-convexity parameter and α its smoothness parameter. At $n\sigma \ge 8\alpha$, Lei et al. (2020, Eq. (4.8)) bound the expected generalization gap by

$$\left(\frac{1024\alpha^2}{n^2\sigma^2} + \frac{64\alpha}{n\sigma}\right)\mathbb{E}F_S(A_\lambda). \tag{L.3}$$

Empirical optimality converts this to unregularized excess with additional bias at most $\lambda/2$. Here $\alpha = 4/5$ and $\sigma = \lambda = c/n$ are the most favorable valid uniform parameters, so the coefficient has constant order.

It remains to verify the actual loss input, including legal shifts. On $E = \{|\bar{Z}| \le 1/2\}$, which has probability at least $1/2$ for large n, put $g_i = f'_{Z_i}(w)$ and $\bar{g} = n^{-1}\sum_i g_i$. For every w,

$$\mathrm{Var}_{\mathrm{emp}}(g_i) = a^2(1-\bar{Z}^2)/c_0^2 \ge 3/100.$$

For $n \ge 3$, the average squared pair gradient is

$$\frac{1}{n(n-1)}\sum_{i\ne j}\left|\frac{g_i+g_j}{2}\right|^2 = \bar{g}^2 + \frac{n-2}{2(n-1)}\mathrm{Var}_{\mathrm{emp}}(g_i) \ge 3/400. \tag{L.4}$$

Shifts leave these gradients unchanged. Global nonnegativity and unit smoothness imply that each shifted loss is at least half its squared gradient norm. Thus the shifted empirical risk satisfies $\widehat{F}_S(w) \ge 3/800$ on E, also at the data-dependent A_λ. The expectation in (L.3) is bounded below by a positive constant. A diameter-based excess-risk cap of 2 leaves order one. □

The comparison fixes λ. The separately tuned penalty in Lei et al. (2020, Eq. (4.9)) changes the learner.

L.2 Bounded trajectories for unprojected updates

Lemma L.2 (Uniformly bounded initial sublevels)**.** *Suppose every empirical objective G_S is convex and 1-smooth and, for every possible sample S,*

$$\{\boldsymbol{w} : G_S(\boldsymbol{w}) \le G_S(\mathbf{0})\} \subseteq B_R.$$

Suppose also that a deterministic population minimizer belongs to B_R. Then unprojected full GD and uniform scalar-coordinate RCD, initialized at zero with $0 < \eta \le 1$, have all states and full averages in B_R. The bound holds in every original and replacement-sample world, including every realization after conditioning on a subset of observations. Consequently the bounded-trajectory ID arguments apply with constants depending only on R.

Proof. For a full step, smooth descent gives $G_S(\boldsymbol{w} - \eta\nabla G_S(\boldsymbol{w})) \le G_S(\boldsymbol{w}) - \eta(1 - \eta/2)\|\nabla G_S(\boldsymbol{w})\|^2$. The same inequality for a selected coordinate replaces the full squared gradient by its coordinate square. Every update therefore stays in its initial sublevel. Convexity of B_R gives the average bound. The hypothesis is pointwise in the entire sample, so it holds separately in every replacement world. No high-probability event is used for this upper-bound argument. After conditioning, the remaining realizations still constitute such complete sample worlds. In the RCD supply the domain is now $\mathbb{R}^D$, a product domain, and the shared-coordinate-mask identity applies directly. Descent, telescoping optimization comparison, and the ID readout use only these verified radius bounds, not an active projection. For the two witnesses below the anchor is zero, and empirical descent even gives $G_S(\boldsymbol{X}) \le G_S(\mathbf{0}) = G_S(\boldsymbol{w}^*)$ for both outputs. □

Let $A_0 = 2048$ and $K_0 = 128A_0$, as in the pair and source constructions, and let $\chi(\boldsymbol{w}) = H((\|\boldsymbol{w}\| - 1)_+)$, where $H(s) = s^2/2$ for $s \le 1$ and $H(s) = s - 1/2$ for $s \ge 1$. For the pair kernel (J.18), all nonaffine terms are nonnegative and its affine drive has norm at most 4. Hence, uniformly in the sample,

$$G_S(\boldsymbol{w}) \ge K_0^{-1}(5\chi(\boldsymbol{w}) - 4\|\boldsymbol{w}\| + 6), \qquad G_S(\mathbf{0}) = 6/K_0.$$

If $r = \|\boldsymbol{w}\| \ge 2$, then $\chi(\boldsymbol{w}) = r - 3/2$. Membership in the initial sublevel forces $r \le 15/2$. Its initial sublevel is thus contained in B_8, for every sample. The population anchor is zero by Lemma J.5. On the pair construction's good event, Proposition J.8 gives the stronger pathwise bounds

$$\|\boldsymbol{x}_t\| \le \frac{88 + \sqrt{2}/16}{A_0} < 0.05, \qquad \|\boldsymbol{y}_t\| \le \zeta\theta.$$

Thus removing interval clipping preserves its lower-bound trajectory. The B_8 bound, which holds on the complementary event as well, supplies the matching upper bound.

For the pointwise source losses (K.25), $\|\boldsymbol{d}_i\| \le 2$ and $\omega_i \le 1$ give

$$G_S(\boldsymbol{w}) \ge 1 + \chi(\boldsymbol{w})/8 - 2\|\boldsymbol{w}\|/K_0, \qquad G_S(\mathbf{0}) = 1.$$

If an initial-sublevel point had $r \ge 2$, then $r \le (3/2)/(1 - 16/K_0) < 2$, a contradiction. These sublevels lie in B_2 for every sample. The source anchor is again zero. On the good event (K.20), the direct full-gradient candidate bounds in (K.31) and Section K.7 keep the entire trajectory strictly inside B_1. Hence the source lower bound also holds for literal unprojected full GD, while Lemma L.2 justifies its upper bound on all sample worlds.

