PERSISTENT OVERFITTING: THE STATISTICAL PRICE OF INTERPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

What is the statistical price of fitting n noiseless training examples exactly? In noiseless linear regression with sufficient dimension, requiring interpolation raises the minimax expected population risk from $\Theta(1/n)$ to $\Theta(\log n/n)$, with the same data and linear predictor class. Gradient descent and fixed-dataset SGD incur this statistical price at convergence, matching the minimax rate among interpolating estimators. Yet, on a single instance, their T-point prefix averages achieve expected population risk $\Theta(\log n/n^{3/2})$ at $T = n$, followed by a persistent $\Theta(\sqrt{n})$ deterioration to $\Theta(\log n/n)$, even though the iterates never move farther from the target in Euclidean distance. We explain this behavior through shared constraints and missing coordinate penalties. For general realizable quadratics, we sharply characterize how the interpolation price depends on dimension, population trace, and sample rank. Finally, we study how accurately the risk of the converged predictor can be certified from observed data. With m fresh validation examples, the optimal certification scale is $\Theta(\min\{\log n/n, 1/m\})$, even when the true risk is zero.

1 INTRODUCTION

The statistical consequences of minimizing training error are central to learning theory (Bartlett et al., 2021). Early stopping uses training duration as regularization (Yao et al., 2007; Raskutti et al., 2014), while benign-overfitting results identify when exact fitting supports accurate prediction (Bartlett et al., 2020; Hastie et al., 2022). These perspectives leave open a basic question in noiseless learning, where every observation is consistent with a common predictor: does requiring exact interpolation preserve the best statistical accuracy attainable from the sample? The existence of a perfect predictor alone does not answer this question. For gradient methods trained on a fixed sample of n examples, this issue appears in a concrete form. Zhang et al. (2025, Section 4) established regression bounds with logarithmic factors and conjectured an $O(1/n)$ expected-risk guarantee for gradient descent (GD) and stochastic gradient descent (SGD) on smooth, convex, realizable losses even as training time tends to infinity. We show that the obstruction is more fundamental than the behavior of either algorithm: exact interpolation itself can impose an unavoidable statistical cost.

Does exact interpolation incur a statistical cost? We study noiseless linear regression $y = \boldsymbol{x}^\top \boldsymbol{w}^*$, with $\|\boldsymbol{x}\| \le 1$ and $\|\boldsymbol{w}^*\| \le 1$, where both the feature distribution and the target are unknown. Population risk measures squared prediction error on fresh observations. We compare learning rules with the same n examples and output space $\mathbb{R}^d$, differing only in whether they must interpolate the training sample. In sufficient dimension, this requirement raises the minimax expected population risk from $\Theta(1/n)$ to $\Theta(\log n/n)$. The gap arises because exact fitting can force the learner to commit along directions that the sample cannot resolve even though they remain penalized by the population loss. The resulting logarithmic penalty persists for every fixed polynomial dimension $d = n^c$, $c > 0$, and the lower bound holds for every possibly randomized interpolating rule. Thus exact fitting has a statistical cost within the same linear predictor class, regardless of the interpolating algorithm.

Can gradient methods lose previously attained accuracy? With zero initialization and a suitable constant step size, GD and fixed-dataset SGD match the minimax rate among all interpolating estimators at convergence. In sufficient dimension, this rate is $\Theta(\log n/n)$, disproving the $O(1/n)$ long-horizon conjecture of Zhang et al. (2025). Yet this endpoint optimality need not preserve the accuracy attained earlier in training. For each sufficiently large n, we construct a fixed regression instance on which, from zero initialization, prefix-averaged GD and fixed-dataset SGD achieve

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.

expected population risk $\Theta(\log n/n^{3/2})$ at $T = n$, followed by a persistent $\Theta(\sqrt{n})$ deterioration to $\Theta(\log n/n)$ throughout late training and at convergence. The distribution and target are fixed across all training horizons, and the iterates never move farther from the target in Euclidean distance. The deterioration instead reflects a mismatch between empirical and population geometry: finite-sample optimization can remove a transient component that initially offsets population error in directions left unresolved by the sample. In Section 4, we make this mechanism explicit using shared constraints and missing coordinate penalties.

Can the observed data certify the quality of the converged predictor? Accurate prediction and risk certification are distinct tasks (Bartlett & Long, 2021; Gastpar et al., 2024). We ask whether the full sample and training trajectory can certify risk below the worst-case scale. We study data-dependent upper confidence bounds, valid uniformly over the model class, and ask how small they can be at zero risk. In sufficient dimension and at fixed confidence, training data alone cannot guarantee a bound below order $\log n/n$, even in this zero-risk case. We quantify how fresh validation observations improve certification accuracy without changing the predictor. Taken together, our results characterize the statistical cost of exact interpolation, the training dynamics that incur it, and the information needed to certify the resulting predictor. Our main contributions are:

1. **Sharp statistical cost of interpolation (§3):** In noiseless regression, the interpolation-constrained minimax expected population risk is $\Theta(\min\{\log(1+n),\, 1+\log d\}/n)$, compared with $\Theta(1/n)$ when exact fit is optional. The penalty is logarithmic in n when d grows as a fixed positive power of n, with $d = \Theta(\sqrt{n})$ already sufficient. We also establish matching dimension, population-trace, and sample-Hessian rank laws for normalized realizable quadratics.
2. **Persistent deterioration and sampling geometry (§4):** On one fixed regression instance, prefix-averaged GD and SGD attain expected population risk $\Theta(\log n/n^{3/2})$ at $T = n$, but their risk at convergence is $\Theta(\log n/n)$, a persistent $\Theta(\sqrt{n})$ deterioration. We obtain a sharp characterization of the risk for every training time T. We further show that different sampling laws can produce different risk curves for the same population loss, and give a sufficient geometric condition for safe continued training.
3. **Sharp limits on risk certification (§5):** With $m \geq 1$ fresh validation examples, the optimal zero-risk certification scale is $\Theta(\min\{\log n/n,\, 1/m\})$ at fixed confidence and sufficient dimension. In particular, certification at tolerance of order $1/n$ requires $\Theta(n)$ validation examples in the worst case, even when the endpoint's true risk is zero. A single endpoint certificate also yields simultaneous risk guarantees for all future iterates and averages.

2 Statistical setup and gradient methods

In this section, we formalize the noiseless regression and realizable quadratic models used throughout the paper, then define the population risk, interpolation requirement, and gradient methods. Throughout, vectors are bold and scalar coordinates are plain. The notation $\|\cdot\|$ denotes the Euclidean norm for vectors and the operator norm for matrices. All logarithms are natural. Unless stated otherwise, implicit constants are universal, and $a = \Theta(b)$ means $cb \leq a \leq Cb$ for positive universal constants c, C, uniformly over the stated parameter ranges.

Observation models. We first consider noiseless linear regression from independent and identically distributed (i.i.d.) observations $S = ((\boldsymbol{x}_i, y_i))_{i=1}^n$. Each feature $\boldsymbol{x} \in \mathbb{R}^d$ satisfies $\|\boldsymbol{x}\| \leq 1$, and its response is $y = \boldsymbol{x}^\top \boldsymbol{w}^*$. The squared loss is $f(\boldsymbol{w}; (\boldsymbol{x}, y)) = \frac{1}{2}(\boldsymbol{x}^\top \boldsymbol{w} - y)^2$. Both the feature distribution and the target $\boldsymbol{w}^*$ are unknown. We also consider i.i.d. observations of realizable quadratic losses $f(\boldsymbol{w}; z) = \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^\top H_z (\boldsymbol{w} - \boldsymbol{w}^*)$, where H_z is the sample Hessian and $0 \preceq H_z \preceq I$. Each observation gives the complete loss function, equivalently its quadratic, linear, and constant coefficients. Neither the target nor the population Hessian is supplied separately. The sampling law and target jointly specify an instance fixed before sampling. Regression is a special case with $H_z = \boldsymbol{x}\boldsymbol{x}^\top$. Optimization takes place on $\mathbb{R}^d$, without projection or a global Lipschitz assumption.

Risk and estimators. For either model, let D denote the distribution of an observation z, with $z = (\boldsymbol{x}, y)$ in regression, and let $S = (z_1, \ldots, z_n) \sim D^n$ be the training sample. The population and empirical risks are $F(\boldsymbol{w}) = \mathbb{E}_{z\sim D}[f(\boldsymbol{w}; z)]$ and $F_S(\boldsymbol{w}) = n^{-1}\sum_{i=1}^n f(\boldsymbol{w}; z_i)$, respectively. The notation F_D makes the dependence on D explicit. A parameter estimator $\widehat{\boldsymbol{w}}(S) \in \mathbb{R}^d$ may depend

measurably on the sample and independent randomness, with no restriction on its output norm. In regression, it predicts $\boldsymbol{x}^\top \widehat{\boldsymbol{w}}(S)$. It interpolates if $F_S(\widehat{\boldsymbol{w}}(S)) = 0$, equivalently, if every training loss is zero. The minimax comparisons impose $\|\boldsymbol{w}^*\| \le 1$.

Gradient methods and outputs. Both methods start from a known deterministic initialization $\boldsymbol{w}_1$, fixed before sampling and satisfying $\|\boldsymbol{w}_1 - \boldsymbol{w}^*\| \le 1$. In the minimax comparisons, $\boldsymbol{w}_1 = 0$. With constant step size η, GD updates by $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla F_S(\boldsymbol{w}_t)$, while SGD uses $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla f(\boldsymbol{w}_t; z_{i_t})$. Unless stated otherwise, we fix $\eta = 1/4$, which simplifies constants without affecting the rates studied here. For SGD, the indices i_t are i.i.d. uniform on $\{1, \dots, n\}$, independent of S, so training repeatedly samples with replacement from the fixed dataset. For either method, the full-prefix average $\bar{\boldsymbol{w}}_T = T^{-1} \sum_{t=1}^{T} \boldsymbol{w}_t$ includes the initialization and $T - 1$ subsequent updates. Expectations are over the training sample and, for SGD, its index-sampling randomness. Let $\mathbb{E}_{\mathrm{alg}}[\cdot \mid S]$ denote expectation over the SGD index sequence conditional on S.

Common interpolating endpoint. Let $H := \mathbb{E}_{z \sim D}[H_z]$ and $H_S := n^{-1} \sum_{i=1}^{n} H_{z_i}$ denote the population and empirical Hessians, respectively. The null space $\ker H_S = \{\boldsymbol{v} \in \mathbb{R}^d : H_S \boldsymbol{v} = 0\}$ consists of directions unconstrained by the training sample. Let P_S be the orthogonal projector onto this space, and let H_S^+ denote the Moore–Penrose pseudoinverse. Both trajectories and their prefix averages converge to the same endpoint, almost surely for SGD:

$$\boldsymbol{w}_\infty = \boldsymbol{w}^* + P_S(\boldsymbol{w}_1 - \boldsymbol{w}^*) = \boldsymbol{w}_1 - H_S^+ \nabla F_S(\boldsymbol{w}_1). \tag{2.1}$$

This is the interpolator closest to $\boldsymbol{w}_1$ in Euclidean distance, and the minimum-norm interpolator when $\boldsymbol{w}_1 = 0$. The pseudoinverse formula uses only the observed sample and known initialization. The projection formula shows that error survives in directions unconstrained by the sample, which may still contribute to population risk.

3 The Statistical Price of Interpolation

We first isolate the cost of exact interpolation in noiseless regression, then determine how dimension, population trace, and sample rank limit this cost for general quadratics.

3.1 Noiseless Regression: A Sharp Minimax Gap

Consider the regression model of Section 2 with $\|\boldsymbol{w}^*\| \le 1$. Let $\mathcal{I}_{n,d}$ contain all measurable, possibly randomized parameter estimators that interpolate almost surely for every admissible distribution. The interpolation-constrained minimax risk is $\mathcal{R}_{\mathrm{int}}(n, d) = \inf_{\widehat{\boldsymbol{w}} \in \mathcal{I}_{n,d}} \sup_D \mathbb{E} F_D(\widehat{\boldsymbol{w}}(S))$, where the supremum ranges over all admissible regression distributions D. The unrestricted minimax risk $\mathcal{R}_{\mathrm{free}}(n, d)$ uses the same observations and output space $\mathbb{R}^d$, with the infimum over all measurable, possibly randomized estimators.

Theorem 3.1 (Sharp price of interpolation in noiseless regression)**.** *Under the noiseless linear regression model of Section 2 with $\|\boldsymbol{x}\| \le 1$ and $\|\boldsymbol{w}^*\| \le 1$, let $\mathcal{R}_{\mathrm{free}}(n, d)$ and $\mathcal{R}_{\mathrm{int}}(n, d)$ denote the minimax expected population risks without and with the interpolation constraint, respectively. Then, for all integers $n, d \ge 1$,*

$$\mathcal{R}_{\mathrm{free}}(n, d) = \Theta\left(\frac{1}{n}\right), \qquad \mathcal{R}_{\mathrm{int}}(n, d) = \Theta\left(\frac{1}{n} \min\{\log(1 + n),\, 1 + \log d\}\right). \tag{3.1}$$

In particular, for every fixed $c > 0$ and all sufficiently large n, the logarithmic gap persists in dimension $d = \lceil n^c \rceil$, even with unit-norm features and rank-one sample Hessians. The gap constants may depend on c. The minimum-norm interpolator achieves the interpolating upper bound, as do zero-initialized GD and fixed-dataset SGD at convergence.

Remark 3.1 (**Why exact interpolation creates the statistical gap**). The two minimax problems use the same observations and linear predictor class, differing only in whether the estimator must interpolate. In the hard instances below, exact fitting forces weight onto coordinates that the sample cannot distinguish, some of which remain penalized by the population loss.

The construction uses $D_0 \le d$ active coordinates and hides the target as $\boldsymbol{w}_J^* = \boldsymbol{e}_J$ for an unknown index $J \in [D_0]$. A shared feature $\boldsymbol{b} = D_0^{-1/2} \mathbf{1}$, with label $D_0^{-1/2}$, reveals only the aggregate

constraint $\sum_{j=1}^{D_0} w_j = 1$. Non-target coordinate features $\boldsymbol{e}_j$, each appearing with probability p and label zero, rule out individual coordinates when observed. Let V contain their observed indices and $U = [D_0] \setminus V$ the remaining candidates. Under a uniform prior on J, used only for the lower-bound argument, the posterior is uniform on U, even after conditioning on the learner's independent randomness. Whenever the shared feature is observed, any interpolating output $\boldsymbol{w}$ therefore satisfies $\boldsymbol{w}_V = 0$ and $\sum_{j \in U} w_j = 1$. With $K := |U|$ and F_J denoting the population risk on instance J, posterior averaging gives

$$\mathbb{E}_J[F_J(\boldsymbol{w}) \mid S, \boldsymbol{w}] = \frac{p}{2}\left(1 - \frac{1}{K}\right) \sum_{j \in U} w_j^2 \geq \frac{p(K-1)}{2K^2}. \tag{3.2}$$

The sampling probability is chosen so that, with constant probability, the shared feature is observed and K lies between two and an absolute constant. Since $p = \Theta(\min\{\log(1+n),\, 1 + \log d\}/n)$, this forces expected population risk of at least order p, matching the interpolation-constrained rate in Theorem 3.1. Averaging over the hidden index yields one fixed hard population and target for each interpolating rule, chosen before sampling. We provide the full calibration, cover the remaining parameter regimes, and establish the matching upper bounds in Appendix L. Furthermore, we show in Proposition K.1 of Appendix K that ridge regression with regularization parameter $\lambda = 1/n$, which does not enforce interpolation, achieves the unrestricted $\Theta(1/n)$ benchmark.

3.2 What limits the interpolation price?

Section 3.1 shows that exact interpolation can incur a logarithmic statistical penalty. We now characterize how ambient dimension and total population curvature limit this cost for general realizable quadratics, with sample rank as a direct specialization. For regression, $\operatorname{tr}(H) = \mathbb{E}\|\boldsymbol{x}\|^2 \leq 1$. Under the quadratic model of Section 2, impose the population-curvature budget $\operatorname{tr}(H) \leq \tau$. Let $\mathcal{R}_{\mathrm{int}}(n, d; \tau)$ and $\mathcal{R}_{\mathrm{free}}(n, d; \tau)$ denote the minimax expected population risks over interpolating and unrestricted parameter estimators, respectively, with $\|\boldsymbol{w}^*\| \leq 1$. The relevant dimension–trace rate is

$$\Psi(n, d, \tau) := \frac{1}{n} \min\{n, \log(1 + n\tau), 1 + \log d\}. \tag{3.3}$$

Theorem 3.2 (Sharp dimension–trace law)**.** *Under the quadratic model of Section 2 with $\operatorname{tr}(H) \leq \tau$, the following hold for all integers $n, d \geq 1$ and $0 < \tau \leq d$.*

(i) Optimal risks with and without interpolation. *The minimax expected population risks satisfy*

$$\mathcal{R}_{\mathrm{int}}(n, d; \tau) = \Theta\big(\Psi(n, d, \tau)\big), \qquad \mathcal{R}_{\mathrm{free}}(n, d; \tau) = \Theta\left(\min\left\{\tau, \frac{1}{n}\right\}\right). \tag{3.4}$$

(ii) Common gradient endpoint. *Let $\boldsymbol{u} := \boldsymbol{w}_1 - \boldsymbol{w}^*$ be the initial error. GD and fixed-dataset SGD converge to the common interpolating endpoint $\boldsymbol{w}_\infty$ in* (2.1)*, almost surely for SGD. Its worst-case expected population risk satisfies*

$$\mathcal{R}_\infty(n, d; \tau) := \sup_{D,\, \|\boldsymbol{u}\| \leq 1} \mathbb{E}F(\boldsymbol{w}_\infty) = \Theta\big(\Psi(n, d, \tau)\big). \tag{3.5}$$

(iii) Sufficiently late gradient averages. *For $\mathsf{A} \in \{\mathrm{GD}, \mathrm{SGD}\}$, define the worst-case expected risk of the T-point prefix average by $\mathcal{R}_{\mathsf{A}}(n, d, T; \tau) := \sup_{D,\, \|\boldsymbol{u}\| \leq 1} \mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}})$. With step size $1/4$,*

$$\mathcal{R}_{\mathsf{A}}(n, d, T; \tau) = \Theta\big(\Psi(n, d, \tau)\big) \qquad \textit{for every integer } T \geq 32nd. \tag{3.6}$$

In (ii)–(iii)*, the suprema range over admissible quadratic laws with $\operatorname{tr}(H) \leq \tau$ and deterministic initial errors $\boldsymbol{u}$ of norm at most one, fixed before sampling. The endpoint limit is taken instancewise before the supremum. For each (n, d, τ), a single fixed instance attains the expected-risk lower bounds for both algorithms at every $T \geq 32nd$ and at convergence. When $\tau = 0$, all of these risks vanish.*

Remark 3.2 (**Why dimension and population curvature control the price**). The key obstruction is a mismatch between what the training sample constrains and what the population loss still penalizes. Exact interpolation removes error along directions detected by the sample, but may leave error in directions with nonzero population curvature. Let $Z_S = \max\{0, \lambda_{\max}(H - 2H_S)\}$ measure the largest population curvature not adequately represented by the sample. The central comparison is

$$\mathbb{E}Z_S \leq C\Psi(n, d, \tau), \qquad F(\boldsymbol{w}) \leq 2F_S(\boldsymbol{w}) + \tfrac{1}{2}Z_S \quad \text{whenever } \|\boldsymbol{w} - \boldsymbol{w}^*\| \leq 1. \tag{3.7}$$

Thus population risk separates into empirical optimization error and curvature missed by the sample. The comparison holds simultaneously for all residuals of norm at most one, including data-dependent outputs. The minimum-norm interpolator satisfies this residual bound and has zero empirical risk. Averaged gradient iterates satisfy the same bound and $\mathbb{E}_{\mathrm{alg}}[F_S(\bar{\boldsymbol{w}}_T) \mid S] = O(1/T)$. Thus minimum-norm interpolation and sufficiently late gradient training share the same missed-curvature bound. A dimension-based matrix bound gives $(1 + \log d)/n$, while a trace-sensitive argument reduces the deviation to a random subspace of expected dimension at most $2n\tau$, yielding $\log(1 + n\tau)/n$. This uses only the population trace budget, without a corresponding sample-level trace or rank bound. On the same residual ball, $F(\boldsymbol{w}) \le \tau/2$ controls the very small-trace regime. Appendix D gives the spectral-rounding and matrix-concentration argument (Theorem D.1).

These bounds clarify the different regimes. For fixed d and fixed $\tau > 0$, the interpolation-constrained risk is $\Theta(1/n)$, matching the unrestricted order. In contrast, $d = n$ and $\tau = 1$ already allow risk $\Theta(\log n/n)$. When $n\tau \le 1$, both minimax risks are $\Theta(\tau)$, while constant-order worst-case interpolation risk requires exponentially large dimension and population trace. A sample-rank constraint is a direct specialization. If $\mathrm{rank}(H_z) \le r$ almost surely, then $0 \preceq H_z \preceq I$ implies $\mathrm{tr}(H) \le r$. The matching lower constructions can also satisfy this rank bound, yielding Corollary 3.3. The full observation model remains unchanged.

Corollary 3.3 (Sharp sample-rank law)**.** *In the quadratic model of Section 2, suppose* $\mathrm{rank}\, H_z \le r$ *almost surely, where* n, d, r *are integers with* $n \ge 1$ *and* $1 \le r \le d$*. The interpolation-constrained minimax risk, the limiting worst-case risk, and the late worst-case risks of the prefix averages of GD and fixed-dataset SGD with step size* $1/4$ *are*

$$\Theta\left(\min\left\{1, \frac{1 + \log \min\{d, nr\}}{n}\right\}\right), \tag{3.8}$$

with the late-time statement holding for every integer $T \ge 32nd$*. Without the interpolation constraint, the minimax risk is* $\Theta(1/n)$*.*

Remark 3.3 (**Persistence under late selection**). The late-time lower bound remains valid even when the training horizon is chosen after inspecting the entire trajectory. On one fixed hard instance, with constant probability, every sufficiently late prefix average and the limiting endpoint simultaneously have risk at least a universal constant times $\Psi(n, d, \tau)$. More precisely, for all integers $n, d \ge 1$ and $0 < \tau \le d$, Lemma E.5 gives an admissible instance and a deterministic unit initial error, fixed before sampling and common to GD and SGD, such that each algorithm with step size $1/4$ satisfies

$$\mathbb{P}\left\{\inf_{T \in \{T_0, T_0+1, \ldots, \infty\}} F(\bar{\boldsymbol{w}}_T) \ge c\,\Psi(n, d, \tau)\right\} \ge \frac{1}{4}, \qquad T_0 = 32nd. \tag{3.9}$$

Here $c > 0$ is universal and $\bar{\boldsymbol{w}}_\infty = \boldsymbol{w}_\infty$. Thus even a data-dependent choice of $T \ge T_0$, made after observing the full trajectory, cannot uniformly avoid this risk scale. Rank-one samples already permit a logarithmic interpolation price, although their worst-case expected risk remains $O(\log(1 + n)/n)$ at convergence and at each deterministic $T \ge n$. The construction behind (3.9) differs from the low-risk-valley instance of Theorem 4.1. The threshold T_0 is sufficient, need not be sharp, and does not rule out effective stopping at earlier times. Theorem E.1 gives the matching confidence law.

4 Training dynamics and sampling geometry

Section 3 characterizes the statistical price of exact interpolation. We now show how gradient training can lose an intermediate advantage and how sampling geometry governs safe continuation.

4.1 Persistent deterioration on a fixed instance

We next show that continued optimization can destroy an accuracy advantage already attained on the same fixed problem. On one noiseless regression instance, the prefix averages of GD and fixed-dataset SGD attain low expected population risk at an intermediate time, then deteriorate by a factor $\Theta(\sqrt{n})$ and remain at the interpolation scale throughout late training and at convergence.

Theorem 4.1 (Complete risk curve on a fixed regression instance)**.** *For every sufficiently large* n*, there exists a noiseless regression instance with* $\|\boldsymbol{x}\| \le 1$ *and* $\|\boldsymbol{w}^*\| \le 1$ *in dimension* $d = \lfloor \sqrt{n} \rfloor + 1$

such that GD and fixed-dataset SGD, both initialized at $\boldsymbol{w}_1 = 0$ and using step size $1/4$, satisfy

$$\mathbb{E}F(\bar{\boldsymbol{w}}_T) = \Theta\left(\frac{1}{n(1+T)^2} + \frac{\log n}{n^{3/2}} + \frac{\log n}{n}\min\left\{\left(\frac{T\log n}{n^{3/2}}\right)^2, 1\right\}\right) \tag{4.1}$$

for every integer $T \geq 1$. The implicit constants are universal. The distribution, target, and initialization are fixed before sampling and are the same for every horizon and both algorithms.

Corollary 4.2 (Persistent deterioration after attaining low risk)**.** *On the fixed regression instance of Theorem 4.1, both zero-initialized GD and fixed-dataset SGD attain their optimal risk scale already at $T = n$:*

$$\mathbb{E}F(\bar{\boldsymbol{w}}_n) = \Theta\left(\frac{\log n}{n^{3/2}}\right), \qquad \inf_{T\in\mathbb{N},\, T\geq 1} \mathbb{E}F(\bar{\boldsymbol{w}}_T) = \Theta\left(\frac{\log n}{n^{3/2}}\right). \tag{4.2}$$

Nevertheless, there exists a universal constant $C_0 > 0$ such that, for every integer $T \geq C_0 n^{3/2}/\log n$,

$$\mathbb{E}F(\bar{\boldsymbol{w}}_T) = \Theta\left(\frac{\log n}{n}\right), \qquad \mathbb{E}F(\boldsymbol{w}_\infty) = \Theta\left(\frac{\log n}{n}\right). \tag{4.3}$$

Thus the expected risk throughout the late-training regime and at convergence is a factor $\Theta(\sqrt{n})$ larger than the risk already attained at $T = n$.

Remark 4.1 (**From a low-risk valley to the interpolation price**). The expected-risk curve in (4.1) separates three stages of training. The first term is a transient contribution that decays with T. Once this transient becomes negligible, the second term sets the intermediate low-risk scale $\Theta(\log n/n^{3/2})$, which is already attained at $T = n$. The third term is initially smaller, but grows with continued optimization and reaches the interpolation scale $\Theta(\log n/n)$ at times $T = \Theta(n^{3/2}/\log n)$. It then remains at this order through convergence. Thus the same fixed instance exhibits a low-risk valley followed by a persistent rise to the expected risk of the interpolating endpoint.

The resulting $\Theta(\sqrt{n})$ deterioration compares two training stages on this fixed instance, unlike the logarithmic gap in Theorem 3.1 between the minimax risks of interpolating and unrestricted estimator classes. Thus the intermediate risk can be $o(1/n)$ even though $\Theta(1/n)$ is the unrestricted minimax benchmark, since the latter is a worst-case guarantee over the full model class rather than a lower bound for every instance. The low-risk valley also does not require early stopping. On the same instance, ridge centered at $\boldsymbol{w}_1$ with $\lambda = 1/n$ can be optimized to convergence while retaining expected risk $\Theta(\log n/n^{3/2})$, as shown in Corollary K.2. Thus, on this instance, regularization avoids the late statistical cost of the unregularized gradient-selected endpoint even at convergence.

Remark 4.2 (**Why continued optimization can increase population risk**). The deterioration in Corollary 4.2 is not caused by the iterates moving farther from the target. Along both algorithms, Euclidean distance to $\boldsymbol{w}^*$ is nonincreasing. The key is that population risk is measured in the geometry induced by the population Hessian H, rather than by Euclidean distance. Let $\boldsymbol{e}_\infty = P_S(\boldsymbol{w}_1 - \boldsymbol{w}^*)$ denote the empirical-kernel component of the error, which is preserved throughout training, and let $\bar{\boldsymbol{r}}_T = \bar{\boldsymbol{w}}_T - \boldsymbol{w}_\infty$ denote the transient component that continued optimization removes. Since $\bar{\boldsymbol{r}}_T \in (\ker H_S)^\perp$,

$$F(\bar{\boldsymbol{w}}_T) = F(\boldsymbol{w}_\infty) + \boldsymbol{e}_\infty^\top H\bar{\boldsymbol{r}}_T + \tfrac{1}{2}\bar{\boldsymbol{r}}_T^\top H\bar{\boldsymbol{r}}_T. \tag{4.4}$$

Although $\boldsymbol{e}_\infty$ and $\bar{\boldsymbol{r}}_T$ are orthogonal in Euclidean geometry, they need not be orthogonal under H. The cross term can therefore be negative, allowing the transient component to partially cancel the persistent endpoint error in population risk. Continued optimization drives $\bar{\boldsymbol{r}}_T$ toward zero and removes this beneficial cancellation, revealing the larger risk already encoded in the empirical-kernel component. Population risk can thus increase even as Euclidean distance to the target decreases.

Mechanism behind the construction. We make the sampling geometry explicit. In error coordinates $\boldsymbol{w} - \boldsymbol{w}^* = (s, \boldsymbol{u}) \in \mathbb{R}\times\mathbb{R}^M$, take $M = \lfloor\sqrt{n}\rfloor$, $q = 64/M$, $p = 1 - q^{1/n}$, and $\boldsymbol{b} = M^{-1/2}\mathbf{1}$ for sufficiently large n. Sample the shared feature $\sqrt{2/3}(\sqrt{q}, -\boldsymbol{b})$ with probability $1/2$, the coordinate feature $(0, \boldsymbol{e}_j)/\sqrt{3}$ with probability p for each $j \in [M]$, where $\boldsymbol{e}_j$ is the j-th standard basis vector in $\mathbb{R}^M$, and zero otherwise. Thus every nonzero sample Hessian has rank one, and each coordinate feature is absent from the entire training sample with probability $q = (1-p)^n = \Theta(n^{-1/2})$. The resulting population loss is

$$F(\boldsymbol{w}) = \frac{1}{6}\left[(\sqrt{q}\, s - \boldsymbol{b}^\top\boldsymbol{u})^2 + p\|\boldsymbol{u}\|^2\right]. \tag{4.5}$$

The shared feature couples all coordinates, while missing coordinate features create the persistent empirical-kernel error. The calibration $Mq = \Theta(1)$ leaves a constant expected number of coordinate penalties unobserved, while $pq = \Theta(\log n/n^{3/2})$ sets the slow scale. Under zero initialization, the required cancellation is encoded by a target fixed before sampling. Appendix G gives the explicit translation and complete construction. For SGD, the conditional mean trajectory given the training sample equals GD, yielding the lower bound by Jensen's inequality, while the matching upper curve requires conditional covariance and energy control for repeated sampling from the fixed dataset.

Robustness of the mechanism. The deterioration is not confined to a single calibration. Varying the shared-feature coupling produces weak- and strong-coupling regimes with $\sup_{T\geq 1} \mathbb{E}F(\bar{\boldsymbol{w}}_T) = O(1/n)$ for the specified initialization family, separated by an intermediate regime with persistent larger risk. Controlled positive-curvature and smooth full-support perturbations preserve long high-risk plateaus before eventual recovery. In these perturbed constructions, the empirical Hessian is positive definite, almost surely in the smooth case, so the interpolating endpoint recovers the target. Appendices G, I, and J give the full phase diagram and robustness results.

4.2 Sampling geometry and safe continuation

Whether optimization destroys beneficial cancellation depends on how population curvature is distributed across individual samples.