L.3 Scalar RCD and the Wu–Hu–Lei expression

For constant unit step, $q = T/D$, unit smoothness, and $f_* = \min_{\mathbb{R}^D} F$, the dependence displayed in Wu et al. (2025, Theorem 3, equation (11)) is, up to fixed numerical constants,

$$\Phi_\gamma = q^{-1} + \frac{q^{-1} + f_*}{\gamma} + (1 + \gamma)(A_q + B_q f_*), \qquad \gamma > 0, \tag{L.5}$$
$$A_q = (q + 1)/n^2, \qquad B_q = q(q + 1)/n^2.$$

Indeed, substitute $T = Dq$, $\sum_t \eta_t = T$, and $\sum_t \eta_t^2 = T$ in that equation. The initial-distance and initial-loss constants in its proof stay bounded for the unshifted witnesses. The coefficient multiplying

$B_q f_*$ is explicitly positive even if the initial point equals the population anchor. Optimizing the displayed normalized expression gives

$$\inf_{\gamma>0} \Phi_\gamma = q^{-1} + A_q + B_q f_* + 2\sqrt{(q^{-1} + f_*)(A_q + B_q f_*)}. \tag{L.6}$$

This equality optimizes the expression as defined, not unspecified constants hidden by big-O notation in the cited theorem. The order comparison below is unchanged by those fixed positive constants.

Proposition L.3 (A canonical pairwise comparison)**.** *For the canonical pair witness $\epsilon = 0$, dimension $D = 66n$, and preset target $q_0 = q \asymp n^{4/3}$, unprojected uniform scalar RCD with unit step has last-point excess risk $\Theta(n^{-2/3})$. The optimized expression in (L.6) does not tend to zero, even after capping it by a fixed-radius trivial risk bound and allowing any integrable, observation-dependent additive loss constants that preserve global nonnegativity.*

Proof. Lemma J.3 gives zero first projection. The pairwise ID upper bound, with the B_8 bridge above, is $O(q^{-1} + n^{-1} + q/n^2) = O(n^{-2/3})$. The scalar-coordinate lower bound and Proposition J.8 give the reverse inequality. The loss and its target horizon are selected before the training sample and coordinate seed.

Write $\widehat{f}_{ij}(\boldsymbol{w}) = f_{ij}(\boldsymbol{w}) + c_{ij}$ for an allowed shift, with c_{ij} fixed before sampling. Gradients, trajectories and excess risks are unchanged. Global nonnegativity and 1-smoothness imply the elementary self-bound

$$\widehat{f}_{ij}(\mathbf{0}) \geq \tfrac{1}{2}\|\nabla f_{ij}(\mathbf{0})\|^2.$$

To see this, apply smooth descent at $-\nabla f_{ij}(\mathbf{0})$ and use nonnegativity there. In the canonical construction $\nabla f_{ij}(\mathbf{0}) = (-\boldsymbol{d}_{ij}, \mathbf{0})/K_0$, $\mathbb{E}\boldsymbol{d}_{ij} = \mathbf{0}$, and $\mathbb{E}\|\boldsymbol{d}_{ij}\|^2 = 2$. Convexity leaves zero a population minimizer after the shift. Therefore

$$\widehat{f}_* \geq K_0^{-2}. \tag{L.7}$$

The lower bound is determined by the gradients, not the arbitrary original offset $6/K_0$.

Allowing the smallest legitimate smoothness parameter cannot remove this obstruction. On any radial segment with $1 < \|\boldsymbol{w}\| < 2$, the contribution $5\chi/K_0$ has second derivative $5/K_0$ and every other term is convex. The full kernel therefore has smoothness at least $5/K_0$, a constant independent of n. Thus for every legitimate common smoothness parameter, the positive $B_q\widehat{f}_*$ contribution has order at least $q^2/n^2 = n^{2/3}$. It cannot be removed by choosing γ. All output excess risks are at most 16 by the B_8 radius and global Lipschitzness, so capping the certificate leaves a constant-order guarantee, while the proven risk tends to zero. □

The next comparison grants the same population anchor and residual reader.

L.4 An explicitly derived signed-anchor baseline

Lemma L.4 (Signed-anchor certificate without interaction absorption)**.** *Under the assumptions of Theorem D.6, run the same actual full PGD or uniform scalar-coordinate RCD, using the same domain, initialization, step $0 < \eta \leq 1$, and last or full post-update average output $\widehat{\boldsymbol{w}}$. Write $h = \kappa\eta$, $q = hT$, and $s_\nu = n^{-1} + \nu n^{-1/2}$. Retaining the anchor covariance but discarding the remaining common-factor dissipation gives the valid bound*

$$\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F(\boldsymbol{w}^*)] \leq C \min\left\{1, q^{-1} + s_\nu + \frac{q s_\nu}{n} + \frac{q^2 s_\nu}{n^2}\right\}. \tag{L.8}$$

This is a bound derived here from the specified sequence of inequalities, not a limitation on all analyses with access to the same anchor.

Proof. Use the pair weights $\pi_e = \binom{n}{2}^{-1}$, incidences $p_i = 2/n$, replacement response $\mathcal{I}_t = \mathcal{I}(\boldsymbol{X}_t)$, and residual energy $U_t = U(\boldsymbol{X}_t)$ defined in Appendix D. In particular, $P_3 = \sum_i p_i^3 = 8/n^2$. Original and ghost runs share the entire coordinate seed. The current mask is independent of the data and past.

A recursion that does not retain the residual dissipation. For replacement i, split the empirical gradient difference into the common-factor part $\boldsymbol{A}_i$ and changed-factor part $\boldsymbol{D}_i$, as in Lemma D.2. With $\boldsymbol{\Delta} = \boldsymbol{x} - \boldsymbol{y}$, the common objective is convex and at most unit smooth, so $\|\boldsymbol{A}_i\|^2 \le \langle \boldsymbol{A}_i, \boldsymbol{\Delta}\rangle$, and $\|\boldsymbol{D}_i\| \le 2p_i$. The elementary two-term squared-norm bound and nonexpansiveness of the full projection give

$$\|T_S(\boldsymbol{x}) - T_{S^{(i)}}(\boldsymbol{y})\|^2 - \|\boldsymbol{\Delta}\|^2 \le -2\eta(1-\eta)\langle \boldsymbol{A}_i, \boldsymbol{\Delta}\rangle - 2\eta\langle \boldsymbol{D}_i, \boldsymbol{\Delta}\rangle + 8\eta^2 p_i^2.$$

Discard the first, nonpositive term. The shared uniform mask multiplies this increment by κ for actual coordinate updates. Summing with weights p_i and exchanging original and replacement observations as in (D.13) yields

$$\mathcal{I}_{t+1} - \mathcal{I}_t \le 4h\mathcal{C}(\boldsymbol{X}_t) + 8h\eta P_3. \tag{L.9}$$

This calculation is also valid at $\eta = 1$. The interaction sign is retained, but the factorwise negative squares are no longer present.

Retaining the anchor contribution. Define $(\mathsf{T}\boldsymbol{X})_e = \sum_{i\in e} p_i \mathsf{Q}_i \boldsymbol{X}$, and split each evaluated factor gradient as $\boldsymbol{a}_e + \nabla r_e(\boldsymbol{X})$. Self-adjointness of Q_i, together with $\mathsf{Q}_i \boldsymbol{a}_e = \boldsymbol{0}$ for $i \notin e$, gives the exact anchor identity

$$-\sum_e \pi_e \langle \boldsymbol{a}_e, (\mathsf{T}\boldsymbol{X})_e\rangle_{L^2} = -\langle \mathsf{N}\boldsymbol{\xi}, \boldsymbol{X}\rangle_{L^2} \le \sigma_1.$$

Here $\|\boldsymbol{X}\| \le 1$, and σ_1 is the same anchor variance as in (D.27). The learner is not given the anchor. It appears only in the analysis. In particular, the empirical affine gradient is not removed from its update.

For the residual part, ordinary weighted Cauchy–Schwarz gives

$$\begin{aligned}\sum_e \pi_e \|(\mathsf{T}\boldsymbol{X})_e\|_{L^2}^2 &\le \sum_e \pi_e \left(\sum_{i\in e} p_i\right) \sum_{i\in e} p_i \|\mathsf{Q}_i \boldsymbol{X}\|_{L^2}^2 \\ &= \frac{8}{n^2}\sum_i p_i \|\mathsf{Q}_i \boldsymbol{X}\|_{L^2}^2 = \frac{4}{n^2}\mathcal{I}(\boldsymbol{X}).\end{aligned}$$

The last equality uses the exact conditional-variance identity for atom replacement. Since $\sum_e \pi_e \|\nabla r_e(\boldsymbol{X})\|_{L^2}^2 \le 2U(\boldsymbol{X})$, we have, with $K_{\mathrm{src}} = 4/n^2$,

$$\mathcal{C}(\boldsymbol{X}_t) \le \sigma_1 + \sqrt{2K_{\mathrm{src}} U_t \mathcal{I}_t}. \tag{L.10}$$

No matching of a residual interaction to another atom's negative square has been used.

Trajectory energy budget and every prefix. The actual-update optimization proof (D.29)–(D.32) is independent of interaction absorption. It gives

$$\mathcal{B}_q := \sum_{t=0}^{T-1} hU_t \le \tfrac{3}{2} + q\sigma_0, \qquad \mathbb{E}[F_S(\widehat{\boldsymbol{w}}) - F_S(\boldsymbol{w}^*)] \le \frac{3}{2q}.$$

These are expectation bounds for the actual trajectory, not pointwise optimization-rate claims conditional on its seed. Let $M_T = \max_{0\le t\le T} \mathcal{I}_t$. Sum (L.9) through any deterministic prefix, apply (L.10), and use nonnegativity of the residual energies to control every prefix by the same $\mathcal{B}_q$. Time Cauchy–Schwarz gives

$$M_T \le 4q\sigma_1 + 8\eta q P_3 + 4\sqrt{2K_{\mathrm{src}} M_T}\sqrt{q\mathcal{B}_q}.$$

The inequality $b\sqrt{M} \le M/2 + b^2/2$ now implies

$$M_T \le 8q\sigma_1 + 16\eta q P_3 + 32K_{\mathrm{src}} q\mathcal{B}_q \le \frac{384 q s_\nu}{n} + \frac{256 q^2 s_\nu}{n^2}. \tag{L.11}$$

For the second inequality use $\sigma_0 \le 2s_\nu$, $\sigma_1 \le 8s_\nu/n$, $\eta \le 1$, and $s_\nu \ge 1/n$. Jensen's inequality along paired original/ghost trajectories gives $\mathcal{I}(\widehat{\boldsymbol{w}}) \le M_T$ for the full average as well as for the last point.

Conversion to population risk. Lemma D.5, with the same anchor and actual output budget, gives $\mathbb{E}[F(\widehat{\boldsymbol{w}}) - F(\boldsymbol{w}^*)] \le 3/q + 2\sigma_0 + \mathcal{I}(\widehat{\boldsymbol{w}})$. Insert (L.11) and then the Lipschitz diameter bound to obtain (L.8). This reader handles the dependence between the fitted output and its training pairs. No pair-independence assumption is made. □

For $q \le n$, both additional time terms in (L.8) are at most s_ν. Both derivations therefore recover $O(q^{-1} + s_\nu)$ throughout this interval and $O(s_\nu)$ at $q = n$, including the canonical $1/n$ statistical scale. At $\nu = 0$ and the same later horizon $q \asymp n^{4/3}$, the displayed signed-anchor certificate has order $n^{-1/3}$, while the ID certificate attains the matching risk order $n^{-2/3}$. The long-time gain comes from the residual-energy charge $q\mathcal{B}_q/n^2$ in (L.11), which interaction absorption replaces by a term of order $\sqrt{q\mathcal{B}_q}/n^{3/2} + q/n^2$ with the learner, horizon, anchor and original-loss conversion held fixed.

L.5 Full GD and actual individual-gradient path energy

For full GD, let β be a common smoothness constant and choose $\boldsymbol{W}_S^* \in \arg\min_{\boldsymbol{w}} G_S(\boldsymbol{w})$. Define the expected empirical optimization error $\epsilon_{\mathrm{opt}} = \mathbb{E}[G_S(\boldsymbol{X}_T) - G_S(\boldsymbol{W}_S^*)]$ and let

$$\epsilon_{\mathrm{path}} = \sum_{t=0}^{T-1} \eta_t \mathbb{E}\left[\frac{1}{n}\sum_{r=1}^{n} \|\nabla f(\boldsymbol{X}_t, Z_r)\|^2\right], \qquad \epsilon_c = \mathbb{E}\min_{\boldsymbol{w}} G_S(\boldsymbol{w}).$$

By permutation symmetry, this is exactly the quantity using any fixed training index in Nikolakakis et al. (2023, Definition 5). It is not the squared norm of the full empirical gradient. In their convex Theorem 10, with $\eta_t \le 1/(2\beta)$ and $q = \sum_t \eta_t$, the generalization certificate in its Eq. (12) is

$$\frac{4\sqrt{2\beta(\epsilon_{\mathrm{opt}} + \epsilon_c)\epsilon_{\mathrm{path}} q}}{n} + \frac{8\beta\epsilon_{\mathrm{path}} q}{n^2}. \tag{L.12}$$

Their Theorem 11, Eq. (14), at $\eta = 1/(2\beta)$, yields

$$8\left(\frac{1}{n} + \frac{2T}{n^2}\right)(3\beta D_0 + T\epsilon_c) + \frac{3\beta D_0}{T}, \qquad D_0 = \mathbb{E}\|\boldsymbol{X}_0 - \boldsymbol{W}_S^*\|^2. \tag{L.13}$$

The hypotheses concern globally nonnegative smooth losses and unprojected full GD. The bridge above verifies that the present witness uses precisely that update. Empirical minimizers exist and can be chosen inside B_2, so $D_0 \le 4$.