Proposition 4.3 (Same population loss, different risk curves)**.** *For every sufficiently large n, the regression instance of Theorem 4.1 admits an alternative sampling law with $\|\boldsymbol{x}\| \leq 1$ in the same dimension and with the same population loss F, target, initialization, and step size $1/4$. Both laws have sample Hessians of rank at most one. For $\mathsf{A} \in \{\mathrm{GD}, \mathrm{SGD}\}$, let $\bar{\boldsymbol{w}}_T^{\mathsf{A},\mathrm{eig}}$ denote the prefix average under the alternative law. Then, uniformly over all integers $T \geq 1$,*

$$\mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A},\mathrm{eig}}) = \Theta\left(\frac{1}{n(1+T)^2} + \frac{\log n}{n^{3/2}}\right). \tag{4.6}$$

The limiting expected risk is $\Theta(\log n/n^{3/2})$. By contrast, Theorem 4.1 gives expected risk $\Theta(\log n/n)$ under the original law for every $T \geq C_0 n^{3/2}/\log n$ and at convergence. Here C_0 and the implicit constants are universal, and expectations include sample and algorithmic randomness.

Thus the population objective, initialization, and algorithm do not determine the risk curve, even up to universal constant factors, across admissible sampling laws. For the alternative law, let $H = \sum_j \lambda_j \boldsymbol{a}_j \boldsymbol{a}_j^\top$ be an orthonormal eigendecomposition. Sample $\boldsymbol{x} = \boldsymbol{a}_j$ with probability λ_j, assign the remaining mass to $\boldsymbol{x} = 0$, and set $y = \boldsymbol{x}^\top \boldsymbol{w}^*$. Since $\mathrm{tr}(H) \leq 1$, this defines a valid regression law with exactly the same population loss F. Every nonzero sample update acts along a population eigendirection, so neither method transfers error across directions and the growing term in (4.1) disappears. The following commutation condition extends this idea by providing a common contracting energy and a statistical guarantee for safe continued training.

Theorem 4.4 (Safe continuation under a common contracting energy)**.** *Under the quadratic model of Section 2, suppose a deterministic PSD matrix G and $K \geq 1$ satisfy $GH_z = H_z G$ almost surely and $G \preceq H \preceq KG$. For any step size $0 < \eta \leq 1$, both GD and fixed-dataset SGD satisfy*

$$\mathbb{E}\left[\sup_{t\geq T} F(\boldsymbol{w}_t)\right] + \mathbb{E}\left[\sup_{k\geq T} F(\bar{\boldsymbol{w}}_k)\right] \leq \frac{CK}{\eta(2-\eta)}\left(\frac{1}{T} + \frac{1}{n}\right)\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2 \tag{4.7}$$

for every integer $T \geq 1$, where C is universal. Hence, for a fixed step size, $K = O(1)$, and a deterministic $T \geq n$, both expected future suprema are $O(1/n)$, independently of dimension.

Remark 4.3 (**Why the condition guarantees safe continuation**). The matrix G defines a common energy $V_G(\boldsymbol{w}) = (\boldsymbol{w} - \boldsymbol{w}^*)^\top G(\boldsymbol{w} - \boldsymbol{w}^*)$ that contracts under every update and remains comparable to population risk. For fixed-dataset SGD, the proof conditions on the index sequence and reveals each independent training example only at its first visit. Later revisits add no independent information, but cannot increase this energy. Counting distinct visited examples yields the $1/T + 1/n$ rate while retaining the actual repeated-data dynamics. The suprema inside the expectations in Theorem 4.4 also cover choices made after inspecting the sample and full trajectory: the same expected-risk bound applies to any selected iterate or prefix average at a horizon no earlier than the deterministic T.

The condition is substantially more flexible than simultaneous diagonalization with the population Hessian. For features supported on orthogonal invariant blocks, a blockwise scalar G allows arbitrary curvature scales across blocks as long as within-block anisotropy is uniformly bounded. Appendix H gives a rank-one example, continuous in angle within each two-dimensional block, with $K = 3$ in which the sample Hessians almost surely do not commute with H, while $\mathrm{rank}(H)$ may exceed n and the global condition number may be arbitrarily large. The theorem nevertheless gives only a sufficient condition for safe continuation. Population risk itself need not decrease. With fixed step size and bounded K, an $O(1/n)$ future guarantee for $T \geq n$ need not preserve an earlier $o(1/n)$ risk. The condition depends on population geometry, and the theorem does not provide a way to verify it from the observed sample alone.

5 Certifying the training-selected predictor

The preceding results characterize the evolution of population risk along training, but do not address how sharply the risk of the training-selected endpoint can be certified from observed data. We therefore fix the observable endpoint $\boldsymbol{w}_\infty(S)$ in (2.1) and study certification of its unknown population risk $R_D(S) = F_D(\boldsymbol{w}_\infty(S))$, without modifying the predictor.

Certification task. We use the regression model of Section 2 with a known deterministic initialization satisfying $\|\boldsymbol{w}_1 - \boldsymbol{w}^*\| \leq 1$. A nonnegative certificate U may depend on the complete training sample, the full training transcript, independent randomness, and m fresh i.i.d. validation observations $Z'_1, \ldots, Z'_m$ from D, independent of training. For a resolution $a \geq 0$ and fixed failure probability $\delta_0 = 1/16$, require

$$\sup_D \mathbb{P}_D\{R_D(S) > U\} \leq \delta_0, \qquad \sup_D \mathbb{P}_D\{R_D(S) = 0,\ U > a\} \leq \delta_0. \tag{5.1}$$

The first condition requires U to be a uniformly valid upper confidence bound on the true risk. The second measures how sharply U can certify a zero-risk endpoint, using a joint, unconditional probability. The suprema range over all admissible populations and targets, and probabilities include sample, algorithmic, and certificate randomness. Define $W(n, m, d)$ as the infimum over resolutions a achievable by certificates satisfying both conditions.

Theorem 5.1 (Sharp certification resolution at zero risk)**.** *For all sufficiently large n, every $d \geq \lceil \sqrt{n} \rceil$, and every integer $m \geq 0$, the optimal certification resolution on zero-risk endpoints satisfies*

$$W(n, m, d) = \Theta\left(\min\left\{\frac{\log n}{n}, \frac{1}{m}\right\}\right), \tag{5.2}$$

where the $1/m$ term is omitted when $m = 0$. An observable certificate achieves this rate. Conversely, for every certificate with uniform coverage, there is an admissible population, target, and initialization, all fixed before sampling, for which, with probability at least $1/8$, the selected endpoint has zero population risk while the certificate is at least a universal positive multiple of the rate in (5.2)*.*

Remark 5.1 (**Why an accurate predictor can still be hard to certify**). The difficulty is informational rather than predictive. A coordinate penalty missing from the training sample may either be absent from the population or simply remain unobserved, so the same training data can select an endpoint with zero risk under one population and positive risk under another. This ambiguity persists if the validation sample also misses the distinguishing observation, which explains why fresh validation improves the optimal resolution only once m exceeds order $n/\log n$. A matching certificate combines a training-based curvature bound with the empirical validation risk $\widehat{R}_V = m^{-1}\sum_{j=1}^m f(\boldsymbol{w}_\infty; Z'_j)$, defined for $m > 0$. For suitable universal constants C_1, C_2, take

$$U = \min\left\{\frac{1}{2},\ C_1\frac{\log(1+n)}{n},\ C_2\left(\widehat{R}_V + \frac{1}{m}\right)\right\}. \tag{5.3}$$

The curvature term controls what can be certified from training alone, while validation tests events that training may have missed. Choosing the training and validation bounds to each fail with probability at most $\delta_0/2$ ensures uniform coverage for their minimum by a union bound. When $R_D(S) = 0$, every validation loss vanishes almost surely, so U achieves the sharp resolution in Theorem 5.1. For $m = 0$, the validation term is omitted. Appendix M gives the matching indistinguishability argument and explicit constants.

The certification rate has two consequences. First, under the conditions of Theorem 5.1, reliable acceptance at tolerance $0 < \varepsilon \le c_0 \log n/n$, for a sufficiently small universal c_0, has optimal validation budget $m = \Theta(1/\varepsilon)$ (Corollary M.3). In particular, certification at the $\Theta(1/n)$ scale requires $\Theta(n)$ validation examples. Second, a single valid endpoint certificate controls the entire future trajectory. On the event $R_D(S) \le U$, simultaneously for every integer $T \ge 1$,

$$\sup_{k \ge T} F_D(\bar{\boldsymbol{w}}_k) \le \left(\sqrt{U} + \frac{1}{\sqrt{2}\,T} \sum_{t=1}^{T} \|\boldsymbol{w}_t - \boldsymbol{w}_\infty\| \right)^2 . \tag{5.4}$$

Theorem M.2 gives the analogous last-iterate guarantee. The horizon may be chosen after inspecting the validation data and realized trajectory without additional confidence cost. Continued optimization can shrink the observable optimization remainder, but it provides no new population information.

6 Related work

We position our results relative to three lines of work: restrictions on learning rules, long-run behavior of gradient methods, and risk certification.

Interpolation and restrictions on learning rules. Logarithmic costs from restricting learning rules are known in consistent and proper learning (Auer & Ortner, 2007; Bousquet et al., 2020a; Høgsgaard et al., 2026), while interpolation has also been studied under label noise (Belkin et al., 2019b; Muthukumar et al., 2019; Oravkin & Rebeschini, 2021; Chinot et al., 2022). Our setting isolates exact fitting in bounded-feature noiseless squared-loss regression. The two estimator classes use the same data and full linear parameter space, differing only in the interpolation requirement. We characterize the sharp dependence on dimension, population trace, and sample-Hessian rank.

Gradient methods and sampling geometry. Stability (Bousquet & Elisseeff, 2002) and early-stopping analyses connect training duration to generalization (Yao et al., 2007; Raskutti et al., 2014; Hardt et al., 2016; Lin & Rosasco, 2017; Charles & Papailiopoulos, 2018; Lei & Ying, 2020), while structured-design and multi-pass analyses study benign interpolation and nonmonotone risk under additional spectral or distributional assumptions (Bartlett et al., 2020; Liang & Rakhlin, 2020; Hastie et al., 2022; Tsigler & Bartlett, 2023; Zou et al., 2022; Paquette et al., 2022). For smooth realizable learning, Zhang et al. (2025) derive logarithmic regression bounds and conjecture a log-free $O(1/n)$ guarantee for arbitrarily long training. Our lower bounds rule out this conjecture. Beyond the limiting rate, we show that sampling laws with the same population loss and sample Hessians of rank at most one can have sharply different risk trajectories, and give a common-energy condition guaranteeing safe continued training. Related multi-pass overfitting and fresh-sample guarantees rely on different smoothness or information assumptions (Vansover-Hager et al., 2025; Attia et al., 2025).

Risk certification and estimability. Bartlett & Long (2021) study limitations of model-dependent bounds for interpolating predictors, while Gastpar et al. (2024; 2025) distinguish learning from risk estimation with full-sample access. We instead fix the gradient-selected endpoint and ask how sharply its population risk can be certified without changing the predictor. Theorem 5.1 gives the sharp dependence on the training size n and validation size m. Appendix B gives detailed comparisons.

7 Discussion and conclusion

In this paper, we show that exact fitting in sufficient dimension raises the minimax expected population risk in noiseless regression from $\Theta(1/n)$ to $\Theta(\log n/n)$, a rate attained by zero-initialized GD and fixed-dataset SGD at convergence. On one fixed instance, however, their prefix averages achieve expected risk $\Theta(\log n/n^{3/2})$ at $T = n$ before a persistent $\Theta(\sqrt{n})$ deterioration to $\Theta(\log n/n)$ throughout late training and at convergence. Sampling laws with the same population objective can yield different risk curves, while a common contracting energy suffices for safe continuation. With m fresh validation examples, the optimal fixed-confidence certification resolution on zero-risk endpoints is $\Theta(\min\{\log n/n, 1/m\})$. Future work could sharpen finite-time risk bounds for general quadratics and broaden the conditions for safe continuation.

AI USE STATEMENT

Generative AI tools were used for language editing, structural organization, presentation improvements, and LaTeX assistance. The research questions, theoretical ideas, mathematical constructions, theorem statements, proof approaches, and mathematical derivations in this work were conceived and developed by the authors. After the proofs had been developed by the authors, AI tools were used as supplementary proof-auditing aids to stress-test the arguments and to flag possible logical gaps, overly compressed reasoning, unstated intermediate steps, implicit assumptions, edge cases, and internal inconsistencies. These checks served only as prompts for further author review, and any resulting corrections or clarifications were independently evaluated and carried out by the authors. The authors verified the final mathematical arguments and take full responsibility for the correctness and content of the manuscript.

ETHICS STATEMENT

This paper presents a mathematical analysis of interpolation and risk certification. It involves no human participants, personal data, or empirical datasets.

REPRODUCIBILITY STATEMENT

Section 2 and the theorem statements specify the assumptions, parameter ranges, and gradient-method conventions. Appendices C–N provide complete proofs and explicit constructions, linked to the main results by the proof guide. Appendix A collects the notation.

REFERENCES

Amit Attia, Matan Schliserman, Uri Sherman, and Tomer Koren. Fast last-iterate convergence of SGD in the smooth interpolation regime. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3507.

Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: PAC learning and online learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2307.03848.

Peter Auer and Ronald Ortner. A new PAC bound for intersection-closed concept classes. *Machine Learning*, 66:151–163, 2007. doi: 10.1007/s10994-006-8638-3. URL https://doi.org/10.1007/s10994-006-8638-3.

Peter L. Bartlett and Philip M. Long. Failures of model-dependent generalization bounds for least-norm interpolation. *Journal of Machine Learning Research*, 22(204):1–15, 2021. URL https://jmlr.org/papers/v22/20-1164.html.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: A statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027. URL https://arxiv.org/abs/2103.09177.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a. doi: 10.1073/pnas.1903070116. URL https://www.cs.columbia.edu/~djhsu/papers/biasvariance-pnas.pdf.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1611–1619, 2019b. URL https://proceedings.mlr.press/v89/belkin19a.html.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002. URL https://www.jmlr.org/papers/v2/bousquet02a.html.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 582–609, 2020a. URL https://proceedings.mlr.press/v125/bousquet20a.html.

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, 2020b. URL https://proceedings.mlr.press/v125/bousquet20b.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.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6): 2879–2912, 2024. URL https://arxiv.org/abs/2210.08571.

Geoffrey Chinot, Matthias Löffler, and Sara van de Geer. On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *The Annals of Statistics*, 50(4):2306–2333, 2022. doi: 10.1214/22-AOS2190. URL https://arxiv.org/abs/2012.00807.

Vitaly Feldman and Jan Vondrák. 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.

Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Fantastic generalization measures are nowhere to be found. In *Proceedings of the International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NkmJotfL42.

Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Which algorithms have tight generalization bounds? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1227. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/33ffc1a4a06e19865ac6cba04c753121-Abstract-Conference.html.

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. URL https://proceedings.mlr.press/v48/hardt16.html.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022. doi: 10.1214/21-AOS2133.

Mikael Møller Høgsgaard, Kasper Green Larsen, and Liang-Yu Zou. The interplay between interpolation and aggregation in regression: Optimal sample complexity. arXiv:2605.29819, version 2, 2026. URL https://arxiv.org/abs/2605.29819v2. Version dated July 27, 2026.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321. URL https://arxiv.org/abs/1808.03204.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 9.1–9.24, 2012. URL https://proceedings.mlr.press/v23/hsu12.html.

Bangti Jin, Zehui Zhou, and Jun Zou. On the saturation phenomenon of stochastic gradient descent for linear inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1553–1588, 2021. doi: 10.1137/20M1374456.

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. URL https://proceedings.mlr.press/v119/lei20c.html.

Jiping Li and Rishi Sonthalia. Risk phase transitions in spiked regression: Alignment driven benign and catastrophic overfitting. *arXiv preprint arXiv:2510.01414*, 2025. URL https://arxiv.org/abs/2510.01414.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020. doi: 10.1214/19-AOS1849. URL https://arxiv.org/abs/1808.00387.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017. URL https://jmlr.org/papers/v18/17-176.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Jaouad Mourtada and Lorenzo Rosasco. An elementary analysis of ridge regression with random design. *Comptes Rendus Mathématique*, 360:1055–1063, 2022. doi: 10.5802/crmath.367.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. arXiv:1903.09139, 2019. URL https://arxiv.org/abs/1903.09139.

Eduard Oravkin and Patrick Rebeschini. On optimal interpolation in linear regression. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2110.11258.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Implicit regularization or implicit conditioning? Exact risk trajectories of SGD in high dimensions. *arXiv preprint arXiv:2206.07252*, 2022. URL https://arxiv.org/abs/2206.07252.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(11):335–366, 2014. URL https://jmlr.org/papers/v15/raskutti14a.html.

James B. Simon, Dhruva Karkada, Nikhil Ghosh, and Mikhail Belkin. More is better in modern machine learning: When infinite overparameterization is optimal and overfitting is obligatory. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OdpIjS0vkO.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023. URL https://jmlr.org/papers/v24/22-1398.html.

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.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. doi: 10.1007/s00365-006-0663-2. URL https://doi.org/10.1007/s00365-006-0663-2.

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.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Risk bounds of multi-pass SGD for least squares in the interpolation regime. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12909–12920, 2022.

Contents of Appendix

The appendices contain related work, full proofs, and extensions. Appendix A collects the main symbols and links to their definitions. Appendix B gives detailed comparisons, and Appendix C supplies the gradient-method foundations.

Proof guide. For the statistical price in Section 3, Appendix L proves the minimax laws using a hidden-target posterior, Appendix D derives the algorithmic resource laws by spectral rounding, and Appendix E strengthens them to a single late-trajectory event. For Section 4, Appendices F–G establish sampling-law separation and the risk curves through spectral and covariance estimates. Appendix H proves safe continuation using first visits and a common energy. Appendix M proves Theorem 5.1 and its extensions by indistinguishability and matching certificates. Appendices I–K treat perturbations and regularization. Appendix N gives a general fluctuation comparison.

A Notation Reference **16**

B Additional related work **18**

C Persistent counterexamples and elementary estimates **19**

C.1 Detailed statements 19

C.2 The quadratic construction and finite-time GD proof 20

C.3 Fixed-dataset SGD and the exact long-time limit 21

C.4 Consequences in the original model 23

C.5 A matching theorem within the quadratic class 23

C.6 A rank-one least-squares version 24

C.7 Gradient energy and the unrestricted finite-time gap 26

D Sharp long-run resource laws **26**

D.1 Preliminary identities 26

D.2 A trace-sensitive relative covariance comparison 27

D.3 The sharp long-time trace-budget law 28

E Sharp confidence bounds for data-dependent late horizons **30**

E.1 An exponential relative-curvature comparison 31

E.2 One optimization event for the entire SGD path 31

E.3 A single event of persistent failure 32

F Same-instance dynamics and sampling-law separation **34**

F.1 A single sparse instance with a matched risk curve 34

F.2 Equal population objectives do not determine the risk curve 38

G Rank-one risk curves across two geometric safety boundaries **40**
G.1 A continuous family with a common initialization 41
G.2 Uniform spectral and GD estimates 42
G.3 A uniform upper bound for fixed-data SGD 43
G.4 Exact critical windows and the occupancy geometry 44

H Statistical safety through a comparison metric **46**
H.1 A commuting comparison metric 46

I Recovery under weak curvature **47**
I.1 Near-zero curvature: a sharp late plateau and recovery law 47

J An overfitting plateau under smooth full-support sampling **49**
J.1 A pathwise feature-perturbation bound 49
J.2 Explicit smooth distributions and time scales 50

K Protection by observable regularization **51**
K.1 An observable objective with an optimal fully optimized guarantee 51
K.2 Ridge retains the low-risk valley under both sampling laws 52

L Interpolation-constrained minimax risk **53**
L.1 Experiment, estimator classes, and main laws 54
L.2 A hidden-target posterior lemma 55
L.3 Calibration of dimension and population trace 55
L.4 An exactly rank-one sampling law 56
L.5 A sample-rank ceiling for arbitrary interpolators 56
L.6 Relationship to prior interpolation and proper-learning bounds 57

M The information required to certify continued optimization **58**
M.1 Experiment and certification frontier 58
M.2 From the endpoint to every future output 60
M.3 A sound acceptance rule with a positive-risk margin 60

N A fluctuation bridge for arbitrary quadratic observations **61**

A NOTATION REFERENCE

The table collects the main notation and links to its definition or first substantive use. Construction-specific symbols retain their local meaning. For example, the candidate set U and its size K differ from the certificate U and energy-comparison constant K.

Table 1: Notation reference.

Notation	Meaning	Defined in
Observation model and risk		
n, d	Training sample size and parameter dimension.	Section 2
$D,\ S = (z_i)_{i=1}^n$	Observation law and i.i.d. sample, $S \sim D^n$.	Section 2
$\boldsymbol{x}, y, \boldsymbol{w}^*$	Feature, response $y = \boldsymbol{x}^\top \boldsymbol{w}^*$, and unknown target.	Section 2
$f(\boldsymbol{w}; z)$	Observed squared or realizable quadratic loss.	Section 2
H_z	Sample Hessian, $0 \preceq H_z \preceq I$. In regression, $H_z = \boldsymbol{x}\boldsymbol{x}^\top$.	Section 2
$F,\ F_D,\ F_S$	Population risk (with law explicit in F_D) and empirical risk.	Section 2
$H,\ H_S$	Population Hessian $\mathbb{E}H_z$ and empirical Hessian $n^{-1}\sum_i H_{z_i}$.	Section 2
$\widehat{\boldsymbol{w}}(S)$	Possibly randomized parameter estimator with unrestricted output norm.	Section 2
$\|\cdot\|,\ \|\cdot\|_H$	Euclidean or operator norm. $\|\boldsymbol{v}\|_H^2 = \boldsymbol{v}^\top H \boldsymbol{v}$ is the population seminorm squared.	Section 2, Appendix G
Gradient methods and outputs		
$\boldsymbol{w}_1, \eta, i_t$	Known deterministic initialization fixed before sampling, step size, and i.i.d. uniform SGD index in the fixed dataset.	Section 2
$\boldsymbol{w}_t,\ \bar{\boldsymbol{w}}_T$	Iterate and prefix average $T^{-1}\sum_{t=1}^T \boldsymbol{w}_t$, including initialization.	Section 2
$\mathbb{E}_{\mathrm{alg}}[\cdot \mid S]$	Expectation over SGD indices conditional on the sample.	Section 2
$P_S,\ H_S^+$	Projector onto $\ker H_S$ and Moore–Penrose pseudoinverse of H_S.	Section 2
$\boldsymbol{w}_\infty$	Common endpoint $\boldsymbol{w}^* + P_S(\boldsymbol{w}_1 - \boldsymbol{w}^*)$.	Eq. (2.1)
Statistical price and resource laws		
$\mathcal{I}_{n,d}$	Possibly randomized estimators that interpolate almost surely under every admissible law.	Section 3.1
$\mathcal{R}_{\mathrm{int}},\ \mathcal{R}_{\mathrm{free}}$	Minimax expected risks with and without the interpolation requirement.	Section 3.1
D_0, J, V, U, K	Hidden-target construction: active dimension, target index, observed set, candidate set, and $K = \|U\|$.	Remark 3.1
$\tau,\ r$	Population trace budget and sample-rank ceiling: $\mathrm{tr}(H) \le \tau$ and $\mathrm{rank}(H_z) \le r$ almost surely.	Section 3.2
$\Psi(n, d, \tau)$	Dimension–trace rate $n^{-1}\min\{n, \log(1+n\tau), 1+\log d\}$.	Eq. (3.3)
$\mathcal{R}_\infty,\ \mathcal{R}_{\mathsf{A}}$	Worst-case expected endpoint and prefix-average risks, $\mathsf{A} \in \{\mathrm{GD}, \mathrm{SGD}\}$.	Theorem 3.2
$\boldsymbol{u}$	Deterministic initial error $\boldsymbol{w}_1 - \boldsymbol{w}^*$, with $\|\boldsymbol{u}\| \le 1$, in resource-law suprema.	Theorem 3.2
Z_S	Missed-curvature deviation $\max\{0, \lambda_{\max}(H - 2H_S)\}$.	Remark 3.2

Continued on the next page.

Table 1. Notation reference (continued).

Notation	Meaning	Defined in
Late-horizon confidence		
T_0	Sufficient late-horizon threshold $32nd$ for the resource-law construction.	Remark 3.3
$\Psi_\delta,\ q_{1-\delta}$	Confidence-dependent resource rate and $(1-\delta)$ quantile.	Appendix E
σ	Data- and trajectory-dependent choice of a prefix-average horizon, restricted to $\sigma \geq T_0$.	Appendix E
Training dynamics and safe continuation		
$\boldsymbol{e}_\infty,\ \bar{\boldsymbol{r}}_T$	Persistent error $P_S(\boldsymbol{w}_1 - \boldsymbol{w}^*)$ and transient $\bar{\boldsymbol{w}}_T - \boldsymbol{w}_\infty$.	Remark 4.2
$M, p, q, \boldsymbol{b}$	Shared-penalty construction: coordinate count, penalty probability, absence probability $q = (1-p)^n$, and $\boldsymbol{b} = M^{-1/2}\mathbf{1}$.	Eq. (4.5)
$\bar{\boldsymbol{w}}_T^{\mathsf{A},\mathrm{eig}}$	Prefix average under the alternative population-eigenvector sampling law.	Proposition 4.3
G, K	Deterministic PSD metric and $K \geq 1$ with $GH_z = H_z G$ almost surely and $G \preceq H \preceq KG$.	Theorem 4.4
$V_G(\boldsymbol{w})$	Common energy $(\boldsymbol{w} - \boldsymbol{w}^*)^\top G(\boldsymbol{w} - \boldsymbol{w}^*)$.	Remark 4.3
Endpoint certification		
$R_D(S)$	Risk $F_D(\boldsymbol{w}_\infty(S))$ of the training-selected endpoint.	Section 5
$m,\ Z'_j$	Validation sample size and fresh observations independent of training.	Section 5, Remark 5.1
U, a, δ_0	Upper confidence certificate, zero-risk resolution, and fixed failure probability $\delta_0 = 1/16$.	Eq. (5.1)
$W(n, m, d)$	Infimum of attainable resolutions under uniform coverage and the joint zero-risk requirement.	Section 5
$\widehat{R}_V$	Validation risk $m^{-1}\sum_{j=1}^m f(\boldsymbol{w}_\infty; Z'_j)$, defined for $m > 0$.	Remark 5.1
Selected appendix notation		
$\boldsymbol{w}_0$	Initial vector in target-zero error coordinates, indexed from zero in Appendix G.	Opening of Appendix G
$s, \rho, \vartheta, \Lambda$	Coupling-family parameter, ratio $\rho = s/q$, factor $\vartheta = \rho/(1+\rho)^2$, and slow scale $\Lambda = p(s+q)$.	Theorem G.2
$\widehat{\boldsymbol{w}}_\lambda, \lambda$	Ridge estimator centered at $\boldsymbol{w}_1$ and positive regularization parameter.	Eq. (K.1)
ε, ϕ	Acceptance tolerance and binary acceptance rule. Validation does not change the endpoint.	Corollary M.3
$d_t, \bar{d}_T$	Distance $\|\boldsymbol{w}_t - \boldsymbol{w}_\infty\|$ and its prefix mean $T^{-1}\sum_{t=1}^T d_t$.	Theorem M.2
$g_{T,\eta}(H_S)$	Averaging filter $T^{-1}\sum_{j=0}^{T-1}(I - \eta H_S)^j$ in target-zero coordinates.	Appendix N
$\Delta_T(S), \Pi$	Conditional SGD–GD excess risk and projector onto $\operatorname{ran} H_S$.	Theorem N.1

B ADDITIONAL RELATED WORK

Interpolation and restrictions on learning rules. Logarithmic penalties arise in several learning problems under different restrictions on the learner. Auer & Ortner (2007) show that some consistent rules incur a logarithmic cost even when a better proper rule exists. Bousquet et al. (2020a) establish obstructions applying to every proper learner on suitable classification classes. Høgsgaard et al. (2026) obtain proper-learning penalties in realizable regression under cutoff loss, while Attias et al. (2023) characterize realizable regression and ERM through combinatorial dimensions. Noise-fitting costs constrain even optimally chosen interpolators (Muthukumar et al., 2019), and optimal response-linear interpolators can exploit covariance and signal-prior information (Oravkin & Rebeschini, 2021). Our comparison fixes the data and full linear output space, changing only the requirement of exact fitting under bounded-feature noiseless squared loss. Appendix L gives the detailed statistical comparison.

Stability and the long-training conjecture. Stability controls generalization through sensitivity to sample perturbations (Bousquet & Elisseeff, 2002), yielding duration-dependent SGD bounds (Hardt et al., 2016), fast rates at low optimal loss (Srebro et al., 2010; Lei & Ying, 2020), and convergence-based guarantees under additional geometry (Charles & Papailiopoulos, 2018). High-probability refinements sharpen the consequences of uniform stability (Feldman & Vondrák, 2019; Bousquet et al., 2020b). Realizability alone does not ensure these stability conditions. Zhang et al. (2025, Section 4) derive regression bounds of $O(1/(\eta T) + \log^3 n/n)$ and conjecture $O(1/n)$ risk as $\eta T \to \infty$. Our $\Theta(\log n/n)$ limiting law refutes the log-free conjecture while remaining consistent with those bounds. The sharp finite-time worst-case risk over arbitrary i.i.d. quadratics remains open, as detailed in Appendix N.

Data reuse, fresh observations, and regularization. Vansover-Hager et al. (2025) show rapid multi-pass overfitting for nonsmooth convex losses. For least squares, Zou et al. (2022) decompose multi-pass SGD risk into GD risk and a nonnegative fluctuation, while Paquette et al. (2022) characterize high-dimensional trajectories under structural assumptions. Our upper risk curve on the same instance requires an additional fluctuation bound. Fresh-observation guarantees (Attia et al., 2025) use independent data at every update. Applied to empirical resampling, they control the empirical objective. Early stopping regularizes gradient methods (Yao et al., 2007; Raskutti et al., 2014; Jin et al., 2021), including through the number of passes (Lin & Rosasco, 2017). Ridge analysis (Mourtada & Rosasco, 2022) supplies our observable unrestricted comparator.

Benign interpolation and sampling geometry. Interpolation can attain optimal nonparametric rates under suitable smoothness and design assumptions (Belkin et al., 2019b), and kernel interpolation can generalize through favorable spectral geometry (Liang & Rakhlin, 2020). Analyses under covariance and design assumptions characterize benign minimum-norm and ridge regression (Bartlett et al., 2020; Hastie et al., 2022; Tsigler & Bartlett, 2023), quantify random-design regularization (Hsu et al., 2012; Cheng & Montanari, 2024), and identify favorable random-feature regimes with optimally tuned regularization (Simon et al., 2024). Double descent and nonmonotonicity as model capacity or spectral structure varies (Belkin et al., 2019a; Hastie et al., 2022; Li & Sonthalia, 2025) are distinct from deterioration along one fixed training problem. Our construction holds the population loss fixed to isolate how curvature is distributed among individual samples. It complements the structured regimes of benign overfitting and implicit regularization surveyed by Bartlett et al. (2021). The common contracting energy gives a sufficient condition for safe continuation, including examples with noncommuting sample and population Hessians. It does not characterize every safe sampling law.