Proposition L.5 (A fixed-source path comparison)**.** *Fix a source satisfying (K.5) and $K_n \gtrsim_p \sqrt{n}$. At $q_0 = q \asymp n^{3/2}/\sqrt{K_n}$ select the rare-band witness with $m \asymp_p K_n$, and run unprojected GD with $\eta = 1/2$ and $T = \lceil 2n^{3/2}/\sqrt{K_n} \rceil$. Its last-point excess risk is $\Theta_p(n^{-1/2})$. The upper is uniform over the fixed-source class subject to the stated bounded-trajectory hypothesis. The matching family uses $O_p(K_n)$ dimensions. The certificates (L.12) and (L.13) remain at least constant order, even with the actual individual-gradient path energy and any integrable observation-dependent additive shifts preserving global nonnegativity. The same is true after a fixed-radius trivial cap.*

Proof. The fixed-source law, its rare-band lower witness, and the unprojected bridge give

$$\mathbb{E}[F(\boldsymbol{X}_T) - F^*] \asymp_p n^{-1/2}, \qquad q\sqrt{m}/n^2 \asymp_p n^{-1/2}.$$

The lower construction is admissible because its warm-up threshold is $C_p n/\sqrt{K_n}$ and the ratio of the chosen time to that scale grows as $\sqrt{n}$. Also $K_n/n \gtrsim_p n^{-1/2}$, so saturation does not remove the lower. We compute the actual inputs to the external expressions. On the event (K.20), for every active label i and every iterate,

$$\nabla_{\boldsymbol{x}} f_i(\boldsymbol{X}_t) = \frac{\omega_i}{K_0}\big(\nabla_{\boldsymbol{x}} P_\ell(\boldsymbol{X}_t) + \lambda_m \boldsymbol{x}_t - \boldsymbol{d}_i\big).$$

The source memory bound and $\|\nabla_{\boldsymbol{x}} P_\ell\| \le a_0\sqrt{2m}$, where $a_0 = 1/(4m)$, show that

$$\|\nabla_{\boldsymbol{x}} P_\ell(\boldsymbol{X}_t) + \lambda_m \boldsymbol{x}_t\| \le \frac{32 + \sqrt{2}/2}{\sqrt{m}} < 1$$

for all sufficiently large m. Because $\|\boldsymbol{d}_i\|^2 = 4 - 2/m$ and $\omega_i \geq 1/2$, we obtain the conservative bound $\|\nabla f_i(\boldsymbol{X}_t)\| \geq 1/(2K_0)$. Let N_A be the number of active observations, including repetitions. As every weight is at most one, $N_A \geq W \geq 0.99m/8$ on the same event. Its probability is at least $1/2$. Consequently, for each t,

$$\mathbb{E}\left[\frac{1}{n}\sum_{r=1}^{n}\|\nabla f_{Z_r}(\boldsymbol{X}_t)\|^2\right] \geq c\frac{m}{n}, \qquad \epsilon_{\text{path}} \geq c\frac{qm}{n}. \tag{L.14}$$

No observation is declared independent of its fitted iterate in this argument, and shifting loss values leaves the calculation unchanged.

To evaluate the smallest achievable ϵ_c, use the empirical stationary point (K.36):

$$\boldsymbol{x}_S = (\boldsymbol{\mu}_S - a_0\boldsymbol{\sigma})/\lambda_m, \qquad \boldsymbol{y}_S = (b_0/\lambda_c)\boldsymbol{c}_{\boldsymbol{\sigma}}.$$

It is strictly inside B_1, has inactive training readouts, and has zero empirical gradient. It is therefore a global empirical minimizer by convexity. For each active sample loss its memory gradient is $\omega_i(\boldsymbol{\mu}_S - \boldsymbol{d}_i)/K_0$. The bound $\|\boldsymbol{\mu}_S\| \leq 32/\sqrt{m}$ again gives norm at least $1/(2K_0)$. For every allowed shifted loss, global nonnegativity and the original unit smoothness imply $\widehat{f}_i(\boldsymbol{x}_S, \boldsymbol{y}_S) \geq \|\nabla f_i(\boldsymbol{x}_S, \boldsymbol{y}_S)\|^2/2$. Shifts preserve all empirical minimizers. Averaging over the sample, then using the same good event, proves

$$\widehat{\epsilon}_c \geq c\frac{m}{n}. \tag{L.15}$$

This calculation allows label-dependent offsets and does not merely retain the original constant one in the losses.

The source losses contain $\chi/8$, so any legitimate common smoothness parameter is at least $1/8$ by the radial second-derivative argument. In particular the common upper bound $\beta = 1$ and step $\eta = 1/2$ meet both cited theorems exactly. The second term of (L.12) is at least cq^2m/n^3 by (L.14). The positive $16T^2\widehat{\epsilon}_c/n^2$ term of (L.13) has the same order by (L.15). With $m \asymp_p K_n$ and $q \asymp n^{3/2}/\sqrt{K_n}$, we have $q^2m/n^3 \asymp_p 1$. Thus both lower bounds are constant-order. The B_2 bridge permits a trivial cap of 4, which still leaves constant order. □

Scope of the source comparison. The source witness has small anchor-gradient variance, so the first-projection ID bound also certifies its comparison rate after the additive pair embedding $\widetilde{f}(\boldsymbol{w}; z, z') = (f_z(\boldsymbol{w}) + f_{z'}(\boldsymbol{w}))/2$. It therefore does not establish that source conditioning is necessary. The fixed-source theorem is uniform over a loss class without that small-variance assumption.

L.6 General GD/RCD coverage and broader source horizons

The scalar family (L.1) also realizes $\Theta(n^{-1/2})$ risk for literal unprojected GD and scalar RCD at $q \asymp n$, without those structural restrictions. For RCD, embed it in the first of any D coordinates. The empirical pair objective is still the same scalar objective. An active-coordinate step is

$$x^+ = x - \frac{\eta}{c_0}(h'(x) - a\bar{Z}).$$

For $0 < \eta \leq 1$, this map is monotone and maps $I = [-r-a, r+a]$ into itself: its endpoint derivatives point inward, and the endpoint images remain in I. When $\bar{Z} \leq 0$, monotonicity and $x^+(0) \leq 0$ also preserve $I \cap (-\infty, 0]$. Unselected coordinates do not move. Thus every original and replacement trajectory, and all empirical and population minimizers chosen in this scalar interval, have norm at most $3/4$. The bounded-trajectory interaction upper gives $O(n^{-1/2})$ at $q = \eta T$ or $\eta T/D \asymp n$. (L.2) gives the lower for both outputs.

This also verifies the generic displayed-bound comparisons. On E, the individual-gradient energy is at least $3/100$ for every state, while (L.4) and self-bounding keep the shifted pair loss positive. At the population optimizer,

$$\mathbb{E}|\nabla f(w^*; Z, Z')|^2 = \frac{a^2(1-\mu^2)}{2c_0^2},$$

so legal nonnegative shifts cannot make the pair population minimum vanish. The $B_q f_*$ term of (L.5) therefore remains constant at $q \asymp n$, even after optimizing γ. For pointwise GD, the same

individual-gradient bound holds at every empirical minimizer, so self-bounding gives $\widehat{\epsilon}_c \gtrsim 1$ and $\epsilon_{\mathrm{path}} \gtrsim q$. The positive terms in (L.12) and (L.13) remain constant at $q \asymp n$. The admissible steps are $\eta = 1$ for RCD and $\eta = 1/2$ for GD.

For the fixed-source row, each source stays fixed across sample sizes. Every fixed Zipf law $p_j \propto j^{-s}$ with $1 < s \leq 2$ satisfies the assumptions, with $q_* \asymp_s n^{3/2-1/(2s)}$. More generally every counting exponent $\alpha > 1/2$ qualifies. At $\alpha = 1/2$ the slowly varying factor must remain bounded below. For any prescribed $q \geq q_*$, the fixed-source class has sharp late risk

$$\Theta_p\left(\min\{K_n/n, q\sqrt{K_n}/n^2\}\right).$$

The same rare-band calculation, with a witness preset for that horizon, keeps the compared formulas constant since $q^2 K_n/n^3 \geq 1$. The full envelope allows different losses at different prescribed times. It does not assert that one loss traces the increasing envelope.

L.7 Other smooth-GD comparisons

Path-dependent upper bounds. Nikolakakis et al. (2023, Theorem 9) bound squared output sensitivity for unprojected convex GD by

$$\epsilon_{\mathrm{stab}} \leq \frac{4q}{n^2}\epsilon_{\mathrm{path}}, \qquad \epsilon_{\mathrm{path}} = \sum_t \eta_t \mathbb{E}\|\nabla f(\boldsymbol{X}_t, Z_i)\|^2. \tag{L.16}$$

Their nonnegative smooth-loss framework permits unbounded gradients. With unit gradients, the substitution $\epsilon_{\mathrm{path}} \leq q$ gives $O(q^2/n^2)$, whereas bounded trajectories and interaction absorption give $O(q/n^{3/2} + Lq/n^2)$ here. These are different assumption sets. Appendix L evaluates the published path and risk expressions on the same bounded, literally unprojected trajectories.

Why the domain matters for a prior lower bound. The non-realizable quadratic witness of Zhang et al. (2025, Appendix B.1, Lemma B.1) is

$$f(w, z) = w^2/(2q) + zw, \qquad z \sim \mathrm{Unif}\{-1, 1\}, \quad q = \eta T \geq 1. \tag{L.17}$$

It is $1/q$-smooth but not globally Lipschitz. Writing $\bar{z} = n^{-1}\sum_i z_i$, unprojected GD in our post-update indexing gives $w_T = -q[1-(1-1/T)^T]\bar{z}$, hence expected risk $\Theta(q/n)$. Its full-average multiplier is $c_T = 1-(1-1/T)[1-(1-1/T)^T] \geq 1/4$, so the same order holds. In contrast, every unit-interval output satisfies

$$F(w) - F^* = w^2/(2q) \leq 1/(2q). \tag{L.18}$$

At $q = n$ the latter is $O(1/n)$: this witness cannot establish constant risk for the bounded-trajectory problem.

Finite-step and oracle lower bounds. Schliserman et al. (2025) give polynomial-dimensional differentiable convex examples driven by $\eta\sqrt{T}$, including constant risk at $T = n, \eta = \Theta(n^{-1/2})$ in $O(n^2)$ dimensions. At fixed q, this term is $\sqrt{\eta q} \to 0$ as $\eta \downarrow 0$. Differentiability also imposes no uniform smoothness budget. Our fast construction preserves a q/n mechanism in that limit, with dimension independent of T. The full-batch oracle lower bounds of Amir et al. (2021a) instead quantify over a broader class of learners. Our laws fix GD or its flow.

M Further consequences and scope

This appendix collects consequences beyond the two core laws: exact threshold conventions, dimension refinements, statistical budgets and scale changes.

M.1 Exact threshold conventions

The flow risks are defined in Appendix C. Fix either output rule $\widehat{\boldsymbol{w}}_q = \boldsymbol{w}(q)$ or $\widehat{\boldsymbol{w}}_q = q^{-1}\int_0^q \boldsymbol{w}(t)\,dt$ for all $q > 0$, and let $\mathcal{R}_{n,d,L}(\widehat{\boldsymbol{w}}_q)$ denote its worst-case expected excess risk. For GD, use the corresponding fixed output rule and take the infimum below over its fixed safe grid $q = \eta T$. The following threshold conventions supplement the full law.