Risk certification and estimability. Bartlett & Long (2021) establish limitations for bounds determined by the least-norm predictor and sample size under bounded antimonotonicity in sample size. Gastpar et al. (2024) study full-sample access and use mixture indistinguishability. Gastpar et al. (2025) further connect estimability to stability and characterize average squared estimation error through conditional variance. Our task fixes the gradient-selected endpoint and requires a uniformly valid one-sided certificate that adapts on zero-risk runs. Theorem 5.1 determines its sharp resolution with separate training and validation budgets, rather than absolute estimation error at every positive risk. Empirical Bernstein bounds yield an $O(1/m)$ validation certificate at zero loss and fixed confidence (Maurer & Pontil, 2009). Our lower bound establishes the matching joint scale

$\Theta(\min\{\log n/n, 1/m\})$. The future-output guarantee propagates a single coverage event through observable distances, without acquiring further population information.

C PERSISTENT COUNTEREXAMPLES AND ELEMENTARY ESTIMATES

We prove nonexpansiveness and convergence for the methods in Sections 2 and 4, then quantify how shared constraints and missing coordinate penalties produce persistent error. These counterexamples concern prescribed gradient methods. The lower bounds for every interpolating rule are proved separately in Appendix L.

Notation across proofs. The empirical Hessian is H_S throughout. Some dynamics proofs use the explicitly rescaled matrix $A = 3H_S$. Proofs indexed from $\boldsymbol{w}_0$ count completed updates: their indices $0, \dots, T-1$ correspond to $1, \dots, T$ in the main text. Both averages contain T points and $T-1$ preceding updates. For constructions in target-zero error coordinates, Appendix G gives the translation to the zero-initialized main instance. All initializations and distributions are fixed before sampling.

For the statements that allow nonquadratic losses, the larger smooth convex class consists of nonnegative convex 1-smooth sample losses $f(\cdot; z) : \mathbb{R}^d \to \mathbb{R}_+$ with a common zero-loss target $\boldsymbol{w}^*$ and a deterministic initialization satisfying $\|\boldsymbol{w}_1 - \boldsymbol{w}^*\| \le 1$. Both algorithms and the averaging convention are those of Section 2. Translating the target to zero in the quadratic proofs is only a change of analysis coordinates and does not supply the target to the learner. The final subsection derives the diagnostic finite-time lower bound recorded in Appendix N.

C.1 DETAILED STATEMENTS

For the larger smooth convex class, let $\mathcal{R}_{\mathsf{A}}^{\mathrm{smooth}}(n, d, T)$ be the supremum of $\mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}})$ in dimension d, for $\mathsf{A} \in \{\mathrm{GD}, \mathrm{SGD}\}$, and set $\mathcal{R}_{\mathsf{A}}^{\mathrm{smooth}}(n, T) = \sup_{d \ge 1} \mathcal{R}_{\mathsf{A}}^{\mathrm{smooth}}(n, d, T)$. For quadratic losses, write $\mathcal{R}_{\mathsf{A}}(n, d, T) = \mathcal{R}_{\mathsf{A}}(n, d, T; d)$, since $H_z \preceq I$ already implies $\operatorname{tr} H \le d$. These are algorithm-specific worst-case risks, with a supremum over instances for a prescribed method. The long-training $O(1/n)$ conjecture is discussed in Section 4 of Zhang et al. (2025).

Theorem C.1 (A fixed instance remains bad)**.** *For every sufficiently large n, there is a quadratic instance in dimension $n+1$ for which both algorithms satisfy*

$$\inf_{T \ge n^2} \mathbb{E}F(\bar{\boldsymbol{w}}_T) = \Omega(\log n/n). \tag{C.1}$$

There is also an instance in dimension $16 \cdot 2^n + 1$ with

$$\inf_{T \ge 16(n+1)2^n} \mathbb{E}F(\bar{\boldsymbol{w}}_T) \ge 1/288. \tag{C.2}$$

Each distribution and initialization is fixed before sampling and works at every stated horizon. The same lower orders hold at the limit, where every training loss is zero.

The following parameterized bound yields both regimes, with explicit constants.

Theorem C.2 (Parameterized persistent lower bound)**.** *Fix $n \ge 1$, $q \in (0, 1/2]$, and an integer M with $Mq \ge 16$. Set $p = 1 - q^{1/n}$. There is a quadratic instance in dimension $d = M + 1$, independent of the realized data, such that for both algorithms*

$$\mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}}) \ge \frac{p}{144} \quad \text{for every integer } T \ge \frac{16(n+1)}{q}. \tag{C.3}$$

For this same instance,

$$\lim_{T \to \infty} \mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}}) \ge \frac{p}{36}. \tag{C.4}$$

Both algorithms converge almost surely to the same empirical-kernel projection.

Corollary C.3 (Failure in linear dimension)**.** *For every integer $n \ge 33$, there is one instance in dimension $n+1$ such that, for both algorithms,*

$$\inf_{T \ge n^2} \mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}}) \ge \frac{1}{144}\left[1 - \left(\frac{16(n+1)}{n^2}\right)^{1/n}\right] = \Omega\left(\frac{\log n}{n}\right).$$

Consequently there is no absolute constant C such that $\mathbb{E}F(\bar{\boldsymbol{w}}_T) \le C(1/T + 1/n)$ for all instances and all $T \ge n$, for either algorithm.

Corollary C.4 (Constant-risk late regime)**.** *For every $n \ge 1$ and every $T \ge 16(n+1)2^n$,*

$$\frac{1}{288} \le \mathcal{R}_{\mathsf{A}}^{\mathrm{smooth}}(n, T) \le \frac{1}{2}, \qquad \mathsf{A} \in \{\mathrm{GD}, \mathrm{SGD}\}.$$

The lower bound uses one instance in dimension $16 \cdot 2^n + 1$, valid for every such T. Its limiting expected risk is at least $1/72$.

Theorem C.5 (Matching quadratic-class dimension dependence)**.** *For every $n, d \ge 1$ and every integer $T \ge 2nd$,*

$$\mathcal{R}_{\mathsf{A}}(n, d, T) = \Theta\left(\min\left\{1, \frac{1 + \log d}{n}\right\}\right), \qquad \mathsf{A} \in \{\mathrm{GD}, \mathrm{SGD}\},$$

with absolute implicit constants. For all $T \ge 1$, the following explicit upper bound holds:

$$\mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}}) \le \min\left\{\frac{1}{2}, \frac{16}{7T} + \frac{14(1 + \log d)}{3n}\right\}. \tag{C.5}$$

Theorem C.6 (Rank-one least-squares counterexample)**.** *For every integer $M \ge 1024$, put $n = \lfloor M \log M \rfloor$. There is a realizable least-squares instance in dimension $M + 1$, with every sample Hessian having rank one and operator norm one, such that for both algorithms*

$$\inf_{T \ge n^2} \mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}}) \ge \frac{1}{144M} = \Omega\left(\frac{\log n}{n}\right).$$

The limiting expected risk is at least $1/(36M)$.

C.2 The quadratic construction and finite-time GD proof

Fix the parameters of Theorem C.2. Set $\alpha = \sqrt{q}$ and $\boldsymbol{b} = M^{-1/2}\mathbf{1} \in \mathbb{R}^M$. A sample $z \in \{0,1\}^M$ has independent Bernoulli(p) coordinates. For $\boldsymbol{w} = (x, \boldsymbol{y}) \in \mathbb{R} \times \mathbb{R}^M$, define

$$f((x, \boldsymbol{y}); z) = \frac{1}{6}\left[(\alpha x - \boldsymbol{b}^\top \boldsymbol{y})^2 + \sum_{j=1}^{M} z_j y_j^2\right]. \tag{C.6}$$

Take $\boldsymbol{w}^* = 0$ and $\boldsymbol{w}_1 = (1, 0)$. The distribution and all coefficients are fixed before drawing S.

C.2.1 Verification of the hypotheses

With $\boldsymbol{v} = (\alpha, -\boldsymbol{b})$, the Hessian is

$$H_z = \nabla^2 f(\cdot; z) = \frac{1}{3}\left(\boldsymbol{v}\boldsymbol{v}^\top + \mathrm{diag}(0, z)\right).$$

Thus $H_z \succeq 0$ and $\|H_z\|_{\mathrm{op}} \le (\|\boldsymbol{v}\|^2 + 1)/3 = (q+2)/3 \le 1$. The losses are globally smooth, convex, nonnegative quadratics, and vanish at zero. The initialization distance is exactly one. The population risk is

$$F(x, \boldsymbol{y}) = \frac{1}{6}\left[(\alpha x - \boldsymbol{b}^\top \boldsymbol{y})^2 + p\|\boldsymbol{y}\|^2\right]. \tag{C.7}$$

In particular, $F(\boldsymbol{w}_1) = q/6$.

For a fixed dataset define

$$a_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij}, \quad D_S = \mathrm{diag}(a_1, \ldots, a_M), \quad U = \{j : a_j = 0\}, \quad V = [M] \setminus U, \quad Q = \frac{|U|}{M}.$$

The coordinates in U lack the *diagonal penalty* in the training set but remain present in the shared linear term. For $j \in V$, $a_j \ge 1/n$.

Lemma C.7 (Unseen-coordinate event)**.** *The event $G = \{q/2 \le Q \le 2q\}$ has probability at least $3/4$.*

Proof. For each coordinate, $\mathbb{P}(a_j = 0) = (1-p)^n = q$. Independence across coordinates gives $MQ \sim \mathrm{Bin}(M, q)$. Consequently,

$$\mathbb{P}(G^c) \le \mathbb{P}(|Q - q| > q/2) \le \frac{4\,\mathrm{Var}(Q)}{q^2} \le \frac{4}{Mq} \le \frac{1}{4}. \qquad \square$$

C.2.2 A POSITIVE INVARIANT AND AN EXACT IDENTITY

Write $r_t = \alpha x_t - \boldsymbol{b}^\top \boldsymbol{y}_t$. GD with step size $1/4$ gives

$$x_{t+1} = x_t - \frac{\alpha}{12} r_t, \tag{C.8}$$

$$y_{j,t+1} = (1 - a_j/12) y_{j,t} + \frac{b_j}{12} r_t. \tag{C.9}$$

Therefore

$$r_{t+1} = \left(1 - \frac{q+1}{12}\right) r_t + \frac{1}{12} \boldsymbol{b}^\top D_S \boldsymbol{y}_t. \tag{C.10}$$

Since $r_1 = \alpha$ and $\boldsymbol{y}_1 = 0$, induction proves $r_t \geq 0$ and $\boldsymbol{y}_t \geq 0$ componentwise. Hence x_t is nonincreasing, and $r_t = \alpha x_t - \boldsymbol{b}^\top \boldsymbol{y}_t \geq 0$ implies $x_t \geq 0$.

For $j \in U$, comparing (C.8) and (C.9) and telescoping yields the exact identity

$$\boldsymbol{y}_{U,t} = \frac{\boldsymbol{b}_U}{\alpha}(1 - x_t). \tag{C.11}$$

Since $\|\boldsymbol{b}_U\|^2 = Q$, we obtain

$$r_t + \boldsymbol{b}_V^\top \boldsymbol{y}_{V,t} = \frac{(q+Q)x_t - Q}{\alpha}. \tag{C.12}$$

In particular, $x_t \geq x_\infty := Q/(q+Q)$.

C.2.3 THE AVERAGED ITERATE, NOT JUST THE ENDPOINT

Let bars denote averages over $t = 1, \ldots, T$. Telescoping the x update gives

$$\bar{r}_T = \frac{12(1 - x_{T+1})}{\alpha T}. \tag{C.13}$$

Telescoping the jth coordinate gives, for $j \in V$,

$$a_j \bar{y}_{j,T} = b_j \bar{r}_T - \frac{12}{T} y_{j,T+1} \leq b_j \bar{r}_T.$$

Define $B_S = \sum_{j \in V} b_j^2/a_j \leq n$. Then $\boldsymbol{b}_V^\top \bar{\boldsymbol{y}}_{V,T} \leq B_S \bar{r}_T$. Combining this with (C.12) and (C.13), and using $1 - x_{T+1} \leq q/(q+Q)$, gives $\bar{x}_T - x_\infty \leq 12(1+B_S)q/[T(q+Q)^2]$. Equivalently,

$$1 - \bar{x}_T \geq \frac{q}{q+Q}\left(1 - \frac{12(1+B_S)}{T(q+Q)}\right). \tag{C.14}$$

On G, if $T \geq 16(n+1)/q$, the parenthetical factor is at least $1/2$. Averaging (C.11) therefore gives

$$\|\bar{\boldsymbol{y}}_{U,T}\|^2 = \frac{Q}{q}(1 - \bar{x}_T)^2 \geq \frac{qQ}{4(q+Q)^2} \geq \frac{1}{18}. \tag{C.15}$$

The last inequality follows from $s/(1+s)^2 \geq 2/9$ for $s \in [1/2, 2]$. Equation (C.7) now gives $F(\bar{\boldsymbol{w}}_T^{\mathrm{GD}}) \geq p/108$ on G. Lemma C.7 proves (C.3) for GD.

C.3 FIXED-DATASET SGD AND THE EXACT LONG-TIME LIMIT

C.3.1 A CONDITIONAL MEAN IDENTITY

For a quadratic instance in the original coordinates, the GD recursion gives

$$\bar{\boldsymbol{w}}_T^{\mathrm{GD}} - \boldsymbol{w}^* = g_T(H_S)(\boldsymbol{w}_1 - \boldsymbol{w}^*), \qquad g_T(x) = \frac{1}{T}\sum_{t=0}^{T-1}(1 - x/4)^t. \tag{C.16}$$

Indeed, $\boldsymbol{w}_t^{\mathrm{GD}} - \boldsymbol{w}^* = (I - H_S/4)^{t-1}(\boldsymbol{w}_1 - \boldsymbol{w}^*)$, and averaging yields the spectral filter. The following identity relates this GD average to fixed-dataset SGD.

Lemma C.8 (Quadratic GD–SGD comparison)**.** *For any quadratic instance with a common minimizer, conditional on S and with identical initialization,*

$$\mathbb{E}_{\mathrm{alg}}[\boldsymbol{w}_t^{\mathrm{SGD}} \mid S] = \boldsymbol{w}_t^{\mathrm{GD}}, \qquad \mathbb{E}_{\mathrm{alg}}[\bar{\boldsymbol{w}}_T^{\mathrm{SGD}} \mid S] = \bar{\boldsymbol{w}}_T^{\mathrm{GD}}.$$

Consequently $\mathbb{E}_{\mathrm{alg}}[F(\bar{\boldsymbol{w}}_T^{\mathrm{SGD}}) \mid S] \geq F(\bar{\boldsymbol{w}}_T^{\mathrm{GD}})$.

Proof. Translate the common minimizer to zero, so that $f_i(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top H_i \boldsymbol{w}$ and $H_S = n^{-1}\sum_i H_i$. Conditional on S and the past, the index is uniform and independent. Thus

$$\mathbb{E}_{\mathrm{alg}}[\boldsymbol{w}_{t+1}^{\mathrm{SGD}} \mid S] = (I - \tfrac{1}{4}H_S)\mathbb{E}_{\mathrm{alg}}[\boldsymbol{w}_t^{\mathrm{SGD}} \mid S].$$

Induction proves the iterate identity, and linearity proves the averaged identity. Conditional Jensen's inequality for the convex population risk gives the claimed lower comparison. No replacement of H_S by the population Hessian is made. □

This proves the finite-time SGD bound in Theorem C.2.

C.3.2 The common limiting projection

Lemma C.9 (Convergence for fixed-dataset quadratic SGD)**.** *In target-zero coordinates, suppose $0 \preceq H_i \preceq I$ and let $K = \bigcap_i \ker H_i = \ker H_S$. GD and fixed-dataset SGD with step size $1/4$ both converge almost surely to $P_K \boldsymbol{w}_1$. Their averaged iterates have the same limit.*

Proof. Let $\mathcal{F}_t = \sigma(S, i_1, \ldots, i_{t-1})$ be the information before drawing i_t, so $\boldsymbol{w}_t$ is $\mathcal{F}_t$-measurable. GD convergence follows by diagonalizing $I - H_S/4$. Each SGD update fixes K pointwise and maps $K^\perp$ into itself, so $P_K \boldsymbol{w}_t = P_K \boldsymbol{w}_1$. Set $\boldsymbol{e}_t = \boldsymbol{w}_t - P_K \boldsymbol{w}_1 \in K^\perp$. If $K^\perp = \{0\}$, convergence is immediate. Otherwise let $\lambda_S > 0$ be the smallest eigenvalue of H_S restricted to $K^\perp$. Since $H_i^2 \preceq H_i$,

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{e}_{t+1}\|^2 \mid S, \mathcal{F}_t] &\leq \|\boldsymbol{e}_t\|^2 - \frac{7}{16}\boldsymbol{e}_t^\top H_S \boldsymbol{e}_t \\ &\leq (1 - 7\lambda_S/16)\|\boldsymbol{e}_t\|^2.\end{aligned}$$

Conditional geometric mean-square convergence follows. Markov's inequality and the first Borel–Cantelli lemma give conditional almost-sure convergence. Cesàro averaging preserves convergence. The positive λ_S is used only for convergence on each fixed finite dataset, not as any uniform conditioning assumption or finite-time rate. □

For (C.6), the empirical kernel is

$$K = \{(x, \boldsymbol{y}) : \boldsymbol{y}_V = 0,\ \alpha x = \boldsymbol{b}_U^\top \boldsymbol{y}_U\}.$$

Orthogonal projection of $(1, 0)$ onto this subspace is

$$x_\infty = \frac{Q}{q+Q}, \qquad \boldsymbol{y}_{\infty,U} = \frac{\alpha}{q+Q}\boldsymbol{b}_U, \qquad \boldsymbol{y}_{\infty,V} = 0. \tag{C.17}$$

Indeed, this vector belongs to K, and its difference from $(1, 0)$ is orthogonal to K. The formula also covers $Q = 0$. Its risk is exactly

$$F(\boldsymbol{w}_\infty) = \frac{p}{6}\frac{qQ}{(q+Q)^2}. \tag{C.18}$$

On G this is at least $p/27$, so its expectation is at least $p/36$. All iterates have norm at most one, and $F(\boldsymbol{w}) \leq \|\boldsymbol{w}\|^2/2$. Dominated convergence therefore justifies taking expectations in the limit, proving (C.4) for both algorithms.

C.4 Consequences in the original model

C.4.1 Polynomial and exponential horizons

For Corollary C.3, choose $q = 16(n+1)/n^2$ and $M = n$. Then $q \le 1/2$ for $n \ge 33$, $Mq \ge 16$, and the finite-time threshold is exactly n^2. Let $a_n = \log[n^2/(16(n+1))]$. Since $a_n \sim \log n$ and $a_n/n \to 0$,

$$p = 1 - e^{-a_n/n} \sim \frac{\log n}{n}.$$

The same instance works for every later horizon, including n^{10} and the limit. Its initialization risk is $q/6 = \Theta(1/n)$, so the deterioration is genuine even relative to the initial predictor.

For Corollary C.4, choose $q = 2^{-n}$, $p = 1/2$, and $M = 16 \cdot 2^n$. The lower bounds follow from Theorem C.2. Here even the initialization risk is exponentially small, namely $2^{-n}/6$.

More generally, for any integer $T_0 \ge 32(n+1)$, choose

$$q = \frac{16(n+1)}{T_0}, \qquad M = \lceil 16/q \rceil.$$

Then the same instance has dimension $\lceil T_0/(n+1) \rceil + 1$ and satisfies

$$\inf_{T \ge T_0} \mathbb{E} F(\bar{\boldsymbol{w}}_T^{\mathsf{A}}) \ge \frac{1}{144} \left[1 - \left(\frac{16(n+1)}{T_0} \right)^{1/n} \right]. \tag{C.19}$$

This gives a logarithmic-in-training lower bound before its constant saturation. A matching general nonlinear upper bound at every intermediate horizon is *not* claimed.

C.4.2 The universal constant upper bound

Lemma C.10 (Pathwise radius and risk bound)**.** *Under the smooth convex assumptions specified at the start of this appendix, both algorithms satisfy $\|\boldsymbol{w}_t - \boldsymbol{w}^*\| \le 1$ and $F(\bar{\boldsymbol{w}}_T) \le 1/2$ pathwise.*

Proof. For a nonnegative 1-smooth convex function g with $g(\boldsymbol{w}^*) = 0$, smoothness at $\boldsymbol{w} - \nabla g(\boldsymbol{w})$ gives $\|\nabla g(\boldsymbol{w})\|^2 \le 2g(\boldsymbol{w})$, while convexity gives $\langle \nabla g(\boldsymbol{w}), \boldsymbol{w} - \boldsymbol{w}^* \rangle \ge g(\boldsymbol{w})$. Consequently, with $\eta = 1/4$,

$$\|\boldsymbol{w} - \eta \nabla g(\boldsymbol{w}) - \boldsymbol{w}^*\|^2 \le \|\boldsymbol{w} - \boldsymbol{w}^*\|^2 - 2\eta(1-\eta) g(\boldsymbol{w}) \le \|\boldsymbol{w} - \boldsymbol{w}^*\|^2.$$

Apply this to $g = F_S$ for GD and $g = f(\cdot; z_{i_t})$ for each SGD update. Averaging preserves the unit ball. Since $\nabla f(\boldsymbol{w}^*; z) = 0$, smoothness also gives $f(\boldsymbol{w}; z) \le \|\boldsymbol{w} - \boldsymbol{w}^*\|^2/2$. Taking a population expectation completes the proof. □

This establishes the matching upper bound in Corollary C.4 without a quadratic restriction.

C.5 A matching theorem within the quadratic class

Translate $\boldsymbol{w}^*$ to zero. A nonnegative quadratic with zero minimum has the form $f(\boldsymbol{w}; z) = \frac{1}{2} \boldsymbol{w}^\top H_z \boldsymbol{w}$, where $0 \preceq H_z \preceq I$. Set $H = \mathbb{E} H_z$.

C.5.1 Empirical optimization error

For an update using H_i, the identity $H_i^2 \preceq H_i$ gives

$$\|\boldsymbol{w} - H_i \boldsymbol{w}/4\|^2 \le \|\boldsymbol{w}\|^2 - \frac{7}{16} \boldsymbol{w}^\top H_i \boldsymbol{w} = \|\boldsymbol{w}\|^2 - \frac{7}{8} f_i(\boldsymbol{w}).$$

The same inequality holds with H_i replaced by H_S for GD. Conditional expectation for SGD, telescoping, and convexity therefore yield

$$\mathbb{E}_{\mathrm{alg}}[F_S(\bar{\boldsymbol{w}}_T) \mid S] \le \frac{8}{7T}. \tag{C.20}$$

C.5.2 A RELATIVE MATRIX DEVIATION BOUND

We use the standard self-adjoint matrix Bernstein inequality (Tropp, 2012, Theorem 1.4): if X_i are independent mean-zero self-adjoint $d \times d$ matrices with $\lambda_{\max}(X_i) \le R$ and $\|\sum_i \mathbb{E}X_i^2\| \le \sigma^2$, then

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_i X_i\right) \ge t\right) \le d \exp\left(-\frac{t^2}{2(\sigma^2 + Rt/3)}\right).$$

We apply this inequality after rescaling by the population curvature.

Lemma C.11. *Let* $Z_S = \max\{0, \lambda_{\max}(H - 2H_S)\}$. *Then* $0 \le Z_S \le 1$ *and*

$$\mathbb{E}Z_S \le \min\left\{1, \frac{28(1+\log d)}{3n}\right\}.$$

Proof. Fix $u \in (0, 1]$ and set $B = H + uI$, $Y_i = B^{-1/2} H_i B^{-1/2}$, and $A = \mathbb{E}Y_i = B^{-1/2} H B^{-1/2}$. Write $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$. Then $0 \preceq A \preceq I$ and $0 \preceq Y_i \preceq I/u$. For $X_i = A - Y_i$,

$$\mathbb{E}X_i = 0, \quad \|X_i\| \le 1/u, \quad \mathbb{E}X_i^2 = \mathbb{E}Y_i^2 - A^2 \preceq A/u \preceq I/u.$$

Matrix Bernstein with $R = 1/u$, $\sigma^2 \le n/u$, and $t = n/2$ yields

$$\mathbb{P}\left(\lambda_{\max}(A - \bar{Y}) > 1/2\right) \le d e^{-3nu/28}.$$

On the complementary event, $2\bar{Y} \succeq 2A - I$. Congruence by $B^{1/2}$ gives $2H_S \succeq H - uI$, so $\mathbb{P}(Z_S > u) \le d e^{-3nu/28}$. Since $H \preceq I$ and $H_S \succeq 0$, $Z_S \le 1$. Integrating the tail bound, and splitting at $28\log d/(3n)$ when this lies in $[0, 1]$, proves the assertion. □

For every $\boldsymbol{w}$ of norm at most one, even when $\boldsymbol{w}$ depends on S,

$$F(\boldsymbol{w}) = \tfrac{1}{2}\boldsymbol{w}^\top H \boldsymbol{w} \le \boldsymbol{w}^\top H_S \boldsymbol{w} + Z_S/2 = 2F_S(\boldsymbol{w}) + Z_S/2.$$

Combining Lemmas C.10 and C.11 with (C.20) proves (C.5).

C.5.3 MATCHING THE DIMENSION DEPENDENCE

For $d \ge 33$, set $M = d - 1$ and $q = 16/M$. Theorem C.2 applies whenever $T \ge (n+1)(d-1)$, in particular when $T \ge 2nd$. It gives

$$\mathcal{R}_{\mathsf{A}}(n, d, T) \ge \frac{1}{144}\left[1 - \exp\left(-\frac{\log((d-1)/16)}{n}\right)\right].$$

For $d \ge 33$, $\log((d-1)/16) \ge (1 + \log d)/8$. Also $1 - e^{-x} \ge (1 - e^{-1})\min\{1, x\}$ for $x \ge 0$. These inequalities give the desired lower bound.

For $1 \le d \le 32$, embed the one-dimensional loss $f(w; z) = zw^2/2$, where $z \sim \mathrm{Bernoulli}(1/(n+1))$ and $w_1 = 1$. If every training sample is zero, both algorithms stay at one forever. This event has probability $(n/(n+1))^n \ge e^{-1}$, so the expected risk is at least $1/[2e(n+1)] \ge 1/(4en)$. Since $1 + \log d$ is bounded in this range, the same rate follows. This completes Theorem C.5.

C.6 A RANK-ONE LEAST-SQUARES VERSION

This section proves Theorem C.6 independently of the high-rank sampling construction. Fix $M \ge 1024$ and $n = \lfloor M \log M \rfloor$. Set $q = (1 - 1/(2M))^n$ and $\alpha = \sqrt{q}$, and define $\boldsymbol{b} = M^{-1/2}\mathbf{1}$ and $\boldsymbol{v} = (\alpha, -\boldsymbol{b})$. Draw a feature $\boldsymbol{a} \in \mathbb{R}^{M+1}$ by

$$\boldsymbol{a} = \frac{\boldsymbol{v}}{\sqrt{1+q}} \quad \text{with probability } \frac{1}{2}, \qquad \boldsymbol{a} = (0, \boldsymbol{e}_j) \quad \text{with probability } \frac{1}{2M}, \quad j \in [M].$$

Let $f(\boldsymbol{w}; \boldsymbol{a}) = \frac{1}{2}(\boldsymbol{a}^\top \boldsymbol{w})^2$, $\boldsymbol{w}^* = 0$, and $\boldsymbol{w}_1 = (1, 0)$. Every feature has norm one, so every sample Hessian is rank one with operator norm one. The population risk is

$$F(x, \boldsymbol{y}) = \frac{1}{4}\left[\frac{(\alpha x - \boldsymbol{b}^\top \boldsymbol{y})^2}{1+q} + \frac{\|\boldsymbol{y}\|^2}{M}\right]. \tag{C.21}$$

Let N_0 count occurrences of the shared feature and N_j count occurrences of coordinate feature j. Write $c = N_0/[n(1+q)]$ and $d_j = N_j/n$. Let $U = \{j : N_j = 0\}$ be the unseen coordinates and $Q = |U|/M$ their fraction. Then $F_S(x, \boldsymbol{y}) = \frac{1}{2}[c(\alpha x - \boldsymbol{b}^\top \boldsymbol{y})^2 + \sum_j d_j y_j^2]$.

C.6.1 THE OCCUPANCY EVENT

For distinct j, k, the unseen indicators satisfy

$$\mathbb{P}(j \in U) = q, \qquad \mathbb{P}(j, k \in U) = (1 - 1/M)^n \le q^2.$$

They are therefore pairwise negatively correlated, and $\mathrm{Var}(Q) \le q/M$. If $Mq \ge 16$, Chebyshev gives $\mathbb{P}(q/2 \le Q \le 2q) \ge 3/4$. Since $\mathbb{P}(N_0 = 0) = 2^{-n}$ and $n \ge 2$, the event $G' = \{q/2 \le Q \le 2q,\ N_0 > 0\}$ has probability at least $1/2$.

C.6.2 FINITE-TIME AVERAGES

Conditional on $N_0 > 0$, put $r_t = \alpha x_t - \boldsymbol{b}^\top \boldsymbol{y}_t$ and $\eta = 1/4$. The GD recurrence is

$$x_{t+1} = x_t - \eta c \alpha r_t, \qquad \boldsymbol{y}_{t+1} = (I - \eta D)\boldsymbol{y}_t + \eta c \boldsymbol{b} r_t,$$

where $D = \mathrm{diag}(d_j)$. Moreover,

$$r_{t+1} = (1 - \eta c(1+q)) r_t + \eta \boldsymbol{b}^\top D \boldsymbol{y}_t.$$

Here $c(1+q) = N_0/n \le 1$, so the same positive invariant holds: $r_t \ge 0$, $\boldsymbol{y}_t \ge 0$, and x_t is nonincreasing. The identities

$$\boldsymbol{y}_{U,t} = \frac{\boldsymbol{b}_U}{\alpha}(1 - x_t), \qquad r_t + \boldsymbol{b}_{U^c}^\top \boldsymbol{y}_{U^c,t} = \frac{(q+Q)x_t - Q}{\alpha}$$

follow exactly as before. Thus $x_t \ge Q/(q+Q)$.

Telescoping now gives

$$\bar{r}_T = \frac{1 - x_{T+1}}{\eta c \alpha T}, \qquad d_j \bar{y}_{j,T} = c b_j \bar{r}_T - \frac{y_{j,T+1}}{\eta T} \le c b_j \bar{r}_T.$$

Set $B = \sum_{j \notin U} b_j^2/d_j \le n$. Combining these identities proves

$$1 - \bar{x}_T \ge \frac{q}{q+Q}\left(1 - \frac{1/c + B}{\eta T (q+Q)}\right). \tag{C.22}$$

If $q \le 1/2$ and $N_0 > 0$, then $1/c + B \le n(1+q) + n \le 5n/2$. On G', for $T \ge 16(n+1)/q$, the parenthetical factor in (C.22) is at least $1/2$. Consequently $\|\bar{\boldsymbol{y}}_{U,T}\|^2 \ge 1/18$. Equation (C.21) gives $F(\bar{\boldsymbol{w}}_T^{\mathrm{GD}}) \ge 1/(72M)$ on G', so $\mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathrm{GD}}) \ge 1/(144M)$. Lemma C.8 transfers this lower bound to fixed-dataset SGD.