Theorem M.1 (Curvature–time threshold). *Let $d_n = \lceil C_0 n^3 \log^2(2n) \rceil$ for a sufficiently large universal C_0. Fix a sufficiently small constant $r_0 > 0$. For $d \geq d_n$ and $L \geq 1$, define*

$$q_{\mathrm{crit}}(n, d, L; r_0) = \inf\{q \geq n : \mathcal{R}_{n,d,L}(\widehat{\boldsymbol{w}}_q) \geq r_0\}.$$

Then

$$q_{\mathrm{crit}}(n, d, L; r_0) = \Theta\Big(\max\Big\{n, \frac{n^2}{\sqrt{n} + L}\Big\}\Big). \tag{M.1}$$

These assertions hold for both smooth output rules and on every fixed safe GD time grid. The necessary bound is dimension independent. The regime $L \leq \sqrt{n}$ needs only linear dimension. Moreover, for either flow output,

$$\inf\{L \geq 1 : \mathcal{R}_{n,d_n,L}(\widehat{\boldsymbol{w}}_n) \geq r_0\} = \Theta(n). \tag{M.2}$$

The restriction $q \geq n$ excludes large early optimization error from the definition of overfitting. No monotonicity of the worst-case risk is assumed: the lower witnesses establish persistence throughout their claimed late intervals. Equation (M.2) concerns the exact time n, not merely an unspecified constant multiple of it. The high-curvature witnesses have polynomial parameter dimension. We do not claim an efficient gradient implementation for that faster construction.

The dimension-dependent slice remains open. At $L = n$ and $q = \eta T = n^{3/4}$, fix a safe step and either PGD output rule $\widehat{\boldsymbol{w}} \in \{\boldsymbol{w}_T, \bar{\boldsymbol{w}}_T\}$. Theorem 2.1 gives $\Theta(n^{-1/4})$ for $d \geq d_n$. With merely $d \geq Cn$, the currently established bounds are

$$cn^{-3/8} \leq \mathcal{R}_{n,d,n}(\widehat{\boldsymbol{w}}) \leq Cn^{-1/4}. \tag{M.3}$$

The lower uses the contained $\sqrt{n}$-smooth class and the low-curvature law of Theorem 2.1. The upper is the classical optimization/stability estimate. Thus it is the additional dimension parameter, not the polynomial-dimensional curvature–time diagram, that remains unresolved. The same distinctions apply to flow.

Proposition M.2 (Nonsmooth contrast). *For globally convex 1-Lipschitz losses and bounded ambient empirical subgradients, in dimension $d \geq d_n$,*

$$\begin{aligned} G_{n,d}(q) &= \Theta\big(\min\{1, q^{-1} + q/n\}\big), \\ \mathcal{R}^{\mathrm{adv}}_{n,d}(\eta, T) &= \Theta\big(\min\{1, \eta\sqrt{T} + q^{-1} + q/n\}\big). \end{aligned} \tag{M.4}$$

Here $G_{n,d}(q)$ is the worst-case expected excess risk of the constrained-flow average. The quantity $\mathcal{R}^{\mathrm{adv}}_{n,d}(\eta, T)$ is the corresponding risk of the full post-update average of zero-initialized projected subgradient descent, with an additional supremum over bounded measurable selectors. The upper bound holds for every such selector. One fixed minimum-norm selector realizes the new q/n lower bound. The $\eta\sqrt{T}$ lower mechanism uses the distinct nonsmooth instability construction of Livni (2024).

M.2 Curvature-dependent statistical and iteration budgets

Theorem 2.1 yields the statistically optimal window in Remark 2.2 by comparing each term with $n^{-1/2}$. The optimization term requires $q \gtrsim \sqrt{n}$. For $L \leq \sqrt{n}$, the growth term is at most order $n^{-1/2}$ precisely when either $q \lesssim n/L$ or $q \lesssim \sqrt{n}$. Their union is $q \lesssim n/L$. For $L \geq \sqrt{n}$, the corresponding alternatives are $q \lesssim \sqrt{n}$ or $q \lesssim n/L$, whose union is $q \lesssim \sqrt{n}$. These equivalences are up to universal constants and describe the sharp envelope in the theorem's sufficient dimension.

Restricting to $q \geq n$ and a sufficiently small fixed positive risk level gives Theorem M.1. Dividing its cumulative time by $\eta = 1/L$ yields update orders $Ln^{3/2}$ for $1 \leq L \leq \sqrt{n}$, n^2 for $\sqrt{n} \leq L \leq n$, and Ln for $L \geq n$. For the convex class in this paragraph, a smaller safe step changes the iteration count, not the law at a prescribed cumulative time. The statements compare nested curvature classes and do not assert an intervention on a fixed loss or a data-dependent stopping rule. The remaining corollaries specialize to unit smoothness where specified.

M.3 A NEAR-ERM BOUND WITH ITS CONFIDENCE PARAMETER INTEGRATED OUT

We derive the following dimension-sensitive bound from the approximate-ERM theorem of Carmon et al. (2024, Theorem 2).

Proposition M.3 (Approximate empirical comparison in finite dimension)**.** *Let $n \geq 2$ and let the losses be convex and 1-Lipschitz on the Euclidean unit ball in $\mathbb{R}^d$. Suppose a possibly randomized output $\boldsymbol{X} \in \mathbb{B}$ satisfies, almost surely,*

$$F_S(\boldsymbol{X}) \leq F_S(\boldsymbol{w}^*) + a, \qquad \boldsymbol{w}^* \in \arg\min_{\boldsymbol{w} \in \mathbb{B}} F(\boldsymbol{w}), \qquad a \geq 0.$$

Then, with a universal constant C,

$$\mathbb{E}[F(\boldsymbol{X}) - F(\boldsymbol{w}^*)] \leq C \min\left\{1, a + \frac{1}{\sqrt{n}} + \frac{d\log(3n)}{n}\right\}. \tag{M.5}$$

No independence between $\boldsymbol{X}$ and S is required.

Proof. Center each loss by its deterministic value at zero: $g_z(\boldsymbol{w}) = f(\boldsymbol{w}, z) - f(\boldsymbol{0}, z)$. On $\mathbb{B}$, $|g_z(\boldsymbol{w})| \leq 1$. Convexity, Lipschitzness, empirical comparisons and population excess risks are unchanged. Thus the boundedness assumption of the cited theorem holds without any assumption on the magnitude of the original loss offsets. For every integer $N \geq 1$, the Euclidean linear Rademacher complexity is at most $N^{-1/2}$. For vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ with $\|\boldsymbol{v}_i\| \leq 1$ and independent Rademacher signs $\epsilon_1, \ldots, \epsilon_N$, Jensen's inequality gives $\mathbb{E}_\epsilon \|N^{-1} \sum_{i=1}^N \epsilon_i \boldsymbol{v}_i\| \leq N^{-1/2}$. Consequently the inverse complexity appearing in that theorem at accuracy $\epsilon/2$ is at most $\lfloor 4/\epsilon^2 \rfloor + 1 \leq 5/\epsilon^2$ when $0 < \epsilon < 1$.