C.6.3 CHECKING THE PARAMETER REGIME AND THE LIMIT

For $M \ge 1024$, elementary logarithmic inequalities give

$$\frac{1}{2\sqrt{M}} \le q \le \frac{2}{\sqrt{M}} \le \frac{1}{2}.$$

For the lower bound, use $\log(1-u) \ge -u/(1-u)$ and $n \le M \log M$ to obtain $q \ge M^{-M/(2M-1)} \ge 1/(2\sqrt{M})$. For the upper bound use $n \ge M \log M - 1$ and $\log(1-u) \le -u$. Thus $Mq \ge \sqrt{M}/2 \ge 16$. Also $n \ge M \log M/2$, so $nq \ge \sqrt{M} \log M/4 \ge 32$. It follows that $n^2 \ge 16(n+1)/q$, and the finite-time result holds for every $T \ge n^2$. Finally, $1/M \sim \log n/n$.

On G', the common empirical kernel has the same form as in (C.17). Its projection has zero shared residual and squared unseen-coordinate norm

$$\frac{qQ}{(q+Q)^2} \ge \frac{2}{9}.$$

Hence its population risk is at least $1/(18M)$ on G', giving an expectation of at least $1/(36M)$. Lemma C.9 and dominated convergence complete the limiting claim.

Remark C.1 (Zero initialization). Nonzero initialization is not essential. Translate the parameter as $\widetilde{\boldsymbol{w}} = \boldsymbol{w} - \boldsymbol{w}_1$ and set $\widetilde{\boldsymbol{w}}^* = -\boldsymbol{w}_1$. The initialization is then zero, the target has norm one, and the losses become ordinary noiseless least-squares losses with labels $-\boldsymbol{a}^\top \boldsymbol{w}_1$. All risks and trajectories are unchanged by this translation.

C.7 Gradient energy and the unrestricted finite-time gap

What forces the failure. The shared residual drives mass into coordinates that have no sampled diagonal penalty. Calibrating $\alpha^2 = q$ makes the squared coupling strength comparable to the total shared-feature weight on the unseen coordinates, $Q = \Theta(q)$. Thus the nearest empirical interpolator to the initialization places a constant amount of squared norm in directions with positive population curvature. No nonlinear switching, exploding iterate, adversarial index schedule, or projection step is needed.

Why finite gradient energy does not prevent it. For the GD construction, smoothness of F_S gives

$$F_S(\boldsymbol{w}_{t+1}) \le F_S(\boldsymbol{w}_t) - \frac{7}{32}\|\nabla F_S(\boldsymbol{w}_t)\|^2.$$

Since $F_S(\boldsymbol{w}_1) = q/6$, summing gives $\sum_{t\ge 1} \|\nabla F_S(\boldsymbol{w}_t)\|^2 \le 16q/21$. The total gradient energy can therefore be exponentially small while the limiting population risk is constant. Square-summability does not control the total accumulated displacement uniformly in time. Thus the closest empirical common minimizer need not have dimension-free $O(1/n)$ population risk.

The diagnostic intermediate-time gap. For sufficiently large n, choose $q = 32/\log n$, $M = \lceil 16/q \rceil$, and $p = 1 - q^{1/n}$ in Theorem C.2. Then $q \le 1/2$, $Mq \ge 16$, and $M + 1 \le n$, so zero-padding gives dimension n. The population trace is $(1 + q + Mp)/3 \le 1$ for all sufficiently large n, because $Mp = O(\log n \log\log n/n) = o(1)$. The theorem's threshold satisfies $16(n+1)/q = (n+1)\log n/2 \le \lceil n \log n \rceil$. Its lower bound $p/144 = \Omega(\log\log n/n)$ therefore applies to both GD and fixed-data SGD at the horizon in (N.1). The upper estimate follows directly from the relative-covariance and empirical-energy bounds in Appendix D, which give $O(1/T + \log(1+n\tau)/n)$ for every T. This proves the displayed band without claiming that either endpoint is the sharp unrestricted finite-time rate.

D Sharp long-run resource laws

This appendix proves parts (ii)–(iii) of Theorem 3.2 and Corollary 3.3: the risk of the prescribed converged interpolator and of all averages in the stated late interval. We combine empirical optimization with a bound on population curvature missed by the sample, then calibrate the dimension, trace, and sample-Hessian rank budgets. Appendix L treats the separate infimum over interpolating rules. Its upper bound uses the curvature comparison proved here and does not assume the main statistical theorem.

D.1 Preliminary identities

A common minimizer is translated to zero, so a sample loss is $f(\boldsymbol{w}; z) = \frac{1}{2}\boldsymbol{w}^\top H_z \boldsymbol{w}$ with $0 \preceq H_z \preceq I$. Write $H = \mathbb{E}H_z$, $H_S = n^{-1}\sum_i H_i$, and $F(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top H \boldsymbol{w}$. Initialization is deterministic and has norm at most one. The step size is $1/4$. The output is $\bar{\boldsymbol{w}}_T = T^{-1}\sum_{t=1}^T \boldsymbol{w}_t$, with $\boldsymbol{w}_1$ denoting initialization. SGD samples independent uniform indices from the same fixed dataset. All unconditional expectations include both dataset and algorithm randomness.

Appendix C shows that neither method increases the distance to the common minimizer, that both converge to $P_{\ker H_S}\boldsymbol{w}_1$, and that $\mathbb{E}_{\mathrm{alg}}[\bar{\boldsymbol{w}}_T^{\mathrm{SGD}} \mid S] = \bar{\boldsymbol{w}}_T^{\mathrm{GD}}$. It also proves the empirical optimization bound

$$\mathbb{E}_{\mathrm{alg}}[F_S(\bar{\boldsymbol{w}}_T) \mid S] \le \frac{8}{7T} \tag{D.1}$$

and the matrix comparison

$$\mathbb{E}\lambda_{\max}^+(H - 2H_S) \le \frac{28(1+\log d)}{3n}. \tag{D.2}$$

For a symmetric matrix, $\lambda_{\max}^+(B) = \max(0, \lambda_{\max}(B))$. The one-dimensional rare-sample lower bound is proved in Appendix C.

We will also use the following scaled version of the trajectory identity. Let $\boldsymbol{b} = M^{-1/2}\mathbf{1}$, $\alpha = \sqrt{q}$, $\boldsymbol{v} = (\alpha, -\boldsymbol{b})$, and consider a dataset objective

$$F_S(x, \boldsymbol{y}) = \frac{1}{2}\left[c(\alpha x - \boldsymbol{b}^\top \boldsymbol{y})^2 + \sum_j a_j y_j^2\right], \qquad c > 0.$$

Let $\boldsymbol{e}_0, \dots, \boldsymbol{e}_M$ denote the coordinate vectors of $\mathbb{R}^{M+1}$, with $\boldsymbol{e}_0 = (1, \mathbf{0})$. For GD initialized at $\boldsymbol{e}_0$, let $U = \{j : a_j = 0\}$, $Q = |U|/M$, and $B = \sum_{j \notin U} b_j^2/a_j$. Whenever the update has the nonnegative coefficients used below, the telescoping calculation in Appendix C gives

$$\boldsymbol{y}_{U,t} = \frac{\boldsymbol{b}_U}{\alpha}(1 - x_t), \qquad 1 - \bar{x}_T \geq \frac{q}{q+Q}\left[1 - \frac{1/c + B}{(1/4)T(q+Q)}\right]. \tag{D.3}$$

For the standard star loss $[(\alpha x - \boldsymbol{b}^\top \boldsymbol{y})^2 + \sum z_j y_j^2]/6$, we use unscaled frequencies a_j. The numerator in the bracket then becomes $12(1 + B)$.

D.2 A TRACE-SENSITIVE RELATIVE COVARIANCE COMPARISON

Theorem D.1. *Let $H_1, \dots, H_n$ be independent identically distributed PSD contractions on $\mathbb{R}^d$, let $H = \mathbb{E}H_i$, and put $\tau = \operatorname{tr} H$. Then*

$$\mathbb{E}\lambda_{\max}^+(H - 2H_S) \leq \min\left\{1, \frac{28(1 + \log d)}{3n}, \frac{9}{n}\log(1 + 2n\tau)\right\}. \tag{D.4}$$

No almost-sure bound on $\operatorname{tr} H_i$ or on $\operatorname{rank} H_i$ is assumed.

Proof. The first bound follows from $H \preceq I$, and the second is Lemma C.11. We prove the trace bound.

Spectral rounding. Conditional on $H_i = \sum_j \lambda_{ij} \boldsymbol{u}_{ij}\boldsymbol{u}_{ij}^\top$, independently round each eigenvalue to a Bernoulli variable of mean λ_{ij}. The resulting matrix P_i is an orthogonal projection with $\mathbb{E}[P_i \mid H_i] = H_i$ and $\mathbb{E}\operatorname{rank} P_i = \tau$. Rounding is independent across samples. Convexity of $A \mapsto \lambda_{\max}^+(H - 2A)$ shows that replacing H_i by P_i can only increase the expected quantity being bounded. Thus it suffices to analyze independent projections with mean H.

Ghost sample and conditional support. Introduce an independent projection sample P_i', and write $\bar{P} = n^{-1}\sum_i P_i$ and $\bar{P}' = n^{-1}\sum_i P_i'$. Jensen gives

$$\mathbb{E}\lambda_{\max}^+(H - 2\bar{P}) \leq \mathbb{E}\lambda_{\max}^+(\bar{P}' - 2\bar{P}).$$

Condition on the unordered pairs $\{P_i, P_i'\}$, and randomize their orientations by independent signs ϵ_i, each uniform on $\{-1, 1\}$. With $C_i = P_i + P_i'$ and $D_i = P_i' - P_i$, the difference has the conditional law

$$\bar{P}' - 2\bar{P} \overset{\text{law}}{=} \frac{3\sum_i \epsilon_i D_i - \sum_i C_i}{2n}.$$

Every matrix on the right acts on the fixed conditional subspace $L = \sum_i(\operatorname{Ran} P_i + \operatorname{Ran} P_i')$. Its dimension k satisfies $k \leq \sum_i(\operatorname{rank} P_i + \operatorname{rank} P_i')$ and $\mathbb{E}k \leq 2n\tau$. If $k = 0$, the difference vanishes.

Offset Laplace bound on this subspace. For independent Hermitian random matrices X_i on L and a deterministic Hermitian offset B, the standard matrix Laplace-transform inequality, a consequence of Lieb's trace concavity (Tropp, 2012), gives

$$\mathbb{E}\operatorname{tr}\exp\left(B + \sum_i X_i\right) \leq \operatorname{tr}\exp\left(B + \sum_i \log \mathbb{E}e^{X_i}\right).$$

Here all traces are on L. Since $D_i^2 \preceq 2C_i$ and $\log\cosh(3\theta D_i) \preceq (9\theta^2/2)D_i^2$, at $\theta = 1/9$ this gives

$$\mathbb{E}_\epsilon \operatorname{tr}\exp\left(\theta\left[3\sum_i \epsilon_i D_i - \sum_i C_i\right]\right) \leq \operatorname{tr}\exp\left(-\theta\sum_i C_i + \sum_i \log\cosh(3\theta D_i)\right) \leq k.$$

For completeness, $D_i^2 \preceq 2C_i$ follows from $(A-B)^2 \preceq 2(A^2+B^2)$ and idempotence of the two projections. The last trace bound uses the PSD order of the exponent, without asserting operator monotonicity of the exponential. The conditional positive-eigenvalue tail is consequently at most $\min\{1, ke^{-2nu/9}\}$ at level $u > 0$. Integrating gives $\mathbb{E}_\epsilon \lambda_{\max}^+(\bar{P}' - 2\bar{P}) \le 9(1+\log k)/(2n)$ for $k \ge 1$. Since $1 + \log k \le 2\log(1+k)$ for integer $k \ge 1$, Jensen finally gives

$$\mathbb{E}\lambda_{\max}^+(H - 2H_S) \le \frac{9}{n}\mathbb{E}\log(1+k) \le \frac{9}{n}\log(1+2n\tau).$$

This also handles $\tau = 0$ exactly. □

Corollary D.2. *For both algorithms, every $T \ge 1$, and $\tau = \operatorname{tr} H$,*

$$\mathbb{E}F(\bar{\boldsymbol{w}}_T) \le \min\left\{\frac{\min(1,\tau)}{2}, \frac{16}{7T} + \min\left\{\frac{14(1+\log d)}{3n}, \frac{9}{2n}\log(1+2n\tau)\right\}\right\}. \tag{D.5}$$

Proof. Set $Z = \lambda_{\max}^+(H - 2H_S)$. Simultaneously for every $\|\boldsymbol{w}\| \le 1$, $F(\boldsymbol{w}) \le 2F_S(\boldsymbol{w}) + Z/2$. Thus the inequality applies to the data-dependent output without any independence assertion. Combine Theorem D.1 with (D.1). The other cap uses $\|H\| \le \min(1,\tau)$ and the radius bound. □

D.3 The sharp long-time trace-budget law

Define $\mathcal{R}_{\mathsf{A}}(n,d,T;\tau)$ as the supremum of $\mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}})$ over the quadratic model above with $\operatorname{tr}(\mathbb{E}H_z) \le \tau$, and deterministic initialization of norm at most one.

Theorem D.3. *There are absolute constants $c, C > 0$ such that, for $n, d \ge 1$, $0 < \tau \le d$, $T \ge 32nd$, and $\mathsf{A} \in \{\mathrm{GD}, \mathrm{SGD}\}$,*

$$c\Psi(n,d,\tau) \le \mathcal{R}_{\mathsf{A}}(n,d,T;\tau) \le C\Psi(n,d,\tau). \tag{D.6}$$

Here Ψ is the resource function in (3.3)*. The lower-bound instance for each (n,d,τ) can be held fixed for all such T. The same order holds for the expected risk of the infinite-time limit. At $\tau = 0$, the risk is zero.*

We first record the occupancy refinement that improves the deterministic estimate $B \le n$.

Lemma D.4. *Suppose masks $z_1, \dots, z_n \in \{0,1\}^M$ are independent across examples, each coordinate has marginal probability $p \in (0,1]$, and within each example distinct coordinate indicators have nonpositive covariance. Let $a_j = n^{-1}\sum_{i=1}^n z_{ij}$ be the empirical frequency of coordinate j. Put $q = (1-p)^n$, $Q = M^{-1}|\{j : a_j = 0\}|$, and $B = \sum_{a_j>0} 1/(Ma_j)$. Then*

$$\mathbb{E}Q = q, \qquad \operatorname{Var} Q \le q/M, \qquad \mathbb{E}B \le 2/p.$$

If $Mq = 64$, the event $Q \in [q/2, 2q]$ has probability at least $15/16$, and $B \le 16/p$ has probability at least $7/8$.

Proof. For two coordinates, the probability both are absent from one example is at most $(1-p)^2$. Independence across examples implies nonpositive covariance for their unseen indicators. This proves the variance bound. Each count $K_j = na_j$ is binomial (n,p), and

$$\mathbb{E}\frac{n\mathbf{1}_{K_j>0}}{K_j} \le 2n\mathbb{E}\frac{1}{K_j+1} = \frac{2n[1-(1-p)^{n+1}]}{(n+1)p} \le 2/p.$$

Chebyshev's and Markov's inequalities give the two event probabilities. □

Proof of Theorem D.3. The upper bound follows from Corollary D.2. If $n\tau < 1$, the trace cap is within a constant of the claimed rate. If $n\tau \ge 1$, the claimed rate is at least $(\log 2)/n$, so the $1/T$ term is absorbed.

We first prove the lower bound in the regimes where the construction below yields $M \ge 256$. Put $q = 64/M$, $\ell = \log(M/64)$, and $p = 1 - e^{-\ell/n}$. The selected M ensures $\ell \le n/4$, so $\ell/(2n) \le p \le \ell/n$. The shared direction, $\boldsymbol{b}$, and initialization $\boldsymbol{e}_0$ are as in (D.3).

Case $0 < \tau < 2$. Let $\rho = \min(\tau, 1)$ and $M = \lfloor \min\{d-1, \sqrt{n\rho}/8\} \rfloor$. Here $\ell = \log(M/64) \le M \le n\rho/(64M) \le n/4$, and $p \le \ell/n \le \rho$, using $n\rho \ge 64M^2$. Use the rank-one Hessian $\boldsymbol{v}\boldsymbol{v}^\top/(1+q)$ with probability $\rho/2$, each coordinate projector $\boldsymbol{e}_j\boldsymbol{e}_j^\top$ with probability p, and the zero matrix otherwise. These probabilities are valid because

$$Mp \le \frac{M\ell}{n} \le \frac{M^2}{n} \le \frac{\rho}{64}.$$

Its mean trace is at most $\rho/2 + \rho/64 \le \tau$. Coordinate categories are exclusive, so Lemma D.4 applies. The number N_0 of shared samples exceeds $n\rho/4$ except with probability at most $e^{-n\rho/16}$. Since $M \ge 256$ implies $n\rho \ge 64M^2$, this failure probability is less than $1/16$. On the resulting joint event of probability at least $3/4$, the empirical shared coefficient satisfies $c \ge \rho/8$, giving

$$\frac{1}{c} + B \le \frac{8}{\rho} + \frac{16}{p} \le \frac{24}{p}.$$

For $T \ge 512/(pq)$, the bracket in (D.3) is at least $1/2$. Hence $\|\bar{\boldsymbol{y}}_U\|^2 \ge 1/18$ and the population risk is at least $p/36$ on this event. We do not optimize these constants.

Case $\tau \ge 2$. Let $r = \lfloor \tau \rfloor$, $k = r - 1 \ge \tau/3$, and $M = \lfloor \min\{d-1, \sqrt{nk}/8, e^{n/4}\} \rfloor$. Let $\mu = Mp \le M^2/n \le k/64$. Choose an integer K supported on $\{\lfloor \mu \rfloor, \lceil \mu \rceil\}$ with mean μ, and conditionally choose a uniformly random K-subset of coordinates. Use its indicator mask in the standard star loss

$$f(\boldsymbol{w}; z) = \tfrac{1}{6}[(\boldsymbol{v}^\top \boldsymbol{w})^2 + \sum_j z_j y_j^2].$$

The mask has marginal p and sample size at most k. The sample Hessian has norm at most one, rank at most r, and trace at most $(1+q+k)/3 \le r \le \tau$. Its coordinate indicators have nonpositive covariance: the adjacent-integer distribution minimizes variance among integer-valued variables of mean μ, so

$$\operatorname{Var} K \le \operatorname{Var} \operatorname{Bin}(M, p) = \mu(1-p).$$

Substituting this bound in $\mathbb{E}[K(K-1)]/[M(M-1)]$ gives a joint inclusion probability at most p^2. Lemma D.4 again applies. On its joint good event, of probability at least $13/16$, the standard-star version of (D.3) has bracket at least $1/2$ whenever $T \ge 512/(pq)$. Thus the population risk is at least $p/108$ there.

Time and scaling. In both cases,

$$\frac{512}{pq} \le \frac{16nM}{\ell} \le 32nd.$$

Moreover, for $M \ge 256$, the chosen parameters satisfy

$$\ell \ge c_0 \min\{n, \log(1+n\tau), 1+\log d\} \tag{D.7}$$

for an absolute $c_0 > 0$. To verify this for all admissible parameters, write the minimum before flooring as s. Then $M \ge s/2$ and $\log(M/64) \ge \frac{1}{4}\log M \ge \frac{1}{8}\log s$. The three possible terms in $\log s$ are $\log(d-1)$, $\frac{1}{2}\log(n\rho) - \log 8$ (or $\frac{1}{2}\log(nk) - \log 8$), and, when present, $n/4$. Since $s \ge 256$, each bounds the corresponding term of (D.7) from below up to an absolute factor, using $\tau \le 2\rho$ or $\tau \le 3k$. This proves the rate.

If $M < 256$ or $d = 1$, use the scalar example: Hessian one with probability $a = \min\{\tau, 1/(n+1)\}$ and zero otherwise. Its expected risk at every horizon is at least

$$\frac{a(1-a)^n}{2} \ge \frac{1}{4e}\min\{\tau, 1/n\}.$$

This matches the displayed rate up to an absolute factor. Indeed, $M < 256$ implies that $d \le 257$, that $n\tau \le 3(8 \cdot 257)^2$, or that $n \le 4\log 257$. In each case, the numerator of the desired rate is bounded by an absolute constant. If $n\tau < 1$, use $\log(1+n\tau) \le n\tau$ instead.

The GD lower bounds transfer to fixed-dataset SGD by the conditional-mean identity and Jensen's inequality. Each construction is fixed before sampling and works at every stated late horizon. Convergence and bounded risk carry these lower bounds to the infinite-time limit with the law and initialization held fixed. For the limiting upper bound, let $P_S = P_{\ker H_S}$. The simultaneous comparison gives $F(P_S\boldsymbol{u}) \le Z_S/2$ because $F_S(P_S\boldsymbol{u}) = 0$. Taking expectation and then the supremum over laws and deterministic $\boldsymbol{u}$ proves the upper bound for $\mathcal{R}_\infty$ in (3.5). Neither argument interchanges a time limit with a worst-case supremum. □

Corollary D.5 (Sample-Hessian rank ceiling)**.** *For the subclass* $\operatorname{rank} H_z \le r$ *almost surely,* $1 \le r \le d$*, its worst-case expected risk, for either algorithm and* $T \ge 32nd$*, is*

$$\Theta\left(\min\left\{1, \frac{1+\log\min\{d, nr\}}{n}\right\}\right).$$

At every horizon, the corresponding upper bound holds with an additional $O(1/T)$ *term. In particular, bounded-norm scalar noiseless linear regression has a time-uniform* $O((1+\log n)/n)$ *upper bound for* $T \ge n$*, regardless of ambient dimension. This order is attained at long times when* d *is sufficiently large.*

Proof. The upper bound uses $\operatorname{tr} H \le r$. The lower constructions in the preceding proof have sample rank at most r when $\tau = r$. For $r = 1$, use the activated rank-one case with $\rho = 1$. The scalar exceptional regime is also rank one. □

E Sharp confidence bounds for data-dependent late horizons

We prove the confidence law summarized in Remark 3.3, strengthening the expected resource law of Section 3 to a single event covering all late horizons. The algorithms and averaging convention remain those of Section 2. The sampling law may be any i.i.d. law of PSD contractions, without independence among coefficients of an individual sample. Write $H = \mathbb{E}H_z$, $\tau_0 = \operatorname{tr} H \le \tau$, and $a = \min\{1, \tau\}$. We work in target-zero error coordinates, with deterministic initialization $\boldsymbol{w}_1 = \boldsymbol{u}$ satisfying $\|\boldsymbol{u}\| \le 1$ and population risk $F(\boldsymbol{w}) = \boldsymbol{w}^\top H \boldsymbol{w}/2$. Define

$$\Psi_\delta(n, d, \tau) = \min\left\{1, \tau, \frac{\log(1+n\tau)+\log(1/\delta)}{n}, \frac{1+\log d+\log(1/\delta)}{n}\right\}. \tag{E.1}$$

For a nonnegative random variable Y, let $q_{1-\delta}(Y) = \inf\{r \ge 0 : \mathbb{P}(Y \le r) \ge 1-\delta\}$. For $T = \infty$, the output is the almost-sure limit $P_{\ker H_S}\boldsymbol{u}$ from Lemma C.9. A horizon rule σ may depend on the dataset and the algorithm's full index sequence. It must select an averaged iterate at a horizon $\sigma \ge T_0$ almost surely, rather than construct a new parameter. No stopping-time assumption on σ is required.

Theorem E.1 (Uniform late-horizon confidence law)**.** *There are absolute constants* $c, C > 0$ *such that, for every* $n, d \ge 1$*,* $0 < \tau \le d$*,* $0 < \delta \le 1/8$*, and either algorithm, with* $T_0 = 32nd$*:*

1. *Every admissible instance satisfies*

$$\mathbb{P}\left\{\sup_{T \in \{T_0, T_0+1, \ldots, \infty\}} F(\bar{\boldsymbol{w}}_T) \le C\Psi_\delta(n, d, \tau)\right\} \ge 1-\delta. \tag{E.2}$$

2. *There is one admissible instance, fixed before sampling and unchanged at all later horizons, such that*

$$\mathbb{P}\left\{\inf_{T \in \{T_0, T_0+1, \ldots, \infty\}} F(\bar{\boldsymbol{w}}_T) \ge c\Psi_\delta(n, d, \tau)\right\} \ge 2\delta. \tag{E.3}$$

The same lower-bound distribution and initialization work for GD and SGD. The probability spaces include the respective algorithm randomness.

Consequently, for every admissible horizon rule $\sigma \ge T_0$*, taking the supremum over laws with* $0 \preceq H_z \preceq I$ *and* $\operatorname{tr} H \le \tau$ *gives*

$$\sup_{D,\, \|\boldsymbol{u}\| \le 1} q_{1-\delta}\big(F(\bar{\boldsymbol{w}}_\sigma)\big) = \Theta\big(\Psi_\delta(n, d, \tau)\big). \tag{E.4}$$

All constants are uniform in δ *and in the horizon rule. At* $\tau = 0$ *every risk is zero. At a fixed confidence level, the resource-driven lower example has probability at least* $1/4$*, not a vanishing exceptional probability.*

The cap by τ in (E.1) is essential at very small trace. The resource construction gives the sufficient threshold $32nd$, without identifying the earliest harmful horizon or ruling out effective stopping before it. The infimum in (E.3) covers a single realized trajectory and therefore gives more than separate marginal lower bounds for different T.

E.1 An exponential relative-curvature comparison

Lemma E.2. *For arbitrary i.i.d. PSD contractions, set* $Z = \lambda_{\max}^{+}(H - 2H_S)$ *and* $D_* = 1 + \min\{d, 2n \operatorname{tr} H\}$. *Then*

$$\mathbb{E} \exp\left(\frac{2n}{9} Z\right) \le D_*. \tag{E.5}$$

In particular, for every $0 < \varepsilon < 1$, *with probability at least* $1 - \varepsilon$,

$$H \preceq 2H_S + \frac{9}{2n} \log(D_*/\varepsilon) I. \tag{E.6}$$

Proof. Apply the projection rounding and unordered-pair argument of Theorem D.1 to the convex map $X \mapsto \exp[(2n/9)\lambda_{\max}^{+}(H - 2X)]$. This bounds the exponential moment directly. Conditional spectral Bernoulli rounding gives

$$\mathbb{E} e^{(2n/9)Z} \le \mathbb{E} e^{(2n/9)\lambda_{\max}^{+}(H - 2\bar{P})}.$$

Here $\bar{P} = n^{-1} \sum_i P_i$, where the i.i.d. random projections P_i have mean H and expected rank $\operatorname{tr} H$. For an independent copy $\bar{P}'$, a second Jensen inequality gives

$$\mathbb{E} e^{(2n/9)\lambda_{\max}^{+}(H - 2\bar{P})} \le \mathbb{E} e^{(2n/9)\lambda_{\max}^{+}(\bar{P}' - 2\bar{P})}.$$

Neither use of Jensen asserts stochastic domination of Z itself.

Condition on the unordered pairs $\{P_i, P_i'\}$. The joint span L is then fixed, has dimension $k \le d$, and satisfies $\mathbb{E}k \le 2n \operatorname{tr} H$. With $C_i = P_i + P_i'$ and $D_i = P_i' - P_i$, put $V = (3 \sum_i \epsilon_i D_i - \sum_i C_i)/9$. The conditional signed-swap representation implies $(2n/9)\lambda_{\max}^{+}(\bar{P}' - 2\bar{P}) \overset{\text{law}}{=} \lambda_{\max}^{+}(V)$. The offset matrix-Laplace calculation in Theorem D.1, based on the inequality of Tropp (2012), gives $\mathbb{E}_\epsilon \operatorname{tr}_L e^V \le k$. For $k > 0$, $e^{\lambda_{\max}^{+}(V)} \le 1 + \operatorname{tr}_L e^V$. For $k = 0$, its value is exactly one. Hence

$$\mathbb{E}_\epsilon e^{\lambda_{\max}^{+}(V)} \le 1 + k, \qquad \mathbb{E} e^{(2n/9)Z} \le 1 + \mathbb{E}k \le D_*.$$

Exponential Markov proves (E.6). The proof uses neither a spectral gap nor a deterministic bound on the sample ranks. □

E.2 One optimization event for the entire SGD path

A nonnegative supermartingale gives a bound uniform over training horizons without a union bound (Howard et al., 2020).

Lemma E.3. *For every fixed training set, every* $\|\boldsymbol{u}\| \le 1$, *and* $0 < \varepsilon < 1$, *conditional on that set, with probability at least* $1 - \varepsilon$,

$$F_S(\bar{\boldsymbol{w}}_T^{\mathrm{SGD}}) \le \frac{16/7 + \log(1/\varepsilon)}{2(1 - e^{-1})T} \quad \textit{simultaneously for all integers } T \ge 1. \tag{E.7}$$

For GD the deterministic bound is $F_S(\bar{\boldsymbol{w}}_T^{\mathrm{GD}}) \le 8/(7T)$ *at all such horizons.*

Proof. Fix the dataset throughout. Let $\mathcal{F}_{t-1}$ include the indices preceding update t, so $\boldsymbol{w}_t$ is measurable, and define

$$Y_t = \boldsymbol{w}_t^\top H_{i_t} \boldsymbol{w}_t \in [0, 1], \qquad m_t = \mathbb{E}[Y_t \mid S, \mathcal{F}_{t-1}] = \boldsymbol{w}_t^\top H_S \boldsymbol{w}_t.$$

Radius nonincrease and $H_{i_t}^2 \preceq H_{i_t}$ imply, on every index path,

$$\|\boldsymbol{w}_{t+1}\|^2 \le \|\boldsymbol{w}_t\|^2 - \frac{7}{16} Y_t, \qquad \sum_{t=1}^{N} Y_t \le 16/7 \quad (N \ge 1).$$

Write $b = 1 - e^{-1}$. Convexity on $[0, 1]$ gives $e^{-y} \le 1 - by$, hence

$$\mathbb{E}[e^{bm_t - Y_t} \mid S, \mathcal{F}_{t-1}] \le e^{bm_t}(1 - bm_t) \le 1.$$

Thus $M_N = \exp(b\sum_{t\le N} m_t - \sum_{t\le N} Y_t)$ is a nonnegative supermartingale with $M_0 = 1$. Its probability of ever exceeding $1/\varepsilon$ is at most ε. Indeed, stopping at the first crossing or a deterministic N gives expected stopped value at most one. Markov's inequality bounds the crossing probability by ε, and letting $N \to \infty$ gives the claim. Outside the crossing event,

$$b\sum_{t=1}^{N} m_t \le 16/7 + \log(1/\varepsilon) \quad \text{for every } N.$$

Jensen's inequality for F_S proves (E.7). The index path can be continued for this proof even if a rule returns its output after $T-1$ updates. The extra update does not affect the first T averaged points. The GD bound is the deterministic empirical energy estimate in (D.1). □

Corollary E.4 (Explicit all-horizon upper bound)**.** *For either algorithm and each $0 < \delta < 1$, with probability at least $1-\delta$, simultaneously for every integer $T \ge 1$,*

$$F(\bar{\boldsymbol{w}}_T) \le \min\left\{\frac{\min\{1,\tau\}}{2}, \frac{16/7 + \log(2/\delta)}{(1-e^{-1})T} + \frac{9}{4n}\log\frac{2[1+\min\{d,2n\tau\}]}{\delta}\right\}. \tag{E.8}$$

The bound passes to the infinite-time limit on the same event, interpreting $1/T = 0$.

Proof. Apply Lemma E.2 at error $\delta/2$ over the dataset. On that event, simultaneously for every $\|\boldsymbol{w}\| \le 1$,

$$F(\boldsymbol{w}) \le 2F_S(\boldsymbol{w}) + \frac{9}{4n}\log\frac{2[1+\min\{d,2n\tau\}]}{\delta}.$$

For SGD, intersect this event with the conditional event of Lemma E.3 at error $\delta/2$. Averaging the conditional probability over datasets and applying a union bound gives total failure probability at most δ. GD needs no second random event and satisfies the same, looser bound. The deterministic cap follows from the radius bound and $\|H\| \le \min\{1,\tau\}$. Almost-sure convergence from Lemma C.9 extends the bound to the limit on this event, without exchanging a supremum and a limit. □

E.3 A SINGLE EVENT OF PERSISTENT FAILURE

An expectation lower bound and Jensen alone do not imply a probability lower bound. We therefore exploit an additional pathwise invariant of the resource construction.

Lemma E.5. *For each (n,d,τ) with $0 < \tau \le d$, there is a fixed instance from the proof of Theorem D.3 for which, for both algorithms,*

$$\mathbb{P}\left\{\inf_{T\ge 32nd} F(\bar{\boldsymbol{w}}_T) \ge cR_0\right\} \ge \frac{1}{4}, \quad R_0 = \min\left\{1, \frac{\log(1+n\tau)}{n}, \frac{1+\log d}{n}\right\}. \tag{E.9}$$

The infimum also includes the infinite-time limit.

Proof. Consider first the two star-construction branches with $M \ge 256$ in the cited proof. Every sample Hessian has the form $c_z\boldsymbol{v}\boldsymbol{v}^\top + \mathrm{diag}(0,\boldsymbol{d}_z)$, with $c_z \ge 0$, $\boldsymbol{d}_z \ge 0$ coordinatewise, $c_z(1+q) \le 1$, and $\|\boldsymbol{d}_z\|_\infty \le 1$. Put $r_t = \sqrt{q}\,x_t - \boldsymbol{b}^\top\boldsymbol{y}_t$. The coordinate recurrences for either a sampled step or a full-batch step preserve $r_t \ge 0$ and $\boldsymbol{y}_t \ge 0$. Indeed,

$$r_{t+1} = [1-\eta c_z(1+q)]r_t + \eta\boldsymbol{b}^\top\mathrm{diag}(\boldsymbol{d}_z)\boldsymbol{y}_t, \qquad y_{j,t+1} = (1-\eta d_{z,j})y_{j,t} + \eta c_z b_j r_t,$$

with the sample coefficients replaced by their empirical means for GD. All displayed coefficients are nonnegative.

Fix the dataset and let U be its unobserved coordinate penalties, $Q = |U|/M$, and $\ell_S = q/(q+Q)$. Starting at $\boldsymbol{e}_0$, every index path satisfies

$$\boldsymbol{y}_{U,t} = \frac{\boldsymbol{b}_U}{\sqrt{q}}(1-x_t), \qquad 1 \ge x_t \ge \frac{Q}{q+Q}, \qquad x_{t+1} \le x_t.$$

The lower bound on x_t follows by writing $r_t + \boldsymbol{b}_{U^c}^\top\boldsymbol{y}_{U^c,t} = ((q+Q)x_t - Q)/\sqrt{q} \ge 0$. Consequently $D_T = 1 - \bar{x}_T$ is nondecreasing in T and takes values in $[0,\ell_S]$.

The proof of Theorem D.3 exhibits a dataset event of probability at least $3/4$ on which $Q \in [q/2, 2q]$ and the exact GD bound at $T_0 = 32nd$ satisfies $D_{T_0}^{\mathrm{GD}} \geq \ell_S/2$. For SGD the conditional-mean identity gives $\mathbb{E}_{\mathrm{alg}}[D_{T_0} \mid S] \geq \ell_S/2$. A random variable in $[0, \ell_S]$ with this mean satisfies $\mathbb{P}_{\mathrm{alg}}\{D_{T_0} \geq \ell_S/4 \mid S\} \geq 1/3$, since its mean is at most $\ell_S/4 + (3\ell_S/4)\mathbb{P}\{D_{T_0} \geq \ell_S/4\}$. On this event, simultaneously for all $T \geq T_0$,

$$\|\bar{\boldsymbol{y}}_{U,T}\|^2 = \frac{Q}{q} D_T^2 \geq \frac{qQ}{16(q+Q)^2} \geq \frac{1}{72}.$$

The population coefficient of these squares is $p/2$ in the low-trace branch and $p/6$ in the high-trace branch. Thus the risk is at least $p/432$ at every late horizon with joint probability at least $(3/4)(1/3) = 1/4$. The calibrated p is at least an absolute constant times R_0. GD satisfies the same conclusion on the larger dataset event itself.

In the small-parameter branches, the scalar construction has Hessian one with probability $a_0 = \min\{\tau, 1/(n+1)\}$ and zero otherwise. The all-zero dataset event has probability at least $\big(n/(n+1)\big)^n \geq e^{-1}$. On this event, both algorithms stay at one for every horizon and incur risk $a_0/2$. The bounded-parameter comparison in Theorem D.3 gives $a_0 \geq cR_0$, including when $d = 1$ or $n\tau < 1$. Continuity of the loss and fixed-dataset convergence extend the pathwise lower event to $T = \infty$. □

Lemma E.6 (Confidence-driven lower example)**.** *For every $n, d \geq 1$, $0 < \tau \leq d$, and $0 < \delta \leq 1/8$, a one-dimensional example, embedded in $\mathbb{R}^d$, satisfies for both algorithms*

$$\mathbb{P}\left\{\inf_{T \in \{1,2,\ldots,\infty\}} F(\bar{\boldsymbol{w}}_T) \geq c \min\{1, \tau, \log(1/\delta)/n\}\right\} \geq 2\delta. \tag{E.10}$$

Proof. Set $p = \min\{\tau, 1/2, \log(1/(2\delta))/(2n)\}$, sample Hessian one with probability p and zero otherwise, and initialize at one. These are bounded realizable quadratics of mean trace $p \leq \tau$. Since $p \leq 1/2$, the all-zero dataset event has probability $(1-p)^n \geq e^{-2np} \geq 2\delta$. On this event, no update moves and the risk is $p/2$ at every horizon. For $\delta \leq 1/8$, $\log(1/(2\delta)) \geq (2/3)\log(1/\delta)$, so $p/2$ is at least an absolute constant times the displayed minimum. This is the standard rare-observation lower-bound argument, applied to the confidence parameter. □

Proof of Theorem E.1. For $T \geq 32nd$, the upper bound (E.8) is at most $C\Psi_\delta$. Indeed,

$$\log[1 + \min\{d, 2n\tau\}] \leq \min\{1 + \log d,\ 2\log(1+n\tau)\},$$

and both $1/T$ and $\log(2/\delta)/T$ are bounded by $C\log(1/\delta)/n$. The deterministic cap handles arbitrarily small trace.

For the lower bound, let $b = n^{-1}\min\{\log(1+n\tau), 1+\log d\}$, $s = \log(1/\delta)/n$, and $a = \min\{1, \tau\}$. Then $R_0 = \min\{a, b\}$ and $\Psi_\delta = \min\{a, b+s\}$. Since $\max\{\min(a,b), \min(a,s)\} \geq \frac{1}{2}\min(a, b+s)$, either Lemma E.5 or Lemma E.6 attains $c\Psi_\delta$. Choose the corresponding instance before sampling. Its lower-bound event covers every late horizon and has probability at least 2δ, since $1/4 \geq 2\delta$. This proves (E.3).

Both events remain valid when a data-dependent horizon is substituted. The lower event has probability strictly greater than δ, so the definition of the quantile gives the lower side of (E.4), including atomic risk distributions. The upper event gives the upper side. The lower instance does not need to know which admissible late-horizon rule will be used. □

Rank restriction. If sample rank is bounded by r, substitute $\tau = r$ in (E.1) and omit the redundant cap τ (since $r \geq 1$). The upper bound follows from $\mathrm{tr}\, H \leq r$. The resource lower examples already satisfy the rank cap, and the confidence-driven example is rank one. Thus the same sharp quantile law holds under a sample-rank restriction. In particular, for bounded-feature scalar regression and sufficient dimension it is $\Theta(\min\{1, [\log(1+n)+\log(1/\delta)]/n\})$ at all prescribed or adaptively selected late horizons. This is a distribution-free algorithmic worst-case statement, not a characterization of each design.

Scope of the confidence law. The matching confidence law directly quantifies risk on rare datasets, without converting an expectation bound by Markov's inequality. It remains valid after selection from the entire late trajectory, even by an oracle that chooses its best average. It supplies neither a certificate computable from the training data nor a solution to the unrestricted finite-time problem in (N.1). The proof combines standard matrix Laplace and supermartingale inequalities with a trace-dependent exponential comparison, a uniform optimization event, and a matching pathwise persistence argument for both algorithms.

F SAME-INSTANCE DYNAMICS AND SAMPLING-LAW SEPARATION

This appendix isolates the slow empirical mode in a rank-two design and proves the sampling-law separation used in Section 4.2. Its count estimates also support the rank-one proof in Appendix G. Throughout this appendix, $A = 3H_S$. The gradient step includes the corresponding scaling factor. The constructions below use target-zero error coordinates. Appendix G gives the translation to the zero-initialized regression instance in Theorem 4.1.

F.1 A SINGLE SPARSE INSTANCE WITH A MATCHED RISK CURVE

F.1.1 CONSTRUCTION AND STATEMENT

For all sufficiently large integers n, set

$$M = \lfloor \sqrt{n} \rfloor, \quad q = 64/M, \quad p = 1 - q^{1/n}, \quad \alpha = \sqrt{q}, \quad \boldsymbol{b} = M^{-1/2}\mathbf{1}. \tag{F.1}$$

A categorical mask equals $\boldsymbol{e}_j$ with probability p for every $j \in [M]$, and is zero otherwise. This is valid because $Mp \to 0$. Define

$$f((x, \boldsymbol{y}); z) = \frac{1}{6}\left[(\alpha x - \boldsymbol{b}^\top \boldsymbol{y})^2 + \sum_j z_j y_j^2\right]. \tag{F.2}$$

Use the deterministic prepared initialization

$$\boldsymbol{w}_1 = \frac{1}{2}\boldsymbol{u} + \gamma \boldsymbol{v}, \qquad \boldsymbol{u} = \left(\frac{1}{1+q}, \frac{\alpha \boldsymbol{b}}{1+q}\right), \qquad \boldsymbol{v} = (\alpha, -\boldsymbol{b}), \quad \gamma = \frac{1}{10\sqrt{n}}. \tag{F.3}$$

Since $\boldsymbol{u} \perp \boldsymbol{v}$, the initial squared norm is at most $1/4 + (1+q)/(100n)$. Every sample Hessian is PSD, has operator norm at most one and rank at most two. The loss is globally nonnegative and C^∞, and has common minimizer zero. No projection or regularization is used.

Theorem F.1 (Complete expected averaged-risk curve)**.** *There are absolute $0 < c < C < \infty$ and n_0 such that, for the single instance* (F.1)–(F.3)*, every $n \geq n_0$, every integer $T \geq 1$, and either algorithm,*

$$c\Phi_n(T) \leq \mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}}) \leq C\Phi_n(T), \qquad \Phi_n(T) = \frac{1}{n(1+T)^2} + pq + p\min\{(pqT)^2, 1\}. \tag{F.4}$$

The same distribution and initialization are used for every horizon.

The theorem determines the curve up to absolute constants, without asserting strict unimodality at successive iterates. For any fixed sufficiently large absolute C_1, it implies

$$F(\boldsymbol{w}_1) = \Theta(1/n), \tag{F.5}$$

$$\mathbb{E}F(\bar{\boldsymbol{w}}_n^{\mathsf{A}}) = \Theta(\log n / n^{3/2}), \tag{F.6}$$

$$\inf_{T \geq C_1 n^{3/2}/\log n} \mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}}) = \Theta(\log n / n) \tag{F.7}$$

The same-instance crossing scale for risk of order $1/n$ is

$$T_\times = \frac{1}{pq\sqrt{np}} = \Theta\left(\frac{n^{3/2}}{(\log n)^{3/2}}\right). \tag{F.8}$$

For $T \geq n$, $T = o(T_\times)$ gives $o(1/n)$ risk, $T = \Theta(T_\times)$ gives $\Theta(1/n)$, and $T/T_\times \to \infty$ gives $\omega(1/n)$. This is an instance-specific sharp crossing, not the earliest worst-case overfitting threshold.

F.1.2 A HIGH-PROBABILITY EMPIRICAL REGULARITY EVENT

Let $N_j := \sum_{i=1}^n z_{ij}$ denote the coordinate counts, put $a_j = N_j/n$, and define $V := \{j : a_j > 0\}$, $U := [M] \setminus V$, $Q := 1 - |V|/M$, $B_1 = \sum_{j\in V} b_j^2/a_j$, and $B_2 = \sum_{j\in V} b_j^2/a_j^2$. Write $D_i := \mathrm{diag}(z_{i1}, \dots, z_{iM})$, $D_S := n^{-1}\sum_i D_i = \mathrm{diag}(a_1, \dots, a_M)$, and $D_V := \mathrm{diag}(a_j : j \in V)$. Set $a_{\max} := \max_{j\in[M]} a_j$ and, when $V \neq \varnothing$, $a_{\min} := \min_{j\in V} a_j$. For scalars c_j, $\mathrm{diag}_V(c_j)$ denotes the $(M+1)\times(M+1)$ diagonal matrix with entries c_j on coordinates $j \in V$ and zero elsewhere.

Lemma F.2. *For the categorical instance there is an event E with $\mathbb{P}(E^c) = O(M^{-10})$ on which*

$$a_{\max} \le 16p, \qquad \frac{1}{M}\sum_j a_j \le 2p, \qquad q + Q \le \frac{1}{32np}, \\ \frac{1}{8p} \le B_1 \le \frac{4}{p}, \qquad B_2 \le \frac{12}{p^2}. \tag{F.9}$$

Also $\mathbb{P}(G) \ge 15/16$ for $G = \{q/2 \le Q \le 2q\}$.

Proof. Lemma D.4 gives the probability bound for G. To construct E, put $\mu = np$. Independently draw $K_j \sim \mathrm{Pois}(\mu)$ and a zero-category count with mean $n(1 - Mp)$. Conditioning their sum to equal n gives the required multinomial distribution. By Stirling's inequality, the conditioning event has probability at least $c/\sqrt{n}$. Since $M^2 \le n < (M+1)^2$, conditioning increases the failure probabilities by at most a factor $O(M)$.

For any integer $m \ge 1$, a Poisson variable K of mean μ satisfies $\mathbb{E}[n^m \mathbf{1}_{K>0}/K^m] \le (m+1)!/p^m$. This follows from $k^{-m} \le (m+1)!/[(k+1)\cdots(k+m)]$ for $k \ge 1$ and

$$\mathbb{E}\frac{1}{(K+1)\cdots(K+m)} = \frac{1}{(m-1)!}\int_0^1 e^{-\mu(1-t)}(1-t)^{m-1}\,dt \le \mu^{-m}.$$

The summands defining B_1 are bounded by n, have mean at most $2/p$, and second moment at most $6/p^2$. Bernstein's scalar inequality therefore bounds failure of $B_1 \le 4/p$ by $\exp[-cM/(1+\mu)]$ before conditioning. For B_2, use a bound n^2, mean $6/p^2$, and second moment $120/p^4$ to obtain failure probability $\exp[-cM/(1+\mu^2)]$ for $B_2 \le 12/p^2$.

The sum of counts is Poisson with mean $M\mu$, giving failure probability $e^{-cM\mu}$ for the empirical-mean bound. The Poisson upper tail gives $\mathbb{P}(\max_j K_j > 16\mu) \le M\exp[-(16\log 16 - 15)\mu]$. Here $\mu = \log(M/64) + o(1)$, so this remains $O(M^{-10})$ after conditioning. Finally, the number of zero counts is binomial with mean $Me^{-\mu} \le 128$ for all sufficiently large n. A binomial tail bounds the failure probability of $Q \le 1/(64\mu)$ by e^{-cM}. Since $q \le 1/(64\mu)$ eventually, this gives the bound on $q+Q$ in (F.9). All of these exponential tails except the maximum-count bound are superpolynomially small. Their union, after conditioning, has probability $O(M^{-10})$. On this event $Q \le 1/2$, and Cauchy–Schwarz gives $B_1 \ge (1-Q)^2/[M^{-1}\sum_j a_j] \ge 1/(8p)$. □

F.1.3 THE SLOW EMPIRICAL MODE AND THE PREPARED TRAJECTORY

For the single coupled star, put $A = 3H_S = \boldsymbol{v}\boldsymbol{v}^\top + \mathrm{diag}(0, D_S)$ and $P = I - A/12$. Consider the auxiliary GD trajectory $\boldsymbol{w}_t = P^{t-1}\boldsymbol{e}_0 = (x_t, \boldsymbol{y}_t)$, with residual $r_t = \boldsymbol{v}^\top \boldsymbol{w}_t$. Write $\kappa = Q/(q+Q)$.

Lemma F.3. *On E, the spectral measure of $\boldsymbol{e}_0$ for A has a zero mass κ, one positive eigenvalue $\lambda < \min_{j\in V} a_j/2$, and all other positive eigenvalues at least $\min_{j\in V} a_j$. The low eigenvalue obeys*

$$\lambda\left(1 + \sum_{j\in V}\frac{b_j^2}{a_j - \lambda}\right) = q + Q, \qquad \frac{p(q+Q)}{9} \le \lambda \le 8p(q+Q). \tag{F.10}$$

Its spectral weight is

$$\omega = \frac{q}{q + Q + \lambda^2 B_2(\lambda)}, \qquad B_2(\lambda) = \sum_{j\in V}\frac{b_j^2}{(a_j - \lambda)^2}, \tag{F.11}$$

and the total weight W on the remaining positive eigenvalues is at most $3072q$. Moreover,

$$\sum_{j\in V}\sup_{t\ge 1} y_{j,t}^2 \le Cq. \tag{F.12}$$

Proof. The secular equation for a diagonal matrix plus $\boldsymbol{v}\boldsymbol{v}^\top$ gives (F.10). Its left side is strictly increasing from zero to infinity on $(0, a_{\min})$. Rank-one interlacing puts the remaining positive eigenvalues at least at $a_{\min}$. Unexcited zero directions have no effect. The projection of $\boldsymbol{e}_0$ on the kernel has squared norm κ, as proved in Appendix C. Since the reciprocal sum in (F.10) is at least B_1, the root is at most $8p(q+Q) \le 1/(4n) \le a_{\min}/4$. The same sum is then at most $2B_1$, proving the lower bound.

If $r = \boldsymbol{v}^\top \boldsymbol{w}$ for a λ-eigenvector, its coordinates are $x = \alpha r/\lambda$, $\boldsymbol{y}_U = -\boldsymbol{b}_U r/\lambda$, and $y_j = b_j r/(a_j - \lambda)$ for $j \in V$. Normalizing gives (F.11). Therefore

$$W = 1 - \kappa - \omega = \frac{q\lambda^2 B_2(\lambda)}{(q+Q)(q+Q+\lambda^2 B_2(\lambda))} \le 3072q,$$

using $B_2(\lambda) \le 4B_2 \le 48/p^2$.

Let λ_k and ω_k denote the positive eigenvalues of A and their spectral weights for $\boldsymbol{e}_0$. The residual has the nonnegative spectral expansion $r_t = \alpha^{-1} \sum_{k:\lambda_k>0} \lambda_k \omega_k (1-\lambda_k/12)^{t-1}$. The slow term is bounded by $8p\alpha$, and the sum over all times of the fast part is exactly $12W/\alpha$. The coordinate recurrence $y_{j,t+1} = (1 - a_j/12) y_{j,t} + b_j r_t/12$, with $y_{j,1} = 0$, thus gives $\sup_t y_{j,t} \le \alpha b_j (8p/a_j + 3072)$. Squaring and summing, then using the bound on B_2, gives (F.12). □

Lemma F.4 (Prepared start and the GD curve)**.** *For the star initialized at $\boldsymbol{u}$ in* (F.3)*,*

$$\mathbb{E}_S F(\bar{\boldsymbol{w}}_T^{u,\mathrm{GD}}) = \Theta\left(pq + p\min\{(pqT)^2, 1\}\right), \qquad T \ge 1.$$

On E, its conditional mean-coordinate envelopes obey

$$\sum_{j \in V} \sup_{t \ge 1} (y_{j,t}^u)^2 \le Cq. \tag{F.13}$$

Proof. Although A depends on S, $A\boldsymbol{e}_0 = \alpha \boldsymbol{v}$ does not. Thus $\boldsymbol{u} = (I - A/(1+q))\boldsymbol{e}_0$ is data independent, and polynomial commutation gives the exact identities

$$\begin{aligned} x_t^u &= x_t - \alpha r_t/(1+q), \\ y_{j,t}^u &= (1 - a_j/(1+q)) y_{j,t} + b_j r_t/(1+q), \\ r_t^u &= \boldsymbol{b}^\top D_S \boldsymbol{y}_t/(1+q). \end{aligned} \tag{F.14}$$

The auxiliary trajectory has $r_t, \boldsymbol{y}_t \ge 0$, $x_t \le 1$, and $\boldsymbol{b}^\top \boldsymbol{y}_t \le \alpha$. The identities in (F.14) therefore give nonnegative prepared coordinates and residual, with $r_t^u \le 16p\alpha$ on E. Lemma F.3 and $r_t \le \alpha$ give (F.13).

Upper bound. Let $g_T(\lambda) = T^{-1} \sum_{t=0}^{T-1} (1-\lambda/12)^t$. Spectral expansion yields

$$1 - \bar{x}_T = \omega(1 - g_T(\lambda)) + \sum_{\text{fast}} \omega_k (1 - g_T(\lambda_k)) \le Cq + CpqT.$$

Here the cancellation $\omega\lambda \le 8pq$ is essential. The unseen-coordinate identity and (F.14) then imply, on E,

$$\|\bar{\boldsymbol{y}}_{U,T}^u\|^2 \le CQq(1+pT)^2, \qquad \|\bar{\boldsymbol{y}}_{U,T}^u\|^2 \le 1.$$

Using $\mathbb{E}Q = q$ gives an expected bound $C[q^2 + \min\{(pqT)^2, 1\}]$. The seen coordinates contribute at most Cpq by (F.13), and the residual contributes at most Cp^2q. Since the risk is bounded and $\mathbb{P}(E^c) = O(M^{-10}) = o(pq)$, this proves the upper bound.

Uniform lower floor. The prepared start still satisfies the exact unseen invariant $\boldsymbol{y}_{U,t}^u = \boldsymbol{b}_U(1 - x_t^u)/\alpha$. Its nonnegative residual and seen coordinates imply $x_t^u \ge \kappa$. Minimizing the population star loss over $\boldsymbol{y}$ at a fixed x gives

$$F(x, \boldsymbol{y}) \ge \frac{pq}{6(1+p)} x^2. \tag{F.15}$$

On G, $\bar{x}_T^u \ge \kappa \ge 1/3$. Thus (F.15) supplies a cpq lower bound at every horizon.

Growing unseen contribution. All positive eigenvalues are at least λ, so

$$1 - \bar{x}_T \ge (1-\kappa)(1 - g_T(\lambda)) \ge c\min\{pqT, 1\} \quad \text{on } E \cap G, \quad T \ge 2.$$

The last inequality follows by averaging $1-(1-s)^t$ with $s=\lambda/12$. For $T\geq 2$, this average is bounded below by an absolute constant times $\min(sT,1)$. Also $1-\kappa\geq 1/3$ and $\lambda\geq pq/6$ on G. Equation (F.14) only increases the unseen coordinates relative to the auxiliary trajectory. Since $Q/q\geq 1/2$, their population risk is at least $cp\min\{(pqT)^2,1\}$. The event $E\cap G$ has probability bounded below by an absolute constant. At $T=1$, this smaller term is absorbed by pq. These estimates give the two lower terms in the curve. □

F.1.4 The conditional covariance principle for multi-pass SGD

We analyze SGD conditional on its fixed dataset. The following bound holds without population resampling or a uniform lower bound on the positive empirical eigenvalues.

Lemma F.5 (Categorical-mask covariance bootstrap)**.** *Fix any dataset of categorical masks (zero or a coordinate vector) in the star model, and let a_j denote its frequencies. Let $\boldsymbol{\mu}_t=\mathbb{E}_{\mathrm{alg}}[\boldsymbol{w}_t\mid S]=P^{t-1}\boldsymbol{w}_1$, and set $m_j=\sup_{t\geq 1}(\mu_{j,t})^2$ for seen $\boldsymbol{y}$-coordinates. Then*

$$0\leq \mathbb{E}_{\mathrm{alg}}[F(\bar{\boldsymbol{w}}_T)\mid S]-F(\bar{\boldsymbol{\mu}}_T)\leq \frac{a_{\max}+p}{120}\sum_{j\in V} m_j,\qquad T\geq 1. \tag{F.16}$$

The same bound holds for the last iterate in place of the average. In particular, on E and at the prepared start, the extra risk is $O(pq)$ uniformly in T.

Proof. All matrices in this proof are conditional on the fixed dataset. If $V=\varnothing$, all sampled updates coincide and the assertion is immediate. Assume $V\neq\varnothing$. The difference between a sampled update and its mean is

$$\boldsymbol{\xi}_t=-\frac{1}{12}(0,(D_{i_t}-D_S)\boldsymbol{y}_t),\qquad \mathbb{E}[\boldsymbol{\xi}_t\mid S,\mathcal{F}_t]=0.$$

Categorical exclusivity gives the PSD inequality

$$\mathbb{E}[\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top\mid S]\preceq\frac{1}{144}\operatorname{diag}_V(a_j\mathbb{E}[y_{j,t}^2\mid S]).$$

Indeed its conditional inner matrix is $\operatorname{diag}(a_jy_j^2)-(D_S\boldsymbol{y})(D_S\boldsymbol{y})^\top$. Noise is supported on seen coordinates and hence on $(\ker A)^\perp$.

Let $\Gamma_t=\operatorname{Cov}(\boldsymbol{w}_t\mid S)$ and $M_j=\sup_t\mathbb{E}[y_{j,t}^2\mid S]$. These suprema are finite by radius nonincrease. Then

$$\Gamma_{t+1}=P\Gamma_tP+\mathbb{E}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top,\qquad \Gamma_t\preceq\bar{\Gamma}:=\frac{1}{144}\sum_{m\geq 0}P^m\operatorname{diag}_V(a_jM_j)P^m.$$

The series converges on $(\ker A)^\perp$.

The pseudoinverse satisfies

$$(A^+)_{VV}=D_V^{-1}. \tag{F.17}$$

To verify it, solve $A\boldsymbol{u}=\boldsymbol{z}$ for a vector $\boldsymbol{z}$ supported on V. The x-equation forces the shared residual to vanish, and the seen equations force $\boldsymbol{u}_V=D_V^{-1}\boldsymbol{z}$. Imposing $\boldsymbol{u}\perp\ker A$ determines the other coordinates: if $s=\boldsymbol{b}_V^\top D_V^{-1}\boldsymbol{z}$, then $u_x=\alpha s/(q+Q)$ and $\boldsymbol{u}_U=-\boldsymbol{b}_Us/(q+Q)$. This identifies the full block in (F.17).

Since $0\preceq A\preceq 3I$, we have $I-P^2=A/6-A^2/144\succeq(7/48)A$. For any seen coordinate j, (F.17) implies

$$\sum_{m\geq 0}\|P^m\boldsymbol{e}_j\|^2\leq\frac{48}{7a_j},\qquad \sum_{m\geq 0}\boldsymbol{e}_j^\top P^mAP^m\boldsymbol{e}_j\leq 48/7. \tag{F.18}$$

These inequalities follow by diagonalizing A on its positive subspace and require no sample-independent spectral gap.

Taking diagonal entries in the covariance bound gives

$$M_j\leq m_j+\sum_{k\in V}L_{jk}M_k,\qquad L_{jk}=\frac{a_k}{144}\sum_{m\geq 0}(P^m)_{jk}^2.$$

By (F.18), each column sum of L is at most $1/21$. Summing and absorbing this term gives $\sum_{j\in V} M_j \le (21/20)\sum_{j\in V} m_j$. The same identities give

$$\operatorname{tr}\bar{\Gamma} \le \frac{1}{21}\sum_j M_j, \qquad \operatorname{tr}(A\bar{\Gamma}) \le \frac{1}{21}\sum_j a_j M_j.$$

The population star Hessian is $(\boldsymbol{v}\boldsymbol{v}^\top + p\,\mathrm{diag}(0, I_M))/3 \preceq (A + pI)/3$. Therefore its quadratic variance cost at any time is at most

$$\frac{1}{6}[\operatorname{tr}(A\bar{\Gamma}) + p\operatorname{tr}\bar{\Gamma}] \le \frac{a_{\max}+p}{126}\sum_j M_j \le \frac{a_{\max}+p}{120}\sum_j m_j.$$

Convexity of the population quadratic gives the same bound for the average of the centered iterates, regardless of their dependence across time. Conditional Jensen gives the lower bound. For the prepared start on E, (F.13) bounds the extra risk by $O(pq)$. □

Proof of Theorem F.1. The $\gamma\boldsymbol{v}$ component supplies a fast transient within the same star. Let $\boldsymbol{w}_t^v = P^{t-1}\boldsymbol{v}$. Since $\boldsymbol{v} = A\boldsymbol{e}_0/\alpha$ and $\boldsymbol{v} \in (\ker A)^\perp$,

$$\frac{1+q}{T} \le \boldsymbol{v}^\top \bar{\boldsymbol{w}}_T^v \le \frac{12}{T}, \qquad \bar{\boldsymbol{w}}_T^v = \frac{12}{\alpha T}(\boldsymbol{e}_0 - P^T\boldsymbol{e}_0). \tag{F.19}$$

For the first inequality, every term $\boldsymbol{v}^\top P^t\boldsymbol{v}$ is nonnegative, the first term is $1+q$, and the infinite sum is $12\boldsymbol{v}^\top A^+\boldsymbol{v} = 12$. The last equality follows from $A\boldsymbol{e}_0 = \alpha\boldsymbol{v}$ and $\boldsymbol{e}_0^\top A\boldsymbol{e}_0 = q$. The vector identity follows by telescoping. Since $p \le q$ for sufficiently large n, (F.19) gives $F(\bar{\boldsymbol{w}}_T^v) \le C/T^2$ for every dataset. Combining $F(\boldsymbol{a}+\boldsymbol{b}) \le 2F(\boldsymbol{a}) + 2F(\boldsymbol{b})$ with Lemma F.4 gives the upper GD curve for initialization $\boldsymbol{u}/2 + \gamma\boldsymbol{v}$.