In the notation of the cited theorem, its sufficient sample threshold is therefore at most

$$\frac{12d}{\epsilon} \log \frac{3}{\epsilon} + \frac{5 + 8\log(4/\delta)}{\epsilon^2}. \tag{M.6}$$

For $0 < \delta < 1$, set

$$\epsilon_\delta = \max\left\{a, 64\left(\frac{d\log(3n)}{n} + \sqrt{\frac{1 + \log(4/\delta)}{n}}\right)\right\}.$$

If $\epsilon_\delta < 1$, then $\epsilon_\delta \geq 1/n$, so $\log(3/\epsilon_\delta) \leq \log(3n)$, and the two terms in (M.6) sum to less than n. The cited theorem supplies one event, of probability at least $1 - \delta$, on which its empirical-to-population implication holds for every point of $\mathbb{B}$. It therefore applies to every realization of the algorithmic seed. It yields $F(\boldsymbol{X}) - F(\boldsymbol{w}^*) \leq 40\epsilon_\delta$. If $\epsilon_\delta \geq 1$, this conclusion is automatic because excess risk on $\mathbb{B}$ is at most 2.

Writing $b = d\log(3n)/n$ and putting $\delta = e^{-u}$, the preceding bound implies, after enlarging a universal constant,

$$\Pr\left\{F(\boldsymbol{X}) - F(\boldsymbol{w}^*) > C\left(a + b + \sqrt{\frac{1+u}{n}}\right)\right\} \leq e^{-u} \qquad (u > 0).$$

Integrate this tail using the increasing threshold $r(u) = C(a + b + \sqrt{(1+u)/n})$:

$$\mathbb{E}[F(\boldsymbol{X}) - F(\boldsymbol{w}^*)] \leq r(0) + \int_0^\infty e^{-u} r'(u)\, du \leq C(a + b + n^{-1/2}) + \frac{C}{2\sqrt{n}}.$$

The trivial bound 2 completes the proof. In particular, confidence integration introduces no second logarithm in the dimension term. □

For safe-step full-batch PGD, both the last point and the full post-update average satisfy the pathwise empirical comparison with $a = 1/(2q)$. Combining Proposition M.3 with the unit-smooth interaction bound gives the following refinement in every dimension:

Corollary M.4 (Dimension and training time)**.** *For either output $\widehat{\boldsymbol{w}}$ of unit-smooth PGD, and likewise for constrained flow,*

$$\mathcal{R}_{n,d,1}(\widehat{\boldsymbol{w}}) \leq C \min\left\{1, \frac{1}{q} + \frac{1}{\sqrt{n}} + \min\left(\frac{q}{n^{3/2}}, \frac{d\log(3n)}{n}\right)\right\}. \tag{M.7}$$

For discrete dynamics $q = \eta T$, with $0 < \eta \leq 1$.

This is an upper bound, not a sharp characterization for each fixed d. It improves the long-time guarantee in suitable low-dimensional regimes. It does not improve the dimension-independent statistical optimum or replace the separate linear-dimension necessity result.

M.4 Accuracy, training windows, and update counts

Corollary M.5 (Samples at a prescribed training horizon)**.** *In the unit-smooth class with sufficiently large linear dimension, the conditions for a worst-case expected excess risk of order at most $\epsilon \in (0, c)$ are, up to universal constants,*

$$q \gtrsim \epsilon^{-1}, \qquad n \gtrsim \max\left\{\epsilon^{-2}, (q/\epsilon)^{2/3}\right\}. \tag{M.8}$$

More precisely, large enough constants suffice, and the matching risk lower bound forces the same conditions with possibly smaller constants.

Proof. For ϵ below the constant saturation scale, each nonnegative term in $q^{-1} + n^{-1/2} + q/n^{3/2}$ must be $O(\epsilon)$, and these conditions also suffice. Solving the latter two inequalities for n gives (M.8). □

At $n \asymp \epsilon^{-2}$, the prescribed learner's training window is $\epsilon^{-1} \lesssim q \lesssim \epsilon^{-2}$.

M.5 Persistent deterioration on one instance

Corollary M.6 (Persistent deterioration on one instance)**.** *For every sufficiently large n, one unit-smooth instance in $O(n)$ dimensions has risk $O(n^{-1/2})$ throughout $\sqrt{n} \le q \le n$ and $\Omega(1)$ at every $q \ge Cn^{3/2}$, for both outputs, all safe steps and constrained flow.*

Proof of Corollary M.6. Fix the persistent unit-smooth witness of Appendix E with $q_0 = C_0 n^{3/2}$. Its lower bound holds at every later time for both outputs, every safe GD step and constrained flow. Theorem 2.1 applied to this same instance gives $O(n^{-1/2})$ throughout $\sqrt{n} \le q \le n$. No loss or distribution is reselected between the early and late intervals. The witness starts at a population minimizer, so this argument does not make it realize the optimization-error branch or the entire growing envelope. □

M.6 Restoring radius and Lipschitz units

Let the feasible ball have radius $R > 0$, the losses be G-Lipschitz with $G > 0$, and their smoothness be bounded by $\beta > 0$. Under $\boldsymbol{x} = \boldsymbol{w}/R$ and $\widetilde{f}(\boldsymbol{x}, z) = f(R\boldsymbol{x}, z)/(GR)$, the normalized parameters are

$$L = \frac{\beta R}{G}, \qquad \widetilde{\eta} = \frac{G\eta}{R}, \qquad \widetilde{q} = \frac{Gq}{R}, \qquad q = \eta T.$$

The update, projection and either output commute exactly with this change of variables, and excess risk is multiplied by GR. In particular, when $1 \le \beta R/G \le \sqrt{n}$, the low-curvature upper bound becomes

$$C \min\left\{GR, \frac{R^2}{q} + \frac{GR}{\sqrt{n}} + GR \min\left(\frac{\beta q}{n^{3/2}}, \sqrt{\frac{Gq}{Rn^{3/2}}}\right)\right\}. \tag{M.9}$$

Matching lower bounds transfer by the same change of variables, in the dimensions specified by the normalized theorem. When $\beta = G/R$, the growing unit-smooth term in the unsaturated range is $G^2 q/n^{3/2}$. The safe-step condition remains exactly $\eta \le 1/\beta$. The high-curvature law follows by the same substitution in Theorem 2.1.