For its lower curve and coordinate envelopes, set $s = 1/[2(1+q)] - \gamma/\alpha$. For all sufficiently large n, $0 < s \le 1/[2(1+q)]$. In terms of the auxiliary $\boldsymbol{e}_0$ trajectory,

$$\boldsymbol{w}_t^{\mathrm{new}} = \tfrac{1}{2}\boldsymbol{w}_t - sA\boldsymbol{w}_t, \qquad y_{j,t}^{\mathrm{new}} = (\tfrac{1}{2} - sa_j)y_{j,t} + sb_j r_t. \tag{F.20}$$

The second identity has nonnegative coefficients and therefore gives $\boldsymbol{y}_{U,t}^{\mathrm{new}} \ge \boldsymbol{y}_{U,t}/2$ and $\sum_{j\in V}\sup_t(y_{j,t}^{\mathrm{new}})^2 \le Cq$ on E. The exact invariant is now $\boldsymbol{y}_{U,t}^{\mathrm{new}} = \boldsymbol{b}_U(1/2 - x_t^{\mathrm{new}})/\alpha$. The shared residual equals $r_t^u/2 + \gamma\boldsymbol{v}^\top P^{t-1}\boldsymbol{v} \ge 0$. Together with the nonnegative seen coordinates, the invariant gives $x_t^{\mathrm{new}} \ge \kappa/2$. Thus (F.15) gives the pq lower floor on G, and the unseen-coordinate argument from Lemma F.4 gives the growing lower term on $E \cap G$. Finally, the nonnegative residual decomposition and (F.19) give

$$F(\bar{\boldsymbol{w}}_T^{\mathrm{new}}) \ge \gamma^2(1+q)^2/(6T^2) = \Theta\left(\frac{1}{n(1+T)^2}\right).$$

Taking the maximum of these three lower bounds is within a constant of their sum.

For SGD, apply Lemma F.5 to the integrated initialization and the coordinate envelope from (F.20). On E, the additional expected risk is $O(pq)$ uniformly in T. On E^c, the deterministic risk cap and $O(M^{-10})$ failure probability give a negligible contribution. The same lower curve follows by conditional Jensen. Finally $p = \Theta(\log n/n)$, $q = \Theta(n^{-1/2})$, and $npq \to 0$ give (F.5)–(F.8). Convergence and boundedness also pass the persistent lower bound to infinite training. □

F.2 Equal population objectives do not determine the risk curve

Theorem F.6. *For the instance in Theorem F.1, there is another distribution of PSD-contraction quadratic losses in the same dimension, with the same population Hessian, initialization, step size, and common minimizer, and sample rank at most one, such that, for both algorithms and every $T \ge 1$,*

$$\mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A},\mathrm{comm}}) = \Theta\left(\frac{1}{n(1+T)^2} + pq\right). \tag{F.21}$$

For every $T \ge C/(pq)$, the ratio of the risk in Theorem F.1 to that in (F.21) *is $\Theta(1/q) = \Theta(\sqrt{n})$.*

Proof. The population Hessian is $H = (\boldsymbol{v}\boldsymbol{v}^\top + p\,\mathrm{diag}(0, I_M))/3$, with trace $(1 + q + Mp)/3 < 1$ for sufficiently large n. Diagonalize it as $H = \sum_j \theta_j \boldsymbol{a}_j \boldsymbol{a}_j^\top$, with orthonormal eigenvectors $\boldsymbol{a}_j$. As a new sample Hessian, choose $\boldsymbol{a}_j \boldsymbol{a}_j^\top$ with probability θ_j and zero otherwise. This is a valid rank-one PSD-contraction distribution with exactly the required population objective.

The initialization lies in the two-dimensional span of $\boldsymbol{e}_0$ and $(0, \boldsymbol{b})$, which is invariant under H. On this span the matrix is

$$\frac{1}{3}\begin{pmatrix} q & -\sqrt{q} \\ -\sqrt{q} & 1+p \end{pmatrix}.$$

Its eigenvalues satisfy $\theta_- = \Theta(pq)$ and $1/3 \le \theta_+ \le 2/3$. Let $\boldsymbol{a}_\pm$ be the corresponding unit eigenvectors and $\beta_\pm = \boldsymbol{a}_\pm^\top \boldsymbol{w}_1$. Then $\beta_-^2 = \Theta(1)$ and $\beta_+^2 = \Theta(\gamma^2)$. For the second estimate, $\boldsymbol{u} \perp \boldsymbol{v}$ and the eigenvector equation give $|\boldsymbol{u}^\top \boldsymbol{a}_+| \le p\alpha$, while $|\boldsymbol{v}^\top \boldsymbol{a}_+|$ is bounded below by a positive constant. To see the latter bound, express H in the orthonormal basis proportional to $\boldsymbol{v}, \boldsymbol{u}$. Its first diagonal is at least $1/3$, its second is $O(pq)$, and the off-diagonal is $O(p\alpha)$. Since $p\alpha = o(\gamma)$, the small prepared component cannot cancel the $\gamma \boldsymbol{v}$ component. The estimate for β_-^2 then follows from $\|\boldsymbol{w}_1\|^2 = \Theta(1)$.

Every sampled update in this eigenbasis preserves each coordinate's sign and decreases its magnitude. The low-coordinate risk is at most Cpq. It is also at least cpq in expectation for every T: with probability $(1 - \theta_-)^n \ge 1 - n\theta_- \ge 1/2$, the low projector is absent from the entire training set and that coordinate never moves.

For the high coordinate, let $a = N_+/n$ be its empirical frequency. With probability at least $1 - e^{-cn}$, $a \ge \theta_+/2 \ge 1/6$. Conditional on this event, the GD average multiplier is at most $4/(Ta)$. For fixed-dataset SGD, the squared coordinate multiplier at time t has expectation $(1 - 7a/16)^{t-1}$. The L^2 triangle inequality bounds the root mean square average multiplier by $T^{-1}\sum_{t=0}^{T-1}(1 - 7a/16)^{t/2} \le C/(Ta)$. Thus the high-coordinate expected risk is at most $C\gamma^2[T^{-2} + e^{-cn}]$. The exponential term is absorbed by pq for sufficiently large n. Conversely, every averaged coordinate retains at least $1/T$ times the magnitude of its initial coordinate, giving a $c\gamma^2/T^2$ lower bound pathwise. Adding the two contributions gives (F.21). The ratio follows from Theorem F.1. □

Remark F.1. This theorem rules out a universal, constant-factor, *instancewise* risk formula based only on the population Hessian and initialization, even under a common constant sample-rank cap. It does not rule out useful worst-case covariance bounds, or covariance-based equivalences under additional design assumptions. It is a separation between sampling laws, not a separation between GD and SGD.

F.2.1 A SHARP COMMUTING-CLASS COMPARISON

Proposition F.7. *Suppose the sample Hessians are simultaneously diagonalizable in a deterministic orthonormal basis, and initialization is deterministic. For both algorithms and every $T \ge 1$,*

$$\mathbb{E}F(\bar{\boldsymbol{w}}_T) \le C(1/T + 1/n),$$

independently of dimension. The worst-case order over this commuting class is $\Theta(1/T + 1/n)$, already in dimension one. Population risk, for both last and averaged iterates, is pathwise nonincreasing with training time.

Proof. Let $h_{ij} \in [0, 1]$ be the diagonal entries, $a_j = n^{-1}\sum_i h_{ij}$, $\mu_j = \mathbb{E}h_{ij}$, and b_j the initialization coordinates. Each coordinate preserves sign and decreases in absolute value, proving pathwise monotonicity, also for its successive averages. The squared average multiplier for GD and its conditional expectation for SGD are bounded, respectively, by

$$\min\left\{1, \left(\frac{4}{Ta_j}\right)^2\right\} \quad \text{for GD,} \qquad \min\left\{1, \left(\frac{32}{7Ta_j}\right)^2\right\} \quad \text{for SGD.}$$

For fixed-dataset SGD, the squared coordinate multiplier at time t has expectation at most $(1 - 7a_j/16)^{t-1}$, and the average bound follows from the L^2 triangle inequality. Interpret both bounds as one if $a_j = 0$. A scalar multiplicative Chernoff bound gives $\mathbb{P}(a_j < \mu_j/2) \le e^{-n\mu_j/8}$. On the

complement, the risk multiplier is bounded by the first expression below, while the exceptional event contributes at most the second:

$$\mu_j \min\left\{1, \left(\frac{64}{7T\mu_j}\right)^2\right\} \le \frac{C}{T}, \qquad \mu_j e^{-n\mu_j/8} \le \frac{8}{en}.$$

Summing with the deterministic weights b_j^2, whose sum is at most one, proves the upper bound. Zero population means contribute zero. For the lower bound, the scalar rare-sample instance supplies c/n. A deterministic scalar Hessian $1/T$ supplies c/T, because every iterate up to T has magnitude at least $3/4$ when started at one. The maximum of the two lower bounds is within a factor two of their sum. □

Within the quadratic model, simultaneous diagonalizability is sufficient to prevent deterioration of population risk along training. This does not make noncommutativity sufficient for overfitting, nor does it address nonquadratic losses.

G Rank-one risk curves across two geometric safety boundaries

This appendix proves the fixed-instance regression curve in Theorem 4.1 and gives the full coupling family discussed in Section 4. Theorem G.1 gives an explicit construction for the main curve, and Theorem G.2 extends it across the coupling family. Using the count estimates from Appendix F, we derive uniform GD bounds, control the SGD fluctuations, and identify the two critical windows.

For the proof, we use target-zero error coordinates and index the initial vector by $\boldsymbol{w}_0$. The corresponding zero-initialized instance in the main text is obtained by the translation below. Set

$$M = \lfloor \sqrt{n} \rfloor, \quad q = 64/M, \quad p = 1 - q^{1/n}, \quad \boldsymbol{b} = M^{-1/2}\mathbf{1}, \quad \alpha = \sqrt{q}, \quad \gamma = \frac{1}{10\sqrt{n}},$$

and use the deterministic initialization

$$\boldsymbol{u} = \frac{(1, \sqrt{q}\,\boldsymbol{b})}{1+q}, \qquad \boldsymbol{v} = (\sqrt{q}, -\boldsymbol{b}), \qquad \boldsymbol{w}_0 = \frac{1}{2}\boldsymbol{u} + \gamma \boldsymbol{v}.$$

Thus $\|\boldsymbol{w}_0\|^2 = 1/[4(1+q)] + \gamma^2(1+q) < 1$. This is exactly the initialization in (F.3). Sample

$$H_z = \begin{cases} (2/3)\boldsymbol{v}\boldsymbol{v}^\top, & \text{with probability } 1/2, \\ \boldsymbol{e}_j \boldsymbol{e}_j^\top/3, & \text{with probability } p \text{ for each } j \in [M], \\ 0, & \text{with probability } 1/2 - Mp. \end{cases} \tag{G.1}$$

All samples are rank-at-most-one PSD contractions, with population Hessian $H = (\boldsymbol{v}\boldsymbol{v}^\top + p\,\mathrm{diag}(0, I_M))/3$. We use step size $1/4$ and average $\boldsymbol{w}_0, \ldots, \boldsymbol{w}_{T-1}$.

To obtain the instance in Theorem 4.1, set $\boldsymbol{a}_n = \boldsymbol{w}_0$ and keep the same feature law, with

$$\widetilde{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \boldsymbol{a}_n, \qquad \widetilde{\boldsymbol{w}}^* = -\boldsymbol{a}_n, \qquad \widetilde{y} = -\boldsymbol{x}^\top \boldsymbol{a}_n.$$

Then $\widetilde{\boldsymbol{w}}_1 = 0$, $\|\widetilde{\boldsymbol{w}}^*\| < 1$, and $\widetilde{y} = \boldsymbol{x}^\top \widetilde{\boldsymbol{w}}^*$. The identity $\widetilde{\boldsymbol{w}}_t - \widetilde{\boldsymbol{w}}^* = \boldsymbol{w}_{t-1}$ preserves the residual dynamics and risks of both methods, including their prefix averages and common endpoint.

Theorem G.1 (Complete same-instance rank-one curve)**.** *For all sufficiently large n, both GD and fixed-dataset SGD on* (G.1) *satisfy, uniformly over integers $T \ge 1$,*

$$\mathbb{E}F(\bar{\boldsymbol{w}}_T) = \Theta\left(\frac{1}{n(1+T)^2} + pq + p\min\{(pqT)^2, 1\}\right). \tag{G.2}$$

Theorem G.1 follows from Theorem G.2 at $s = q$, using its initialization extension. The uniform proof below therefore establishes both results.

G.1 A CONTINUOUS FAMILY WITH A COMMON INITIALIZATION

Keep $M = \lfloor \sqrt{n} \rfloor$, $q = 64/M$, $p = 1 - q^{1/n}$, and $\boldsymbol{b} = M^{-1/2}\mathbf{1}$, but let $p \le s \le \sqrt{q}$. Define

$$\begin{aligned} \boldsymbol{v}_s &= (\sqrt{s}, -\boldsymbol{b}), \qquad \boldsymbol{u}_s = \frac{(1, \sqrt{s}\,\boldsymbol{b})}{1+s}, \\ a_s &= \frac{\sqrt{1+s}}{2}, \qquad \beta_s = \frac{\gamma}{\sqrt{1+s}}, \\ \boldsymbol{w}_0^{(s)} &= a_s \boldsymbol{u}_s + \beta_s \boldsymbol{v}_s. \end{aligned} \tag{G.3}$$

Replace $\boldsymbol{v}$ by $\boldsymbol{v}_s$ in (G.1). The resulting objective is

$$F_s(x, \boldsymbol{y}) = \frac{1}{6}\left[(\sqrt{s}x - \boldsymbol{b}^\top \boldsymbol{y})^2 + p\|\boldsymbol{y}\|^2\right].$$

Write $H_s := \nabla^2 F_s = (\boldsymbol{v}_s \boldsymbol{v}_s^\top + p\,\mathrm{diag}(0, I_M))/3$ for its population Hessian. Every sample remains a rank-at-most-one PSD contraction, and $\|\boldsymbol{w}_0^{(s)}\|^2 = 1/4 + \gamma^2$. These are deterministic initializations. They can be made identical across the family: on $\mathrm{span}\{\boldsymbol{e}_0, (0, \boldsymbol{b})\}$ choose the continuous orthogonal map R_s sending $\boldsymbol{u}_s/\|\boldsymbol{u}_s\|$ to $\boldsymbol{e}_0$ and $\boldsymbol{v}_s/\|\boldsymbol{v}_s\|$ to $(0, \boldsymbol{b})$, and extend it by the identity. Conjugating the sample Hessians by R_s sends every initialization to $\boldsymbol{e}_0/2 + \gamma(0, \boldsymbol{b})$ and preserves all risks and dynamics. Thus only the rank-one feature geometry varies.

The ratio has a coordinate-invariant interpretation: if V_g is the span of the gate features, then $s = \tan^2 \angle(\boldsymbol{v}_s, V_g)$, so s/q is a sample-size-calibrated geometric quantity. The common-energy comparison requires a large K even at the safe extremes of this family. Commuting with every gate projector makes a symmetric G diagonal in the root-and-gate basis. Commuting also with $\boldsymbol{v}_s \boldsymbol{v}_s^\top$, whose coordinates are all nonzero, forces G to be scalar. The population block on $\mathrm{span}\{\boldsymbol{e}_0, (0, \boldsymbol{b})\}$ is

$$\frac{1}{3}\begin{pmatrix} s & -\sqrt{s} \\ -\sqrt{s} & 1+p \end{pmatrix},$$

with determinant $ps/9$ and bounded positive trace, and transverse eigenvalue $p/3$. Thus $G \preceq H_s \preceq KG$ requires $K \ge \mathrm{cond}(H_s)$, where $\mathrm{cond}(H_s) = \Theta(1/(ps))$.

Theorem G.2 (Uniform coupling curve)**.** *Put $L = np$, $\rho = s/q$, $\vartheta = \rho/(1+\rho)^2$ and $\Lambda = p(s+q)$, and define $\Psi_{n,s}(T) := 1/[n(1+T)^2] + ps + p\vartheta \min\{(\Lambda T)^2, 1\}$. Absolute constants give, for all sufficiently large n, simultaneously for $p \le s \le \sqrt{q}$, $T \ge 1$, and GD or fixed-dataset SGD,*

$$\mathbb{E}F_s(\bar{\boldsymbol{w}}_T) = \Theta(\Psi_{n,s}(T)). \tag{G.4}$$

The same conclusion holds uniformly if $\boldsymbol{w}_0^{(s)}$ is replaced by $a\boldsymbol{u}_s + \beta \boldsymbol{v}_s$, with $a \in [1/3, 1]$ and $\beta \in [\gamma/2, 2\gamma]$.

Here the occupancy probability q is fixed while the coupling energy s varies independently. Since $Ls \le L\sqrt{q} \to 0$ and $L^2 q \to 0$,

$$\begin{aligned} F_s(\boldsymbol{w}_0^{(s)}) &= \Theta(n^{-1}), \\ \mathbb{E}F_s(\bar{\boldsymbol{w}}_n) &= \Theta(ps), \qquad \inf_{T \ge 1} \mathbb{E}F_s(\bar{\boldsymbol{w}}_T) = \Theta(ps), \\ \mathbb{E}F_s(\boldsymbol{w}_\infty) &= \Theta(p\vartheta). \end{aligned} \tag{G.5}$$

The last order holds uniformly at every $T \ge C/\Lambda$. In particular, $\sup_T \mathbb{E}F_s(\bar{\boldsymbol{w}}_T) = \Theta(n^{-1} + p\vartheta)$. Within this initialized family, safety at every deterministic horizon is therefore equivalent to $L\vartheta = O(1)$: its two branches are $\rho \lesssim L^{-1}$ and $\rho \gtrsim L$. Both lie inside the allowed range $p/q \le \rho \le q^{-1/2}$. The central region $L^{-1} \ll \rho \ll L$ has persistent $\omega(1/n)$ risk. The matching time scales are

$$T_{\mathrm{sat}} = \frac{1}{pq(1+\rho)}, \qquad T_\times = \frac{1}{pq\sqrt{L\rho}}, \qquad T_\times / T_{\mathrm{sat}} = (L\vartheta)^{-1/2}. \tag{G.6}$$

When $L\vartheta \to \infty$, for $T \ge n$ the regimes $T = o(T_\times)$, $T = \Theta(T_\times)$, and $T/T_\times \to \infty$ have respectively $o(1/n)$, $\Theta(1/n)$, and $\omega(1/n)$ risk. Equivalently, writing $B = L\rho/(1+\rho)^2 = L\vartheta$, the maximum risk takes the compact form

$$\sup_{T \ge 1} \mathbb{E}F_s(\bar{\boldsymbol{w}}_T) = \Theta\left(\frac{1+B}{n}\right). \tag{G.7}$$

The time scales in (G.6) describe a crossing when $B \to \infty$, in which case $T_{\mathrm{sat}}/T_{\times} = \Theta(\sqrt{B})$. The directed ratio ρ controls both magnitude and time: $\vartheta(\rho) = \vartheta(1/\rho)$ alone cannot determine the latter. Safety here concerns $\sup_T \mathbb{E}$, not $\mathbb{E}\sup_T$, pathwise monotonicity, or arbitrary initializations.

G.2 Uniform spectral and GD estimates

Let N_c, N_j count shared and gate observations, and put $h = 2N_c/n$, $a_j = N_j/n$, $V = \{j : a_j > 0\}$, $U = [M] \setminus V$, $Q = |U|/M$, $B_k = \sum_{j\in V} b_j^2/a_j^k$ for $k = 1, 2$. Here $D_S := \mathrm{diag}(a_1, \dots, a_M)$, $D_V := \mathrm{diag}(a_j : j \in V)$, $a_{\max} := \max_{j\in[M]} a_j$, and $a_{\min} := \min_{j\in V} a_j$ when $V \neq \varnothing$. We use the zero-extension notation diag_V defined in Appendix F. Lemma F.2, including its Poissonization proof, applies to the gate counts after merging the shared and zero categories. Its stronger zero-count estimate is $Q \le (64np)^{-1}$. Since $\sqrt{q} = o((np)^{-1})$, that event intersected with $1/2 \le h \le 3/2$ gives an event E, independent of s, on which

$$\begin{aligned} a_{\max} \le 16p, \quad (8p)^{-1} \le B_1 \le 4/p, \quad B_2 \le 12/p^2, \quad s + Q \le (32np)^{-1}, \\ \mathbb{P}(E^c) = O(M^{-10}), \qquad \mathbb{P}(E \cap G) \ge c > 0, \quad G = \{q/2 \le Q \le 2q\}. \end{aligned} \tag{G.8}$$

No independence between h and the gate counts is used. Also $\mathbb{E}Q = q$, and, for every $s > 0$,

$$\mathbb{E}\frac{sQ}{(s+Q)^2} = \Theta(\vartheta), \qquad \vartheta = \frac{sq}{(s+q)^2}. \tag{G.9}$$

The lower bound uses G. For $s \ge q$, the upper bound is $q/s \le 4\vartheta$. For $s \le q$, use the discrete fact $Q > 0 \Rightarrow Q \ge q/64$, which bounds the integrand by $64s/q \le 256\vartheta$. The case $Q = 0$ contributes zero. Consequently,

$$\mathbb{E}[sQ \min\{t^2, (s+Q)^{-2}\}] \le C\vartheta \min\{[(s+q)t]^2, 1\}. \tag{G.10}$$

This estimate also controls atypical occupancy patterns.

Fix $S \in E$ and write $A_h = 3H_S = h\boldsymbol{v}_s\boldsymbol{v}_s^\top + \mathrm{diag}(0, D_S)$, $P = I - A_h/12$, $P^t\boldsymbol{e}_0 = (x_t, \boldsymbol{y}_t)$ and $r_t = \boldsymbol{v}_s^\top P^t \boldsymbol{e}_0$. Set $d = s+Q$ and $\kappa = Q/d$. The same secular calculation as in Lemma F.3, retaining s and Q separately, gives one low positive eigenvalue and its weight:

$$\lambda\left(1/h + \sum_{j\in V} \frac{b_j^2}{a_j - \lambda}\right) = d, \qquad pd/10 \le \lambda \le 8pd \le a_{\min}/4, \tag{G.11}$$

$$\omega = \frac{s}{d + \lambda^2 B_2(\lambda)}, \qquad W = 1 - \kappa - \omega \le 3072s. \tag{G.12}$$

Indeed $B_2(\lambda) \le 48/p^2$ and $W = s\lambda^2 B_2(\lambda)/[d(d + \lambda^2 B_2(\lambda))]$. Interlacing puts the remaining positive eigenvalues at least at $a_{\min}$. The $\boldsymbol{e}_0$ kernel weight is κ. The nonnegative spectral expansion of r_t shows that its slow contribution to hr_t is at most $8p\sqrt{s}$, whereas the time sum of its fast contribution is $12W/\sqrt{s}$. The coordinate recursion $y_{j,t+1} = (1 - a_j/12)y_{j,t} + hb_j r_t/12$ therefore gives

$$\sup_t y_{j,t} \le \sqrt{s}\, b_j(8p/a_j + 3072), \qquad \sum_{j\in V} \sup_t y_{j,t}^2 \le Cs. \tag{G.13}$$

Moreover $r_t, \boldsymbol{y}_t \ge 0$, x_t is nonincreasing, and $\boldsymbol{b}^\top \boldsymbol{y}_t \le \sqrt{s}$.

Although A_h is random, $A_h\boldsymbol{e}_0 = h\sqrt{s}\,\boldsymbol{v}_s$, so the prepared vector $\boldsymbol{u}_s = (I - A_h/[h(1+s)])\boldsymbol{e}_0$ is deterministic. Its trajectory obeys

$$\begin{aligned} x_t^u &= x_t - \sqrt{s}\, r_t/(1+s), \\ y_{j,t}^u &= (1 - a_j/[h(1+s)])y_{j,t} + b_j r_t/(1+s), \\ r_t^u &= \boldsymbol{b}^\top D_S \boldsymbol{y}_t/[h(1+s)]. \end{aligned} \tag{G.14}$$

The displayed coordinate coefficients are nonnegative on E, and

$$\sum_{j\in V} \sup_t (y_{j,t}^u)^2 \le Cs, \qquad \sup_t (r_t^u)^2 \le Cp^2 s. \tag{G.15}$$

For $g_T(z) = T^{-1}\sum_{t<T}(1-z/12)^t$, spectral cancellation $\omega\lambda \le 8ps$ gives $1-\bar{x}_T \le Cs + CpsT$. The kernel projection also gives $1-\bar{x}_T \le s/d$. The prepared correction is at most s, so the unseen invariant yields

$$\|\bar{\boldsymbol{y}}^u_{U,T}\|^2 \le CQs + CsQ\min\{(pT)^2, d^{-2}\}. \tag{G.16}$$

Equations (G.10) and (G.15) prove the prepared GD upper bound $C[ps + p\vartheta\min\{(\Lambda T)^2, 1\}]$, up to the rare complement of E.

A different argument is needed for the uniform lower floor when $s \gg q$. Minimizing the population risk over $\boldsymbol{y}$ at fixed x gives $F_s(x,\boldsymbol{y}) \ge psx^2/[6(1+p)]$. If the unseen invariant is $\boldsymbol{y}_U = \boldsymbol{b}_U(a-x)/\sqrt{s}$, it also gives $F_s(x,\boldsymbol{y}) \ge pQ(a-x)^2/(6s)$. Half their sum, minimized over x, is

$$F_s(x,\boldsymbol{y}) \ge \frac{pa^2}{12}\frac{sQ}{s^2+(1+p)Q} \ge cps \quad \text{on } G, \quad s^2 \le q,\ a \in [1/3,1]. \tag{G.17}$$

The unseen-coordinate bound supplies the lower floor when the kernel x-coordinate is small.

For the general initialization $a\boldsymbol{u}_s + \beta\boldsymbol{v}_s$ in the theorem, set $c_s = a/(1+s) - \beta/\sqrt{s}$. Since $s \ge p \gg 1/n$, it is uniformly positive and

$$\boldsymbol{w}_t^{\mathrm{new}} = aP^t\boldsymbol{e}_0 - (c_s/h)A_hP^t\boldsymbol{e}_0, \qquad y_{j,t}^{\mathrm{new}} = (a - c_sa_j/h)y_{j,t} + c_sb_jr_t. \tag{G.18}$$

On E, the unseen coordinates dominate $a\boldsymbol{y}_{U,t}$, the seen envelope is Cs, and $\boldsymbol{y}_U^{\mathrm{new}} = \boldsymbol{b}_U(a - x^{\mathrm{new}})/\sqrt{s}$. Equation (G.17) gives the floor. For $T \ge 2$,

$$1-\bar{x}_T \ge \frac{s}{d}(1-g_T(\lambda)) \ge c\frac{s}{d}\min\{pdT, 1\}. \tag{G.19}$$

On $E \cap G$ the unseen risk is consequently at least $cp\vartheta\min\{(\Lambda T)^2, 1\}$. At $T = 1$, that term is absorbed by ps. The fast GD trajectory $\boldsymbol{w}_t^v := P^t\boldsymbol{v}_s$ satisfies

$$\bar{\boldsymbol{w}}_T^v = \frac{12}{h\sqrt{s}\,T}(I-P^T)\boldsymbol{e}_0, \qquad \frac{1+s}{T} \le \boldsymbol{v}_s^\top\bar{\boldsymbol{w}}_T^v \le \frac{12}{hT}. \tag{G.20}$$

Its upper population risk is C/T^2 because $p \le s$. The combined residual $ar_t^u + \beta\boldsymbol{v}_s^\top P^t\boldsymbol{v}_s$ is a nonnegative sum, so the fast term also gives $c\gamma^2/T^2$ from below. Quadratic subadditivity for the upper bound and the maximum of these three lower bounds establish (G.4) for GD.

G.3 A UNIFORM UPPER BOUND FOR FIXED-DATA SGD

Throughout this subsection, we condition on a fixed dataset $S \in E$. For a deterministic initial vector $\boldsymbol{w}_{\mathrm{init}}$, let $\boldsymbol{\mu}_t = P^t\boldsymbol{w}_{\mathrm{init}}$, $\boldsymbol{\xi}_t = \boldsymbol{w}_{t+1} - P\boldsymbol{w}_t$, and $r_t = \boldsymbol{v}_s^\top\boldsymbol{w}_t$. Conditional on the filtration before draw i_t, $\mathbb{E}[\boldsymbol{\xi}_t \mid \mathcal{F}_t] = 0$, and categorical exclusivity gives

$$\mathbb{E}[\boldsymbol{\xi}_t\boldsymbol{\xi}_t^\top \mid \mathcal{F}_t] \preceq \frac{1}{144}\{2hr_t^2\boldsymbol{v}_s\boldsymbol{v}_s^\top + \mathrm{diag}_V(a_jy_{j,t}^2)\}. \tag{G.21}$$

The omitted term is $(A_h\boldsymbol{w}_t)(A_h\boldsymbol{w}_t)^\top/144 \succeq 0$. Write $m_y = \sum_{j\in V}\sup_t\mu_{j,t}^2$, $m_r = \sup_t(\boldsymbol{v}_s^\top\boldsymbol{\mu}_t)^2$, $M_j = \sup_t\mathbb{E}y_{j,t}^2$ for $j \in V$, $M_y = \sum_{j\in V}M_j$, and $M_r = \sup_t\mathbb{E}r_t^2$. Radius nonincrease makes these finite. For $\Gamma_t := \mathrm{Cov}(\boldsymbol{w}_t \mid S)$, the covariance recursion gives

$$\Gamma_t \preceq \bar{\Gamma} = \frac{1}{144}\sum_{m\ge0}P^m[2hM_r\boldsymbol{v}_s\boldsymbol{v}_s^\top + \mathrm{diag}_V(a_jM_j)]P^m. \tag{G.22}$$

All inputs lie in $(\ker A_h)^\perp$, and the exact identities are $(A_h^+)_{VV} = D_V^{-1}$ and $\boldsymbol{v}_s^\top A_h^+\boldsymbol{v}_s = 1/h$. For the first, solve $A_h\boldsymbol{w} = \boldsymbol{z}$ with $\boldsymbol{z}$ supported on V: its root coordinate forces the residual to vanish and its seen coordinates then force $\boldsymbol{w}_V = D_V^{-1}\boldsymbol{z}$. Orthogonality to the kernel determines the other coordinates. The second identity follows from $A_h\boldsymbol{e}_0 = h\sqrt{s}\boldsymbol{v}_s$. Since $I - P^2 \succeq 7A_h/48$,

$$\sum_m\|P^m\boldsymbol{e}_j\|^2 \le \frac{48}{7a_j}, \quad \sum_m\boldsymbol{e}_j^\top P^mA_hP^m\boldsymbol{e}_j \le \frac{48}{7}, \quad \sum_m\|P^m\boldsymbol{v}_s\|^2 \le \frac{48}{7h}. \tag{G.23}$$

These estimates require no sample-independent empirical spectral gap. Taking seen diagonals and the residual form in (G.22) yields

$$\begin{aligned} M_y &\le m_y + 2M_r/21 + M_y/21, \\ M_r &\le m_r + 2(1+s)M_r/21 + a_{\max}M_y/(21h). \end{aligned} \tag{G.24}$$

Using $h \geq 1/2$, $s \leq 1/2$ and $a_{\max} \leq 1$, we can absorb the right-hand terms to obtain $M_y \leq 9m_y/8 + m_r/8$ and $M_r \leq 2m_r + a_{\max} m_y$. The same series give

$$\operatorname{tr} \bar{\Gamma} \leq (2M_r + M_y)/21, \qquad \operatorname{tr}(A_h \bar{\Gamma}) \leq [2h(1+s)M_r + \sum_{j \in V} a_j M_j]/21. \tag{G.25}$$

As $H_s \preceq (A_h/h + pI)/3$, the population variance cost is at most $C[M_r + (a_{\max} + p)M_y]$. Convexity gives the same bound for a prefix average of the centered iterates, regardless of their dependence across time.

Prepared component. For $\boldsymbol{u}_s$, (G.15) gives $m_y = O(s)$, $m_r = O(p^2 s)$, so this extra risk is $O(ps)$ uniformly in time.

Fast component. Denote the deterministic auxiliary trajectory by $(x_t^0, \boldsymbol{y}_t^0) := P^t \boldsymbol{e}_0$ and its residual by $r_t^0 := \boldsymbol{v}_s^\top P^t \boldsymbol{e}_0$. For initialization $\boldsymbol{v}_s$, the identity $y_{j,t}^v = -b_j r_t^0/\sqrt{s} + a_j y_{j,t}^0/(h\sqrt{s})$ and (G.13) give $m_y, m_r = O(1)$ and hence $M_y, M_r = O(1)$. Near the valley, we need a sharper bound on the accumulated shared noise. For $A_i := 3H_i$, the inequality $A_i^2 \preceq 3A_i$ gives the conditional norm decrease $\mathbb{E}[\|\boldsymbol{w}_{t+1}\|^2 \mid \mathcal{F}_t] \leq \|\boldsymbol{w}_t\|^2 - (7/48)\boldsymbol{w}_t^\top A_h \boldsymbol{w}_t$. Consequently,

$$\sum_{t \geq 0} \mathbb{E} r_t^2 \leq \frac{48}{7h} \|\boldsymbol{v}_s\|^2 \leq C, \qquad \|K_l \boldsymbol{v}_s\|_{H_s}^2 \leq \frac{48}{h^2}(1 + p/s) \leq C, \quad K_l = \sum_{m=0}^{l} P^m. \tag{G.26}$$

Here $\|\boldsymbol{v}\|_{H_s}^2 := \boldsymbol{v}^\top H_s \boldsymbol{v}$. For the second bound use $K_l \boldsymbol{v}_s = 12(I - P^{l+1})\boldsymbol{e}_0/(h\sqrt{s})$, $\boldsymbol{v}_s^\top K_l \boldsymbol{v}_s \leq 12/h$, and $p \leq s$. The exact centered average is $T^{-1} \sum_{t=0}^{T-2} K_{T-2-t} \boldsymbol{\xi}_t$. Cross covariances vanish by conditioning before the later draw. Equations (G.21) and (G.26) bound its shared-source variance by C/T^2. For the diagonal source, for every t, j use

$$\left\| \frac{1}{T} \sum_{r=t+1}^{T-1} H_s^{1/2} P^{r-1-t} \boldsymbol{e}_j \right\|^2 \leq \frac{1}{T} \sum_{r=t+1}^{T-1} \|H_s^{1/2} P^{r-1-t} \boldsymbol{e}_j\|^2.$$

After multiplication by $a_j \mathbb{E} y_{j,t}^2/288$ and reordering sums, its cost is bounded by the diagonal-input part of $\operatorname{tr}(H_s \bar{\Gamma})/2$, at most $C(a_{\max} + p)M_y \leq Cp$. The two contributions come from decomposing the PSD upper bound. The underlying noise components need not be independent. Including the mean gives

$$\mathbb{E}_{\text{alg}}[F_s(\bar{\boldsymbol{w}}_T^{v,\text{SGD}}) \mid S] \leq C(T^{-2} + p). \tag{G.27}$$

Combining the components. For the same sampled indices, linearity and quadratic subadditivity combine the prepared and fast bounds. Since $\gamma^2 p \leq ps$, the result is the upper curve (G.4) on E. On E^c, the initial radius is uniformly bounded, and $M^{-10} = o(p^2) = o(\min_{s \geq p} ps)$, uniformly in T. Conditional Jensen transfers the GD lower bound to SGD. Together with the upper bound, this proves Theorem G.2 and its specialization Theorem G.1.

G.4 Exact critical windows and the occupancy geometry

For the normalized initialization (G.3), the preceding estimates give the sharper, additive approximation

$$\left| \mathbb{E} F_s(\bar{\boldsymbol{w}}_T) - \frac{p(1+s)}{24} \mathbb{E}\left[\mathbf{1}_E \frac{sQ}{(s+Q)^2} (1 - g_T(\lambda))^2 \right] \right| \leq C\{ps + \gamma^2/T^2 + M^{-10}\}. \tag{G.28}$$

Indeed $W \leq Cs$ implies $1 - \bar{x}_T^0 = (s/d)(1 - g_T(\lambda)) + O(s)$. The combined unseen displacement is $a_s(1 - \bar{x}_T^0) + c_s\sqrt{s}\,\bar{r}_T^0$, with second term $O(s)$. Its squared unseen norm is therefore $a_s^2 sQ d^{-2}(1 - g_T(\lambda))^2 + O(s)$: the cross error is at most CsQ/d and the squared error at most CsQ. Seen and residual costs are $O(ps + \gamma^2/T^2)$. For SGD, let $\boldsymbol{Z}^u, \boldsymbol{Z}^v$ denote the centered averaged outputs on the same index path. Its centered output is exactly $a_s \boldsymbol{Z}^u + \beta_s \boldsymbol{Z}^v$, whose population second moment is at most twice the sum of the two component second moments. The established bounds make this $O(ps + \gamma^2/T^2 + \gamma^2 p)$. This is an additive error, so the leading GD mean in (G.28) keeps its exact coefficient.

Let $K = MQ$ and $Z_n = Q/q = K/64$. Write $(k)_r := k(k-1)\cdots(k-r+1)$ for the falling factorial, with $(k)_0 := 1$. For fixed nonnegative integers r, $\mathbb{E}(K)_r = (M)_r(1 - rp)^n \to 64^r$, since

$(1-p)^n = q$ and $np^2 \to 0$. Hence $K \Rightarrow K_\infty \sim \mathrm{Pois}(64)$ and $Z_n \Rightarrow Z = K_\infty/64$. Also $\mathbb{E}Z_n = 1$ and $\mathbb{E}Z_n^2 \le 1 + 1/64$. We need the more precise spectral estimate

$$pB_1 \longrightarrow 1, \qquad \lambda/[p(s+Q)] \longrightarrow 1 \quad \text{in probability.} \tag{G.29}$$

For the first limit, a single gate count has mean $L = np$. On $N_j \ge L/2$, $|L/N_j - 1| \le 2|N_j - L|/L$, of mean $O(L^{-1/2})$. The remaining nonzero counts cost at most $(L+1)e^{-cL}$ and zero counts cost q. Averaging proves L^1 convergence without independence. For the second, on E, $\lambda/a_{\min} \le 8L(s+Q) = o_{\mathbb{P}}(1)$, since $Ls \to 0$ and $\mathbb{E}LQ = Lq \to 0$. Thus the shifted reciprocal sum is asymptotic to B_1, and (G.11) proves the claim.

Theorem G.3 (Two critical-window limits)**.** *Let $\mathcal{G}(z) = (1-e^{-z})/z$ for $z > 0$, with $\mathcal{G}(0) = 1$. For any fixed $a, \tau > 0$, both GD and fixed-data SGD satisfy the following limits as $n \to \infty$.*

Weak-coupling window. *For $\rho = a/L$ and $T = \lfloor \tau/(pq) \rfloor$,*

$$n\mathbb{E}F_s(\bar{\boldsymbol{w}}_T) \longrightarrow \frac{a}{24}\mathbb{E}\left[\frac{(1-\mathcal{G}(\tau Z/12))^2}{Z}; Z > 0\right]. \tag{G.30}$$

Strong-coupling window. *For $\rho = aL$ and $T = \lfloor \tau/(ps) \rfloor$,*

$$n\mathbb{E}F_s(\bar{\boldsymbol{w}}_T) \longrightarrow \frac{1}{24a}(1-\mathcal{G}(\tau/12))^2. \tag{G.31}$$

Proof. Both parameter sequences eventually lie in $p \le s \le \sqrt{q}$. In the weak window, the normalized weight in (G.28), apart from the factor $1 + s \to 1$, satisfies

$$0 \le \frac{aZ_n}{(a/L + Z_n)^2} \le 64a,$$

because positive $Z_n \ge 1/64$. Its distributional limit equals a/Z on $Z > 0$ and zero at $Z = 0$. Equation (G.29) gives $\lambda T/12 \Rightarrow \tau Z/12$, and the discrete prefix filter converges to $\mathcal{G}$. Joint distributional convergence and uniform boundedness, using the lattice gap at zero, prove (G.30).

In the strong window, the weight $Z_n/[a(1+Z_n/(aL))^2]$ is uniformly integrable by its second-moment bound. We have

$$\mathbb{E}\left[\frac{Z_n}{a(1+Z_n/(aL))^2}\right] \longrightarrow \frac{1}{a}, \qquad \lambda T/12 \longrightarrow \tau/12 \quad \text{in probability,}$$

which proves (G.31). The remainder in (G.28), multiplied by n, vanishes: $nps = aq$ and aL^2q, respectively. The fast and rare-event terms also vanish. Thus both GD and fixed-dataset SGD have the stated limits. □

Only fixed $\tau > 0$ is asserted: the original fast transient gives a boundary layer at $\tau = 0$. In the strong window each dataset still has random amplitude Z. The time kernel becomes deterministic, and taking expectation uses $\mathbb{E}Z = 1$.

Finally, when $N_c > 0$, projection onto the empirical kernel gives

$$\mathbb{E}F_s(\boldsymbol{w}_\infty) = \frac{p(1+s)}{24}\mathbb{E}\frac{sQ}{(s+Q)^2} + O(2^{-n}). \tag{G.32}$$

Indeed the projected initialization is $a_s(Q/d, \sqrt{s}\,\boldsymbol{b}_U/d, 0_V)$. The event $N_c = 0$, of probability 2^{-n}, is handled by the radius bound. The case $Q = 0$ contributes zero. Both algorithms converge to this projection by Lemma C.9. For $\Pi = P_{\ker H_S}$ and $E_{00} = \boldsymbol{e}_0\boldsymbol{e}_0^\top$ on $N_c > 0$,

$$\|[\Pi, E_{00}]\|_{\mathrm{op}}^2 = \kappa(1-\kappa) = \frac{sQ}{(s+Q)^2},$$

so the endpoint is exactly the risk cost of this kernel-direction mixing, up to the stated rare event. The critical endpoint limits are $a\,64\mathbb{E}[1/K_\infty; K_\infty > 0]/24$ on the weak side and $1/(24a)$ on the strong side. The inverse occupancy moment on the weak side shows why the common magnitude descriptor does not determine the time mixture.

H STATISTICAL SAFETY THROUGH A COMPARISON METRIC

This appendix proves Theorem 4.4 using a common quadratic energy. The sufficient condition permits sample Hessians that do not commute with the population covariance, including in rank-one regression. It does not characterize all safe sampling laws. Throughout, $H = \mathbb{E}H_z$, $H_S = n^{-1}\sum_i H_i$, $F(\boldsymbol{w}) = \boldsymbol{w}^\top H\boldsymbol{w}/2$, $\boldsymbol{w}_0 = \boldsymbol{u}$ is deterministic, and $\bar{\boldsymbol{w}}_T = T^{-1}\sum_{t<T}\boldsymbol{w}_t$. Both GD and fixed-dataset SGD use $0 < \eta \le 1$. Let $a_\eta := \eta(2-\eta)$, so $a_\eta = 7/16$ at $\eta = 1/4$.

H.1 A COMMUTING COMPARISON METRIC

Theorem H.1 (Safety through a comparison metric)**.** *Let $0 \preceq H_z \preceq I$. Suppose there is a deterministic PSD matrix G and a constant $K \ge 1$ such that $GH_z = H_zG$ almost surely and $G \preceq H \preceq KG$. Then an absolute constant C gives, for both algorithms,*

$$\mathbb{E}\sup_{s\ge t} F(\boldsymbol{w}_s) \le \frac{CK\,\|\boldsymbol{u}\|^2}{a_\eta}\left(\frac{1}{t+1}+\frac{1}{n}\right), \qquad t \ge 0, \tag{H.1}$$

$$\mathbb{E}\sup_{m\ge T} F(\bar{\boldsymbol{w}}_m) \le \frac{CK\,\|\boldsymbol{u}\|^2}{a_\eta}\left(\frac{1}{T}+\frac{1}{n}\right), \qquad T \ge 1. \tag{H.2}$$

The chosen late horizon may depend on the dataset and the full index path. Neither population risk nor the risk of successive prefix averages is asserted to decrease on every path. A comparable quadratic energy decreases and supplies the uniform bounds.

Proof. Define $V(\boldsymbol{w}) := \boldsymbol{w}^\top G\boldsymbol{w}$ and $R := \|\boldsymbol{u}\|$. The claim is immediate when $R = 0$, so assume $R > 0$. Because G commutes with H_z and $H_z^2 \preceq H_z$,

$$(I-\eta H_z)G(I-\eta H_z) = G - 2\eta GH_z + \eta^2 GH_z^2 \preceq G - a_\eta GH_z \preceq G. \tag{H.3}$$

Every update also contracts Euclidean norm. Commutation passes to the expectation H, so $G \preceq H$ implies $GH \succeq G^2$. For a fresh observation independent of the current iterate, therefore,

$$\mathbb{E}[V(\boldsymbol{w}') \mid \boldsymbol{w}] \le V(\boldsymbol{w}) - a_\eta \boldsymbol{w}^\top GH\boldsymbol{w} \le V(\boldsymbol{w}) - a_\eta\,\|G\boldsymbol{w}\|^2 \le V(\boldsymbol{w}) - a_\eta V(\boldsymbol{w})^2/R^2. \tag{H.4}$$

The last step is Cauchy–Schwarz and $\|\boldsymbol{w}\| \le R$.

Condition on an entire index sequence and reveal each training matrix at its first visit. This conditioning is independent of the data. Let D_t count the different indices among the first t updates. Between first visits V cannot increase by (H.3). If b_k, c_k are its expected values immediately before and after the kth first visit, then $b_k \le c_{k-1}$ and Jensen's inequality in (H.4) gives $c_k \le b_k - a_\eta b_k^2/R^2$. For positive means take reciprocals to obtain $c_k^{-1} \ge b_k^{-1} + a_\eta/R^2 \ge c_{k-1}^{-1} + a_\eta/R^2$. If a mean is zero, all later means are zero. Since $V(\boldsymbol{u}) \le R^2$ by $G \preceq H \preceq I$,

$$\mathbb{E}_S[V(\boldsymbol{w}_t) \mid \text{indices}] \le \frac{R^2}{1+a_\eta D_t}. \tag{H.5}$$

The required occupancy estimate is

$$\mathbb{E}\frac{1}{1+D_t} \le C\left(\frac{1}{t+1}+\frac{1}{n}\right). \tag{H.6}$$

Indeed, for $n \ge 2$ and $t \ge 1$, $\mu = \mathbb{E}D_t = n[1-(1-1/n)^t] \ge (1-e^{-1})\min\{t,n\}$. The indicators that two different indices have been visited have nonpositive covariance, since $(1-2/n)^t \le (1-1/n)^{2t}$. Thus $\mathrm{Var}(D_t) \le \mu$. Splitting according to $D_t \ge \mu/2$ and applying Chebyshev gives $\mathbb{E}(1+D_t)^{-1} \le \min\{1, 6/\mu\}$. The cases $n = 1$ and $t = 0$ are immediate. Combining (H.6) with $0 < a_\eta \le 1$ yields

$$\mathbb{E}V(\boldsymbol{w}_t) \le \frac{CR^2}{a_\eta}\left(\frac{1}{t+1}+\frac{1}{n}\right). \tag{H.7}$$

For $s \ge t$, $F(\boldsymbol{w}_s) \le KV(\boldsymbol{w}_s)/2 \le KV(\boldsymbol{w}_t)/2$ pathwise, which proves (H.1). The sequence $v_t = \sqrt{V(\boldsymbol{w}_t)}$ is nonincreasing. Thus, for all $m \ge T$,

$$\sqrt{F(\bar{\boldsymbol{w}}_m)} \le \sqrt{K/2}\,\frac{1}{m}\sum_{t<m} v_t \le \sqrt{K/2}\,\frac{1}{T}\sum_{t<T} v_t.$$

Minkowski's inequality, (H.7), and $T^{-1}\sum_{t<T}(t+1)^{-1/2} \le 2/\sqrt{T}$ prove (H.2) for SGD. Conditional Jensen transfers (H.7) to GD, whose updates contract V because $H_S G = G H_S$. The same two envelopes prove both GD bounds. This argument uses independence only at first visits. □

Rank-one interpretation and a continuous example. For rank-one $H_z = \boldsymbol{x}\boldsymbol{x}^\top$, commutation with G says that $\boldsymbol{x}$ lies inside an eigenspace of G. It does not say that $\boldsymbol{x}$ is an eigenvector of H. In particular, suppose the support splits across orthogonal subspaces E_b, every feature lies in one such subspace, and the unconditional population blocks satisfy $\lambda_b I_{E_b} \preceq H_b \preceq K\lambda_b I_{E_b}$. Then $G = \bigoplus_b \lambda_b I_{E_b}$ satisfies the conditions of Theorem H.1. The relative scales of the blocks and their total dimension are unrestricted.

For an explicit continuous distribution, choose a two-dimensional block b with arbitrary probability $\pi_b > 0$, and in that block take $\boldsymbol{x} = (\cos\theta, \sin\theta)$ with density $(1+\cos(2\theta))/(2\pi)$ on $[0, 2\pi)$. Direct integration gives $\mathbb{E}[\cos^2\theta] = 3/4$, $\mathbb{E}[\sin^2\theta] = 1/4$, and $\mathbb{E}[\sin\theta\cos\theta] = 0$. Hence $H_b = \pi_b \operatorname{diag}(3/4, 1/4)$ and $G_b = (\pi_b/4) I_2$ give $K = 3$. Every sample has unit norm and rank one. Except on a null set, its Hessian does not commute with H. Arbitrarily many blocks allow $\operatorname{rank} H > n$, and arbitrary block probabilities allow arbitrarily large global condition number. The theorem assumes bounded anisotropy within invariant components. It does not cover an arbitrarily anisotropic irreducible sample family.

I Recovery under weak curvature

This appendix proves the weak-curvature robustness results discussed in Section 4. A small positive curvature preserves a finite high-risk plateau but eventually restores zero risk, at different rates for last iterates and prefix averages.

Throughout this appendix, H_z is an i.i.d. PSD contraction, $H = \mathbb{E}H_z$, $H_S = n^{-1}\sum_{i=1}^n H_i$, and $F(\boldsymbol{w}) = \boldsymbol{w}^\top H \boldsymbol{w}/2$. The common minimizer is zero and the initialization $\boldsymbol{w}_{\text{init}}$ is deterministic. We index the initial point by $\boldsymbol{w}_0 = \boldsymbol{w}_{\text{init}}$ and average $\boldsymbol{w}_0, \dots, \boldsymbol{w}_{T-1}$. This is the same output as in the main text. Except where a general step size is explicitly stated, $\eta = 1/4$.

I.1 Near-zero curvature: a sharp late plateau and recovery law

Use the rank-two categorical loss (F.2) with parameters (F.1) and initial vector $\boldsymbol{w}_{\text{init}} = \boldsymbol{u}/2 + \gamma\boldsymbol{v}$ from (F.3). Retain the count and diagonal-matrix notation of Appendix F. For $\varepsilon > 0$, define

$$H_z^{(\varepsilon)} = \frac{H_z + \varepsilon I}{1+\varepsilon}, \quad H^{(\varepsilon)} = \frac{H + \varepsilon I}{1+\varepsilon}, \quad F_\varepsilon(\boldsymbol{w}) = \tfrac{1}{2}\boldsymbol{w}^\top H^{(\varepsilon)} \boldsymbol{w}. \tag{I.1}$$

This is a perturbation of the problem distribution around its common minimizer, which is unknown in the original coordinates. It is not proposed as a directly implementable regularizer.

Theorem I.1 (From a permanent plateau to delayed recovery)**.** *For all sufficiently large n, set $t_0 := \lceil 160/(pq) \rceil$. Fix $0 < \varepsilon \le pq/1000$ and $\eta = 1/4$, and define $d_\varepsilon := \eta\varepsilon/(1+\varepsilon)$ and $r_\varepsilon := 1 - d_\varepsilon$. For GD and fixed-dataset SGD on* (I.1)*, there are absolute constants $c, C > 0$ such that*

$$cp\, r_\varepsilon^{2t} \le \mathbb{E}F_\varepsilon(\boldsymbol{w}_t^{(\varepsilon)}) \le Cp\, r_\varepsilon^{2t}, \qquad t \ge t_0, \tag{I.2}$$

$$cp\left(\frac{1 - r_\varepsilon^T}{T d_\varepsilon}\right)^2 \le \mathbb{E}F_\varepsilon(\bar{\boldsymbol{w}}_T^{(\varepsilon)}) \le Cp\left(\frac{1 - r_\varepsilon^T}{T d_\varepsilon}\right)^2, \qquad T \ge 2t_0. \tag{I.3}$$

The distribution, perturbation, and initialization are fixed for all horizons. In particular the averaged risk is $\Theta(p\min\{1, (\varepsilon T)^{-2}\})$ in this late-time range. Its high-risk plateau lasts from order $(pq)^{-1}$ to order ε^{-1}. The last iterate then recovers exponentially, whereas the prefix average recovers quadratically in T^{-1}. For every $\varepsilon > 0$ the eventual risk is zero. The nonzero permanent plateau occurs at $\varepsilon = 0$.

Proof. Define $\eta' := \eta/[1 + (1-\eta)\varepsilon]$. For every dataset and index path,

$$I - \eta H_i^{(\varepsilon)} = r_\varepsilon(I - \eta' H_i), \qquad \boldsymbol{w}_t^{(\varepsilon)} = r_\varepsilon^t \boldsymbol{w}_t^{\eta'}. \tag{I.4}$$

The same identity holds for GD. Since $\varepsilon \le 1$, $\eta/2 \le \eta' \le \eta$. We establish all auxiliary estimates uniformly over this step-size interval.

Write $A_0 = 3H_S = \boldsymbol{v}\boldsymbol{v}^\top + \mathrm{diag}(0, D_S)$, $P = I - \zeta A_0$ and $\zeta = \eta'/3 \in [1/24, 1/12]$. For the auxiliary trajectory $P^t \boldsymbol{e}_0$, all residuals r_t and coordinates $y_{j,t}$ are nonnegative, and x_t is nonincreasing. The original unseen-coordinate identity and the telescoping calculation give

$$1 - \bar{x}_T \ge \frac{q}{q+Q}\left[1 - \frac{1+B_1}{\zeta T(q+Q)}\right]. \tag{I.5}$$

On $E \cap G$ from Lemma F.2, $B_1 \le 4/p$ and $q/2 \le Q \le 2q$. Hence $1 + B_1 \le 5/p$, and the fraction subtracted in (I.5) is at most $80/(pqT)$. At $T = t_0$ this is at most $1/2$. Monotonicity therefore implies $1 - x_t \ge 1/6$ for every $t \ge t_0$. The prepared-start identity, whose coefficients are independent of ζ, gives

$$\boldsymbol{y}_{U,t}^{\eta'} \ge \frac{1}{2}\boldsymbol{y}_{U,t}^{\boldsymbol{e}_0} \ge \frac{\boldsymbol{b}_U}{12\alpha}, \qquad t \ge t_0, \tag{I.6}$$

componentwise. Also $\|\boldsymbol{b}_U\|^2/\alpha^2 = Q/q \ge 1/2$ on this event. For large n, $\Pr(E \cap G) \ge 7/8$.

To bound the last-iterate risk uniformly, first observe that on E the deterministic prepared trajectory satisfies

$$0 \le r_t^u = \frac{\boldsymbol{b}^\top D_S \boldsymbol{y}_t}{1+q} \le 16p\alpha, \qquad 0 \le \boldsymbol{v}^\top P^t \boldsymbol{v} \le 1 + q.$$

Together with $\|\boldsymbol{w}_t\| \le 1$, its population risk is at most $C(p + p^2 q + 1/n) \le Cp$. These estimates are uniform in ζ. The observed-coordinate envelope used for SGD is likewise uniform: the slow-mode residual bound is unchanged, while the sum of fast residuals is $W/(\zeta\alpha)$ and enters the coordinate recurrence multiplied by ζ. Thus the cancellation in the proof of Lemma F.3 gives $\sum_{j\in V} \sup_t (y_{j,t}^{\eta',\mathrm{GD}})^2 \le Cq$.

For the SGD covariance estimate, condition on S. The driving noise is supported on observed coordinates, and its covariance is bounded by $\zeta^2 \,\mathrm{diag}_V(a_j \mathbb{E} y_{j,t}^2)$. Since $\|A_0\| \le 3$,

$$\sum_{m\ge 0} P^{2m} \preceq \frac{1}{\zeta(2-3\zeta)} A_0^+, \qquad \sum_{m\ge 0} P^m A_0 P^m \preceq \frac{1}{\zeta(2-3\zeta)} I$$

on the positive subspace. The identity $(A_0^+)_{VV} = D_V^{-1}$ therefore bounds each covariance-feedback column sum by $\zeta/(2-3\zeta) \le 1/21$. The same bootstrap as in Lemma F.5 bounds $\sum_{j\in V} \sup_t \mathbb{E} y_{j,t}^2$ by $21/20$ times its deterministic envelope. It gives an additional last-iterate population risk at most $C(a_{\max} + p)q \le Cpq$. The rare complementary event has probability $O(M^{-10}) = o(p)$ and every risk is at most $1/2$. Consequently,

$$\mathbb{E} F(\boldsymbol{w}_t^{\eta',\mathsf{A}}) \le Cp, \qquad t \ge 0, \quad \mathsf{A} \in \{\mathrm{GD}, \mathrm{SGD}\}, \tag{I.7}$$

uniformly over $\eta' \in [\eta/2, \eta]$.

Combining (I.4), (I.7), $\varepsilon \le p$, and $\left\|\boldsymbol{w}_t^{\eta'}\right\| \le 1$ proves the upper bound in (I.2) for every t. For GD, the population term $p\|\boldsymbol{y}\|^2/[6(1+\varepsilon)]$ and (I.6) prove its matching lower bound for $t \ge t_0$. Conditional Jensen transfers that lower bound to SGD.

Minkowski's inequality applied to the last-iterate upper bound proves the upper bound in (I.3). For the lower bound, nonnegativity and (I.6) imply, on $E \cap G$,

$$\bar{\boldsymbol{y}}_{U,T}^{(\varepsilon),\mathrm{GD}} \ge \frac{\boldsymbol{b}_U}{12\alpha T} \sum_{t=t_0}^{T-1} r_\varepsilon^t.$$

Since $d_\varepsilon t_0 \le 1/4$, Bernoulli's inequality gives $r_\varepsilon^{t_0} \ge 3/4$. For $T \ge 2t_0$,

$$\frac{\sum_{t=t_0}^{T-1} r_\varepsilon^t}{\sum_{t=0}^{T-1} r_\varepsilon^t} = r_\varepsilon^{t_0} \frac{1 - r_\varepsilon^{T-t_0}}{1 - r_\varepsilon^T} \ge \frac{3}{8}.$$

This proves the GD lower bound, which conditional Jensen transfers to SGD. The displayed rates follow from $[1-(1-d)^T]/(Td) = \Theta(\min\{1, (dT)^{-1}\})$ for $0 < d < 1$, and $d_\varepsilon = \Theta(\varepsilon)$. All constants and events are independent of the eventual horizon. Positive curvature makes every update a strict Euclidean contraction, proving convergence to zero for every $\varepsilon > 0$ as asserted. □

J AN OVERFITTING PLATEAU UNDER SMOOTH FULL-SUPPORT SAMPLING

The discrete rank-one construction admits a full-dimensional smoothing with an explicit, polynomially long high-risk plateau. Every sample still has rank exactly one almost surely. The density can be made strictly positive at every point of the open unit ball. Exact interpolation eventually recovers the target, while the GD and SGD trajectories retain a finite-time high-risk plateau.

J.1 A PATHWISE FEATURE-PERTURBATION BOUND

Lemma J.1 (Perturbing feature operators along arbitrary update paths)**.** *Let $B_t, \widetilde{B}_t$ be linear operators between finite-dimensional Euclidean spaces, with common domain, satisfying $\|B_t\|_{\mathrm{op}}, \|\widetilde{B}_t\|_{\mathrm{op}} \le 1$ and $\|\widetilde{B}_t - B_t\|_{\mathrm{op}} \le \rho$. For $0 < \eta \le 1$, start $\boldsymbol{e}_0 = \widetilde{\boldsymbol{e}}_0 = \boldsymbol{u}$ and set*

$$\boldsymbol{e}_{t+1} = (I - \eta B_t^* B_t)\boldsymbol{e}_t, \qquad \widetilde{\boldsymbol{e}}_{t+1} = (I - \eta \widetilde{B}_t^* \widetilde{B}_t)\widetilde{\boldsymbol{e}}_t.$$

Then, for every integer $t \ge 0$,

$$\|\widetilde{\boldsymbol{e}}_t - \boldsymbol{e}_t\| \le 2\rho\|\boldsymbol{u}\|\sqrt{\frac{\eta t}{2-\eta}}. \tag{J.1}$$

The same right-hand side with $t = T$ bounds the difference of the prefix averages of the first T iterates. No independence assumption on the operators or update order is needed.

Proof. All update matrices are contractions. In particular $\|\boldsymbol{e}_t\| \le R$ for $R = \|\boldsymbol{u}\|$, and the usual energy identity gives

$$\eta(2-\eta)\sum_{s<t}\|B_s\boldsymbol{e}_s\|^2 \le R^2. \tag{J.2}$$

Define $\Delta_t := \widetilde{B}_t - B_t$ and $\widetilde{P}_t := I - \eta\widetilde{B}_t^*\widetilde{B}_t$. The exact error recursion is

$$\boldsymbol{d}_{t+1} = \widetilde{P}_t\boldsymbol{d}_t - \eta\widetilde{B}_t^*\Delta_t\boldsymbol{e}_t - \eta\Delta_t^* B_t\boldsymbol{e}_t, \qquad \boldsymbol{d}_t = \widetilde{\boldsymbol{e}}_t - \boldsymbol{e}_t.$$

Split $\boldsymbol{d}_t = \boldsymbol{z}_t + \boldsymbol{h}_t$, starting both sequences at zero, with

$$\boldsymbol{z}_{t+1} = \widetilde{P}_t\boldsymbol{z}_t - \eta\widetilde{B}_t^*\Delta_t\boldsymbol{e}_t, \qquad \boldsymbol{h}_{t+1} = \widetilde{P}_t\boldsymbol{h}_t - \eta\Delta_t^* B_t\boldsymbol{e}_t.$$

For any contraction B and vectors $\boldsymbol{z}, \boldsymbol{a}$, putting $\boldsymbol{r} = B\boldsymbol{z}$ gives

$$\begin{aligned}\|(I-\eta B^*B)\boldsymbol{z} - \eta B^*\boldsymbol{a}\|^2 - \|\boldsymbol{z}\|^2 &\le -\eta(2-\eta)\|\boldsymbol{r}\|^2 - 2\eta(1-\eta)\langle \boldsymbol{r}, \boldsymbol{a}\rangle + \eta^2\|\boldsymbol{a}\|^2 \\ &\le \frac{\eta}{2-\eta}\|\boldsymbol{a}\|^2.\end{aligned}$$

The last line completes the square and uses $(1-\eta)^2 + \eta(2-\eta) = 1$. Applying this with $\boldsymbol{a} = \Delta_t\boldsymbol{e}_t$ and telescoping gives $\|\boldsymbol{z}_t\| \le \rho R\sqrt{\eta t/(2-\eta)}$. Contraction, Cauchy–Schwarz, and (J.2) give

$$\|\boldsymbol{h}_t\| \le \eta\rho\sum_{s<t}\|B_s\boldsymbol{e}_s\| \le \rho R\sqrt{\frac{\eta t}{2-\eta}}.$$

Adding the two estimates proves (J.1). The averaging assertion follows by the triangle inequality and $T^{-1}\sum_{t<T}\sqrt{t} \le \sqrt{T}$. □

For GD, take $B_t = B$ with rows $\boldsymbol{x}_i^\top/\sqrt{n}$. A paired rowwise perturbation of radius ρ changes this operator by at most ρ in operator norm. For SGD, take $B_t = \boldsymbol{x}_{i_t}^\top$ and use the same index path in both experiments. This coupling preserves repeated sampling from each fixed training set.