M.7 The same sharp laws with globally nonnegative losses

Global nonnegativity can be imposed while preserving the projected trajectories up to a constant time rescaling.

Proposition M.7 (A globally nonnegative extension of each witness)**.** *Let every $f_z : \mathbb{R}^d \to \mathbb{R}$ be finite, convex, globally 1-Lipschitz and L-smooth, where $L \geq 1$, and normalize $f_z(\mathbf{0}) = 0$ by subtracting its own constant. There are globally nonnegative, convex, globally 1-Lipschitz and L-smooth losses g_z such that on $\mathbb{B}$,*

$$g_z(\boldsymbol{w}) = \tfrac{1}{2} f_z(\boldsymbol{w}) + \tfrac{3}{4}, \qquad \nabla g_z(\boldsymbol{w}) = \tfrac{1}{2} \nabla f_z(\boldsymbol{w}). \tag{M.10}$$

Consequently PGD for g with step η has exactly the same iterates as PGD for f with step $\eta/2$, for every $0 < \eta \leq 1/L$. Both its last-output and full-average excess risks are exactly half those for f at cumulative time $q/2$, where $q = \eta T$. The analogous statement holds for constrained flow and its full average.

Proof. Define $g_z = \psi \circ f_z$, where

$$\psi(u) = \begin{cases} 0, & u \leq -2, \\ (u+2)^2/4, & -2 \leq u \leq -1, \\ u/2 + 3/4, & u \geq -1. \end{cases}$$

The function is nonnegative and C^1, with nondecreasing, $1/2$-Lipschitz derivative in $[0, 1/2]$. Thus g_z is convex, nonnegative and $1/2$-Lipschitz. The chain rule gives

$$\|\nabla g_z(\boldsymbol{x}) - \nabla g_z(\boldsymbol{y})\| \leq \tfrac{1}{2}\|\nabla f_z(\boldsymbol{x}) - \nabla f_z(\boldsymbol{y})\| + \tfrac{1}{2}|f_z(\boldsymbol{x}) - f_z(\boldsymbol{y})|\,\|\nabla f_z(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|,$$

using $(L+1)/2 \leq L$. On $\mathbb{B}$, $|f_z| \leq 1$, so the linear branch gives (M.10), including its endpoint. Every PGD query therefore uses half the original gradient. Induction proves the trajectory identity. The affine loss identity gives the same minimizers and half the excess risk for either output. For flow, positive scaling leaves the normal cone invariant, so $\boldsymbol{u}_g(t) = \boldsymbol{u}_f(t/2)$ solves the constrained inclusion. Uniqueness and a change of variables give the flow and time-average identities. □

The original upper bounds apply to this subclass, and transforming its lower witnesses preserves dimension and all stated output and safe-step quantifiers. Hence the unit-smooth and low-curvature sharp laws retain their orders under global nonnegativity. This does not imply interpolation: the transformed losses are at least $1/4$ on $\mathbb{B}$.

M.8 Dimension requirements and uniform stability

A standard parameter-space chaining argument (Bartlett & Mendelson, 2002) gives

$$\mathbb{E} \sup_{\boldsymbol{w} \in \mathbb{B}} [F(\boldsymbol{w}) - F_S(\boldsymbol{w})] \leq C\sqrt{d/n} \tag{M.11}$$

for globally 1-Lipschitz losses indexed by $\boldsymbol{w} \in \mathbb{B} \subset \mathbb{R}^d$. To verify this, normalize $f(0, z) = 0$ and use successive 2^{-j}-nets of the ball, of size at most $(1 + 2^{j+1})^d$. Conditional Rademacher increments are subgaussian with variance proxy $C2^{-2j}/n$, so their expected maximum at level j is at most $C2^{-j}\sqrt{d(j+1)/n}$. Summing this convergent series and symmetrizing gives (M.11). Lipschitz continuity makes the approximation remainder vanish.

Combining with optimization and classical stability yields

$$G_{n,d}(q) \leq \min\left\{2, \frac{1}{2q} + \frac{q}{n}, \frac{1}{2q} + C\sqrt{d/n}\right\}. \tag{M.12}$$

For a fixed sufficiently small risk level r_0, let $d_{\mathrm{crit}}(n; r_0) = \min\{d : G_{n,d}(n) \geq r_0\}$. Equation (M.12) and Appendix F give

$$c(r_0)n \leq d_{\mathrm{crit}}(n; r_0) \leq Cn^3 \log^2(2n). \tag{M.13}$$

This fast-horizon gap remains open. In contrast, at $q \geq Cn^{3/2}$, constant risk in the 1-smooth class requires $d = \Omega(n)$ by (M.12) and is attained with $d = O(n)$ by Appendix E.

Uniform parameter stability is a different quantity. For the affine loss $f(w, z) = -zw$ on $[-1, 1]$, with $z \in \{0, 1\}$, compare the all-zero sample with a sample containing one 1. Their flow averages differ by $q^{-1}\int_0^q \min\{t/n, 1\}\, dt \asymp \min\{q/n, 1\}$. The last flow point and both GD outputs have the same order. Thus even unique affine gradients preclude a sharper universal uniform parameter-stability law. This does not contradict the i.i.d. average-squared influence and smooth excess-risk bounds used here.

Why the trajectory constraint also matters for the gap. For i.i.d. Rademacher z, the same affine loss has $F \equiv 0$. Unprojected GD has $w_T = q\bar{z}$ and signed expected gap $\mathbb{E}[\bar{z}w_T] = q/n$, while excess risk is zero. On the unit interval, $w_T = \min\{q|\bar{z}|, 1\}\,\mathrm{sign}(\bar{z})$, so that gap is at most $\mathbb{E}|\bar{z}| \le n^{-1/2}$. This example separates both the risk quantities and the trajectory contracts. In our general bounded smooth class, (C.15) likewise improves the signed expected gap to $O(n^{-1/2})$ at $L = 1$, $q = n$. No absolute-gap bound is inferred from that one-sided inequality.