Corollary J.2 (Root-risk transfer)**.** *Suppose two bounded-feature laws admit an i.i.d. coupling $\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\| \le \rho$ almost surely, with the same target and initial error $\boldsymbol{u}$. At step $\eta = 1/4$, let $R_0(T), \widetilde{R}(T)$ be their expected population half-squared losses for prefix-averaged GD or fixed-dataset SGD. Then*

$$\left|\sqrt{\widetilde{R}(T)} - \sqrt{R_0(T)}\right| \le \frac{\rho\|\boldsymbol{u}\|}{\sqrt{2}}\left(1 + 2\sqrt{T/7}\right), \qquad T \ge 1. \tag{J.3}$$

Proof. Couple the training observations and, for SGD, the independent index randomness. Couple an additional fresh test pair independently of both. The expected half-squared risk is one half the squared L^2 norm of the test prediction error. The coupled prediction errors satisfy

$$|\widetilde{\boldsymbol{x}}^\top \widetilde{\bar{\boldsymbol{e}}}_T - \boldsymbol{x}^\top \bar{\boldsymbol{e}}_T| \le \|\widetilde{\bar{\boldsymbol{e}}}_T - \bar{\boldsymbol{e}}_T\| + \rho\|\boldsymbol{u}\|.$$

The reverse triangle inequality in L^2 and Lemma J.1 therefore give (J.3). □

J.2 Explicit smooth distributions and time scales

Use $M, q, p, \boldsymbol{b}, \boldsymbol{v}, \boldsymbol{u}, \gamma$ from (F.1)–(F.3), and let $\boldsymbol{u}_0 := \boldsymbol{u}/2 + \gamma\boldsymbol{v}$ denote the initial error. In $d = M+1$ coordinates, the base feature law corresponding to (G.1) is

$$\boldsymbol{x} = \begin{cases} \sqrt{2/3}\,\boldsymbol{v}, & \text{with probability } 1/2, \\ \boldsymbol{e}_j/\sqrt{3}, & \text{with probability } p \text{ for each } j \in [M], \\ 0, & \text{otherwise.} \end{cases} \tag{J.4}$$

Let $\boldsymbol{\xi}, \boldsymbol{\xi}'$ be independent random vectors with common radial smooth density

$$\psi_d(\boldsymbol{z}) = c_d \exp\left(-\frac{1}{1-\|\boldsymbol{z}\|^2}\right) \mathbf{1}_{\{\|\boldsymbol{z}\|<1\}},$$

where c_d is the normalizing constant. Fix $0 < \varepsilon \le 1/2$ and set $\zeta = \varepsilon^2/n$. Define

$$\boldsymbol{x}_\varepsilon = \begin{cases} (\boldsymbol{x} + \varepsilon\boldsymbol{\xi})/(1+\varepsilon), & \text{with probability } 1-\zeta, \\ \boldsymbol{\xi}', & \text{with probability } \zeta, \end{cases} \tag{J.5}$$

with independent $\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\xi}'$ and branch selection. Every draw and every training observation uses new independent variables. The law is fixed before the sample and is the same at every algorithmic horizon. Use noiseless labels $y_\varepsilon = \boldsymbol{x}_\varepsilon^\top \boldsymbol{w}^*$. One may take initialization zero and $\boldsymbol{w}^* = -\boldsymbol{u}_0$, so that the initial error is $\boldsymbol{u}_0$ and $\|\boldsymbol{w}^*\| < 1$.

Theorem J.3 (A smooth full-support design with a long high-risk plateau)**. Design regularity.** *For all sufficiently large n, the features in* (J.5) *have norm at most one, have a C^∞ density on $\mathbb{R}^d$ strictly positive throughout $\{\boldsymbol{x} : \|\boldsymbol{x}\| < 1\}$, and are nonzero almost surely. Thus the sample Hessians have rank exactly one almost surely.*

Finite-time risk transfer. *Let $R_\varepsilon(T)$ be the expected population half-squared loss of prefix-averaged GD or fixed-dataset SGD with $\eta = 1/4$. Let $R_0(T)$ denote the base risk in Theorem G.1, and let c_*, C_* be its constants. For every integer $T \ge 1$,*

$$\left|\sqrt{R_\varepsilon(T)} - \sqrt{R_0(T)}\right| \le 6\varepsilon\sqrt{T}. \tag{J.6}$$

In particular there are absolute constants $a, c, C > 0$ such that

$$c\Phi_n(T) \le R_\varepsilon(T) \le C\Phi_n(T) \quad \text{if } 1 \le T \le apq/\varepsilon^2, \tag{J.7}$$

$$cp \le R_\varepsilon(T) \le Cp \quad \text{if } (pq)^{-1} \le T \le ap/\varepsilon^2. \tag{J.8}$$

These bounds apply whenever the corresponding interval is nonempty. The constants are independent of n, ε, T, and the algorithm.

For the explicit choice $\varepsilon = n^{-2}$, the same smooth distribution therefore satisfies

$$\begin{aligned} R_\varepsilon(1) &= \Theta(1/n), \\ R_\varepsilon(n) &= \Theta(\log n/n^{3/2}), \\ R_\varepsilon(T) &= \Theta(\log n/n) \quad \text{uniformly for } \frac{An^{3/2}}{\log n} \le T \le a'n^3 \log n, \end{aligned} \tag{J.9}$$

for suitable absolute $A, a' > 0$. The complete curve is preserved through $T \le a''n^{5/2}\log n$.

Eventual recovery. *For each fixed n and every admissible $\varepsilon > 0$, the empirical covariance is positive definite almost surely. Every exact parameter-valued interpolator equals $\boldsymbol{w}^*$ and has zero population risk. Both GD and SGD converge to $\boldsymbol{w}^*$ and $R_\varepsilon(T) \to 0$ as $T \to \infty$.*

Proof. The base features have norm at most one for the stated large-n range. Both branches in (J.5) retain this bound. The first branch is a finite mixture of translates and rescalings of ψ_d, hence has a globally smooth density. The second branch has positive mass and a density strictly positive on the open unit ball. Both branches are absolutely continuous, so the features are nonzero almost surely and their Hessians have rank one.

First omit the second branch and denote the resulting risk by $R_{\mathrm{loc}}(T)$. Its defining coupling satisfies $\|(\boldsymbol{x}+\varepsilon\boldsymbol{\xi})/(1+\varepsilon)-\boldsymbol{x}\| \le 2\varepsilon$. Corollary J.2, $\|\boldsymbol{u}_0\| \le 1$, and $T \ge 1$ give

$$|\sqrt{R_{\mathrm{loc}}(T)} - \sqrt{R_0(T)}| \le 3\varepsilon\sqrt{T}. \tag{J.10}$$

Now couple the local and two-branch training samples by replacing each observation independently with probability ζ. Their trajectories agree when no replacement occurs, an event with complement of probability at most $n\zeta$. On its complement, their average errors differ by at most $2\|\boldsymbol{u}_0\| \le 2$. Couple a fresh test observation in the same way, independently. Another L^2 triangle inequality gives, uniformly over all T,

$$|\sqrt{R_\varepsilon(T)} - \sqrt{R_{\mathrm{loc}}(T)}| \le \sqrt{2}(\sqrt{n\zeta}+\sqrt{\zeta}) \le 2\sqrt{2}\,\varepsilon.$$

Combining this with (J.10) proves (J.6).

Theorem G.1 gives $c_*\Phi_n(T) \le R_0(T) \le C_*\Phi_n(T)$, with $\Phi_n(T) \ge pq$. If $T \le c_* pq/(144\varepsilon^2)$, the perturbation in (J.6) is at most $\sqrt{c_*\Phi_n(T)}/2$. Squaring the resulting upper and lower root bounds proves (J.7). For $T \ge (pq)^{-1}$, $p \le \Phi_n(T) \le 3p$ for large n. The same argument with $T \le c_* p/(144\varepsilon^2)$ proves (J.8). Finally $q = \Theta(n^{-1/2})$ and $p = \Theta(\log n/n)$, yielding all the explicit scales.

To prove eventual recovery, note that $d = M+1 \le n$ for all large n. Independent vectors with densities span $\mathbb{R}^d$ after d draws almost surely: condition on the first $k < d$ draws, whose span is a Lebesgue-null proper subspace, and the next draw lies outside it almost surely. Thus the only solution of all noiseless sample equations is $\boldsymbol{w}^*$. Let $H_{S,\varepsilon}$ denote the positive-definite empirical covariance. GD contracts geometrically on each fixed sample. For SGD, conditional on that sample,

$$\mathbb{E}[\|\boldsymbol{e}_{t+1}\|^2 \mid \boldsymbol{e}_t, S] \le [1-\eta(2-\eta)\lambda_{\min}(H_{S,\varepsilon})]\|\boldsymbol{e}_t\|^2.$$

Geometric second-moment convergence and Borel–Cantelli give $\boldsymbol{e}_t \to 0$ almost surely. Prefix averages also converge to zero. All risks are bounded by $\|\boldsymbol{u}_0\|^2/2$, so dominated convergence proves the asserted expected-risk limit for both algorithms. □

The smoothing radius and the global-mixture mass decrease with n. The theorem gives one smooth distribution per n and controls its risk uniformly over polynomially expanding time intervals, beyond continuity at a fixed n and T. It does not establish the same plateau at a fixed isotropic perturbation level, under uniformly well-conditioned designs, or for a broad collection of natural covariates. The sharp recovery time beyond the guaranteed plateau is also left open.

K Protection by observable regularization

Ridge regularization around the known initialization avoids the selected interpolator's worst-case logarithmic loss while permitting complete optimization of an observable objective. Its $O(1/n)$ guarantee matches the known rate from stopping at $T = \Theta(n)$. Proposition K.1 supplies the free-estimator upper bound in Section 3 and the protection comparison in Section 4. Its proof is independent of the interpolation lower bounds.

K.1 An observable objective with an optimal fully optimized guarantee

In the original, unshifted coordinates define

$$\widehat{\boldsymbol{w}}_\lambda = \arg\min_{\boldsymbol{w}} \left\{ F_S(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}-\boldsymbol{w}_1\|^2 \right\}, \qquad \lambda > 0. \tag{K.1}$$

This is ridge regularization around the known initialization. It requires neither the population Hessian nor the unknown common minimizer. After translation solely for analysis, its residual is

$\widehat{\boldsymbol{w}}_\lambda - \boldsymbol{w}^* = \lambda(H_S + \lambda I)^{-1}(\boldsymbol{w}_1 - \boldsymbol{w}^*)$. The leave-one-out proof uses the established exchangeability and matrix-perturbation approach to random-design ridge analysis (Mourtada & Rosasco, 2022). The result adapts this analysis to the present normalization. Neither the algorithm nor the general proof approach is new.

Proposition K.1 (Distribution-free ridge protection)**.** *For every i.i.d. quadratic model of the paper and every $\lambda > 0$,*

$$\mathbb{E}F(\widehat{\boldsymbol{w}}_\lambda) \le \frac{1}{8}\left(\lambda + \frac{2}{n} + \frac{1}{n^2\lambda}\right). \tag{K.2}$$

In particular, $\lambda = 1/n$ gives $\mathbb{E}F(\widehat{\boldsymbol{w}}_{1/n}) \le 1/(2n)$ in every dimension and for every trace or sample-rank budget. This rate is minimax optimal up to constants even over one-dimensional bounded-feature noiseless regression, when the common target is unknown.

Proof. Write $\boldsymbol{u} = \boldsymbol{w}_1 - \boldsymbol{w}^*$, $\|\boldsymbol{u}\| \le 1$, and introduce $n+1$ i.i.d. Hessians. Set

$$B = \lambda I + \frac{1}{n}\sum_{i=1}^{n+1} H_i, \quad \boldsymbol{v} = \lambda B^{-1}\boldsymbol{u}, \quad C_i = B - H_i/n, \quad \boldsymbol{v}_{-i} = \lambda C_i^{-1}\boldsymbol{u}.$$

The held-out Hessian H_i is independent of $\boldsymbol{v}_{-i}$, which has the law of the ridge residual trained on n observations. The resolvent identity implies

$$H_i^{1/2}\boldsymbol{v}_{-i} = \left(I + \frac{1}{n}H_i^{1/2}C_i^{-1}H_i^{1/2}\right)H_i^{1/2}\boldsymbol{v}.$$

Since $C_i \succeq \lambda I$ and $H_i \preceq I$, the operator in parentheses has norm at most $1+1/(n\lambda)$. Exchangeability gives

$$\begin{aligned}
\mathbb{E}F(\widehat{\boldsymbol{w}}_\lambda) &= \frac{1}{2(n+1)}\mathbb{E}\sum_{i=1}^{n+1}\boldsymbol{v}_{-i}^\top H_i \boldsymbol{v}_{-i} \\
&\le \frac{n}{2(n+1)}\left(1 + \frac{1}{n\lambda}\right)^2 \mathbb{E}[\boldsymbol{v}^\top(B - \lambda I)\boldsymbol{v}] \\
&\le \frac{\lambda}{8}\left(1 + \frac{1}{n\lambda}\right)^2.
\end{aligned}$$

The last step applies the scalar bound $\lambda^2 a/(a+\lambda)^2 \le \lambda/4$ to each eigenvalue of $B - \lambda I$.

For the minimax lower bound let $w_1 = 0$ and choose the unknown target $w^* \in \{-1, 1\}$. Draw $a = 1$ with probability $p = 1/(n+1)$, and $a = 0$ otherwise, with observed response $y = aw^*$ and loss $(aw - y)^2/2$. On the event of n zero features, the observations do not depend on w^*. Any possibly randomized output W therefore satisfies $\frac{1}{2}\mathbb{E}[(W-1)^2 + (W+1)^2 \mid a_1 = \cdots = a_n = 0] \ge 1$. Averaging the risks over the two targets therefore gives at least $p(1-p)^n/2 \ge 1/(4en)$. At least one target attains this bound. This $1/n$ lower bound permits arbitrary parameter estimators in the same linear class. Requiring interpolation raises the worst-case rate to $\log n/n$ in sufficient dimension (Theorem 3.1). The converged unregularized gradient methods attain that constrained rate. □

The guarantee is achievable by an ordinary convergent optimization algorithm. For example, GD on the objective in (K.1) with step $1/(1+\lambda)$ contracts to its minimizer by the factor $(1+\lambda)^{-1}$ per update. Its kth update $\boldsymbol{w}^{(k)}$ satisfies $\mathbb{E}F(\boldsymbol{w}^{(k)}) \le 2\mathbb{E}F(\widehat{\boldsymbol{w}}_\lambda) + (1+\lambda)^{-2k}$. Indeed, the initial distance to the ridge solution is at most one, and a quadratic risk obeys $F(\boldsymbol{a}+\boldsymbol{b}) \le 2F(\boldsymbol{a}) + \|\boldsymbol{b}\|^2$ in translated coordinates. Thus with $\lambda = 1/n$, complete optimization and all sufficiently late iterates retain an $O(1/n)$ guarantee. This is an algorithmic intervention, unlike adding curvature around the unknown target in a theoretical perturbation.

K.2 Ridge retains the low-risk valley under both sampling laws

The general ridge bound does not resolve a valley substantially below $1/n$. The next corollary gives a sharper guarantee for the original rank-two construction and for its rank-one splitting, which have the same population Hessian and deterministic initialization (Theorems F.1 and G.1).

Corollary K.2 (Preserving the valley under complete optimization)**.** *For either the rank-two instance* (F.1)–(F.3) *or the rank-one splitting* (G.1)*, the following bound holds for every* $0 < \lambda \le 1$*:*

$$\mathbb{E}F(\widehat{\boldsymbol{w}}_\lambda) \le C\left[pq + \frac{\lambda^2 \log(2/\lambda)}{n} + p\min\left\{\left(\frac{pq}{\lambda}\right)^2, 1\right\}\right]. \tag{K.3}$$

In particular, for all sufficiently large n*,*

$$\mathbb{E}F(\widehat{\boldsymbol{w}}_{1/n}) = \Theta(pq) = \Theta(\log n/n^{3/2}), \qquad \frac{\mathbb{E}F(\boldsymbol{w}_\infty)}{\mathbb{E}F(\widehat{\boldsymbol{w}}_{1/n})} = \Theta(\sqrt{n}).$$

Each expectation is taken under the selected sampling law. The regularization strength uses only n*. For both laws this matches the order achieved by the early average at* $T = n$*, so complete optimization with ridge retains that early-stopping benchmark.*

Centering ridge at the prepared initialization becomes centering at zero under the translation in Appendix G. Thus the same risk guarantee applies to the zero-initialized instance of Theorem 4.1.

Proof. Let $\boldsymbol{u}_{\mathrm{init}}$ denote the common prepared initial error in the target-zero coordinates of the two constructions. Let $\eta = 1/4$, $P = I - \eta H_S$, and $r = (1+\eta\lambda)^{-1}$. A geometric series, followed by summation of the average filters, gives the exact identity

$$\lambda(H_S + \lambda I)^{-1}\boldsymbol{u}_{\mathrm{init}} = (1-r)\sum_{t\ge 0} r^t P^t \boldsymbol{u}_{\mathrm{init}} = \sum_{T\ge 1} T(1-r)^2 r^{T-1}\bar{\boldsymbol{w}}_T^{\mathrm{GD}}. \tag{K.4}$$

The weights $a_T = T(1-r)^2 r^{T-1}$ are nonnegative and sum to one. Convexity of F and the full expected averaged-GD curve under the selected law (Theorem F.1 or G.1) therefore give the claimed bound, using

$$\sum_{T\ge 1}\frac{a_T}{(1+T)^2} \le \frac{(1-r)^2}{r}\log\frac{1}{1-r} \le C\lambda^2\log(2/\lambda), \qquad \sum_{T\ge 1} a_T T^2 \le \frac{C}{(1-r)^2} \le \frac{C}{\lambda^2}.$$

The last term follows by taking the smaller of the second-moment bound and the bound by one. At $\lambda = 1/n$, the extra terms are $O(n^{-3}\log n + p^3q^2n^2) = o(pq)$ under the specified calibration.

For the matching lower bound, write $H = (\boldsymbol{v}\boldsymbol{v}^\top + p\,\mathrm{diag}(0, I_M))/3$. On the two-dimensional space spanned by $\boldsymbol{e}_0$ and $(0, \boldsymbol{b})$, the unscaled matrix has determinant pq and trace $1 + p + q \le 3$. On the orthogonal y space it equals pI, so $H \succeq (pq/9)I$ for large n. For either law write $3H_S = h\boldsymbol{v}\boldsymbol{v}^\top + \mathrm{diag}(0, \boldsymbol{a})$, where $h = 1$ for the rank-two law and $h = 2N_c/n$ for the rank-one law, with N_c counting shared observations. Put $U = \{j : a_j = 0\}$, $Q = |U|/M$, $G = \{q/2 \le Q \le 2q\}$, and $G' = G \cap \{h \ge 1/2\}$. The gate-count marginal law is identical in both constructions, so $\mathbb{P}(G) \ge 15/16$. For the rank-one law, $N_c \sim \mathrm{Bin}(n, 1/2)$ and Hoeffding's inequality gives $\mathbb{P}(h < 1/2) \le e^{-n/8}$. Consequently $\mathbb{P}(G') \ge 15/16 - e^{-n/8} \ge 3/4$ for sufficiently large n, without requiring independence between the two events.

On G', the shared coefficient is positive, and therefore

$$K = \ker H_S = \{(x, \boldsymbol{y}) : \boldsymbol{y}_{U^c} = 0,\ \boldsymbol{v}^\top(x, \boldsymbol{y}) = 0\}, \qquad \|P_K \boldsymbol{e}_0\|^2 = \frac{Q}{q+Q} \ge \frac{1}{3}.$$

The common initialization satisfies $P_K\boldsymbol{u}_{\mathrm{init}} = P_K\boldsymbol{e}_0/2$, since $\boldsymbol{u}_{\mathrm{init}} - \boldsymbol{e}_0/2$ is a multiple of $\boldsymbol{v} \perp K$. Every ridge residual preserves this same K component, and hence has norm squared at least $1/12$. The population spectral bound and the probability of G' now give $\mathbb{E}F(\widehat{\boldsymbol{w}}_\lambda) \ge cpq$ under either law. The corresponding unregularized limiting rates follow from their full expected risk curves and bounded convergence. □

L Interpolation-constrained minimax risk

This appendix proves the statistical statements in Theorem 3.1, Theorem 3.2, and Corollary 3.3. We first specify the observation experiment and estimator classes and state the corresponding minimax laws. Both classes use the same data and may output any vector in $\mathbb{R}^d$. They differ only in the requirement to interpolate.

The upper bounds use Lemma C.9, the relative-curvature bounds proved in Appendix D, and Proposition K.1. For the lower bounds, averaging over a hidden-target family identifies a fixed hard instance for each estimator. These arguments establish the main statistical statements independently.

L.1 Experiment, estimator classes, and main laws

An instance is a distribution Q of matrices $0 \preceq H_z \preceq I$ and a fixed target $\boldsymbol{w}^* \in \mathbb{R}^d$ with $\|\boldsymbol{w}^*\| \le 1$, where $\operatorname{tr}(\mathbb{E}_Q H_z) \le \tau$. The observation is the quadratic loss

$$f(\boldsymbol{w}; z) = \tfrac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^\top H_z (\boldsymbol{w} - \boldsymbol{w}^*).$$

The learner may observe the entire loss function, equivalently its quadratic, linear, and constant coefficients. Neither Q nor $\boldsymbol{w}^*$ is given to the learner. Both are fixed jointly as the instance before the i.i.d. training sample is drawn. The initialization is zero. Let $F(\boldsymbol{w}) := \mathbb{E}_{H_z \sim Q}[f(\boldsymbol{w}; z)]$ denote the population risk of this instance.

Let $\mathcal{I}_{n,d,\tau}$ contain all measurable parameter estimators that interpolate every observed loss almost surely on every admissible instance. Their independent randomization is unrestricted, and their output need not satisfy a norm constraint. Define

$$\mathcal{R}_{\mathrm{int}}(n, d; \tau) = \inf_{\widehat{\boldsymbol{w}} \in \mathcal{I}_{n,d,\tau}} \sup_{Q, \boldsymbol{w}^*} \mathbb{E} F(\widehat{\boldsymbol{w}}(S)).$$

Let $\mathcal{R}_{\mathrm{free}}$ denote the same infimum over all parameter estimators, without requiring interpolation. In both definitions the output is a vector in the same space $\mathbb{R}^d$, rather than an arbitrary nonlinear prediction function.

The first law gives part (i) of Theorem 3.2.

Theorem L.1 (Sharp interpolation-constrained resource law)**.** *For all $n, d \ge 1$ and $0 < \tau \le d$,*

$$\mathcal{R}_{\mathrm{int}}(n, d; \tau) = \Theta\left(\min\left\{1, \frac{\log(1 + n\tau)}{n}, \frac{1 + \log d}{n}\right\}\right), \tag{L.1}$$

$$\mathcal{R}_{\mathrm{free}}(n, d; \tau) = \Theta(\min\{\tau, 1/n\}). \tag{L.2}$$

All constants are universal, and both risks are zero at $\tau = 0$. The minimum-norm empirical solution attains the upper bound in (L.1)*. Ridge centered at the known initialization, with $\lambda = 1/n$, attains* (L.2)*. The same laws hold if every estimator is additionally required to output a vector of norm at most one.*

For scalar regression, the law takes the form stated in Theorem 3.1.

Theorem L.2 (The price of interpolation in noiseless scalar regression)**.** *For noiseless scalar regression $y = \boldsymbol{x}^\top \boldsymbol{w}^*$ with $\|\boldsymbol{x}\| \le 1$ and $\|\boldsymbol{w}^*\| \le 1$, the interpolation-constrained minimax expected half-squared loss in dimension d is*

$$\Theta\left(\frac{\min\{\log(1 + n), 1 + \log d\}}{n}\right). \tag{L.3}$$

Without the interpolation constraint, the minimax risk over parameter estimators is $\Theta(1/n)$. The logarithmic separation already occurs in dimension $\lfloor \sqrt{n} \rfloor + 1$ using unit-norm features, so every sample Hessian in this lower bound has rank exactly one.

A sample-rank budget gives the statistical part of Corollary 3.3.

Theorem L.3 (Sharp interpolation-constrained sample-Hessian rank law)**.** *Replace the trace budget in the quadratic experiment by the condition $\operatorname{rank} H_z \le r$ almost surely, with no additional trace constraint. For integers $n, d \ge 1$ and $1 \le r \le d$, its interpolation-constrained minimax risk is*

$$\Theta\left(\min\left\{1, \frac{1 + \log \min\{d, nr\}}{n}\right\}\right). \tag{L.4}$$

The minimax risk without requiring interpolation is $\Theta(1/n)$. Both conclusions continue to hold with an additional unit-norm constraint on the estimator's output.

The upper estimators are the minimum-norm solution identified by Lemma C.9 and the ridge estimator in Proposition K.1. Their bounds follow from Corollary D.2 and the ridge proposition. Ridge and its exchangeability analysis are established tools (Mourtada & Rosasco, 2022). The new lower bound quantifies the restriction to empirical interpolators: for every estimator there exists a fixed hard instance. The true $\boldsymbol{w}^*$ remains an empirical minimizer with zero population risk on every sample, so the lower bound cannot apply to every empirical minimizer of a given sample.

L.2 A HIDDEN-TARGET POSTERIOR LEMMA

Fix an integer $m \geq 2$, use $m+1$ active coordinates, and let $\boldsymbol{b} = (m+1)^{-1/2}\mathbf{1}$ on these coordinates. Zero-padding embeds the construction in any $d \geq m+1$. A hidden index $J \in [m+1]$ specifies the fixed target $\boldsymbol{w}_J^* = \boldsymbol{e}_J$. On instance J, draw $Z_J = 0$ deterministically and let the other Z_j be independent Bernoulli(p) variables. For $c > 0$ put

$$f_J(\boldsymbol{w}; \boldsymbol{Z}) = \frac{c}{2}\left(\boldsymbol{b}^\top \boldsymbol{w} - \frac{1}{\sqrt{m+1}}\right)^2 + \frac{1}{4}\sum_{j \neq J} Z_j w_j^2, \qquad H_{\boldsymbol{Z}} = c\boldsymbol{b}\boldsymbol{b}^\top + \tfrac{1}{2}\operatorname{diag}(\boldsymbol{Z}). \tag{L.5}$$

This loss has common zero-loss target $\boldsymbol{e}_J$. Its linear term is $-\boldsymbol{w}^\top H_{\boldsymbol{Z}}\boldsymbol{e}_J$, where $H_{\boldsymbol{Z}}\boldsymbol{e}_J = c\boldsymbol{b}/\sqrt{m+1}$ is independent of J. The constant coefficient likewise does not reveal J.

For a sample let V contain the coordinates having a nonzero mask in at least one observation, put $U = [m+1] \setminus V$, and write $K = |U|$. Under its fixed instance the true index J always belongs to U. Introduce a uniform prior on J solely for the lower-bound proof. The likelihood of any ordered mask transcript under a candidate $J \in U$ is $p^{N_1}(1-p)^{nm-N_1}$, where N_1 counts all nonzero mask entries. It is zero for $J \in V$. Thus the conditional posterior of J given the complete observed losses is uniform on U. Conditioning further on the estimator's independent random seed preserves this posterior.

Every interpolating vector $\boldsymbol{w}$ satisfies $\boldsymbol{w}_V = 0$ and $\sum_{j \in U} w_j = 1$. Averaging its population loss over that posterior therefore gives

$$\mathbb{E}_J[F_J(\boldsymbol{w}) \mid S, \boldsymbol{w}] = \frac{p}{4}\left(1 - \frac{1}{K}\right)\sum_{j \in U} w_j^2 \geq \frac{p(K-1)}{4K^2}. \tag{L.6}$$

The inequality follows from Cauchy–Schwarz and applies to arbitrary real output coordinates, without any norm or sign restriction.

Choose $p = 1 - m^{-1/n}$. The number $L = K - 1$ of absent non-target coordinates is then $\mathrm{Bin}(m, 1/m)$ under every fixed J. In particular $\mathbb{E}L = 1$, $\mathbb{E}L^2 \leq 2$, and

$$\mathbb{P}(1 \leq L \leq 8) \geq \frac{(\mathbb{E}L)^2}{\mathbb{E}L^2} - \frac{\mathbb{E}L}{9} \geq \frac{7}{18}.$$

On $2 \leq K \leq 9$, $(K-1)/(4K^2) \geq 2/81$. It follows that every interpolating estimator satisfies

$$\frac{1}{m+1}\sum_{J=1}^{m+1} \mathbb{E}_{S|J,\mathrm{alg}} F_J(\widehat{\boldsymbol{w}}(S)) \geq \frac{7p}{729} \geq \frac{p}{128}. \tag{L.7}$$

At least one fixed J attains this lower bound. That index, its target, and its Hessian distribution are selected before the training sample. The estimator may know the displayed family but is not given its unknown index. The argument requires no symmetry of the estimator.

L.3 CALIBRATION OF DIMENSION AND POPULATION TRACE

Proof of the lower bound in Theorem L.1. First suppose $n \geq 16$, $B = n\tau \geq 16$, and $d \geq 3$. Set $m = \lfloor \min\{d-1, \sqrt{B}, e^{n/4}\} \rfloor$ and $c = \min\{\tau/2, 1/2\}$. Choose $p = 1 - m^{-1/n}$. Then $m \geq 2$, $\log m/n \leq 1/4$, and every Hessian in (L.5) is a PSD contraction. Moreover, $\operatorname{tr}(\mathbb{E}H_{\boldsymbol{Z}}) = c + mp/2 \leq \tau$ because $p \leq \log m/n$ and $m \log m \leq \sqrt{B}\log B/2 \leq B$.

For $x \geq 2$, $\lfloor x \rfloor \geq \sqrt{x}$. Also $\log(d-1) \geq (1+\log d)/4$ for $d \geq 3$ and $\log B/2 \geq \log(1+B)/4$ for $B \geq 16$. Thus $\log m \geq \frac{1}{8}\min\{1 + \log d, \log(1+B), n\}$. Since $1 - e^{-x} \geq x/2$ for $0 \leq x \leq 1/4$,

$$p \geq \frac{\log m}{2n} \geq \frac{1}{16}\min\left\{1, \frac{\log(1+n\tau)}{n}, \frac{1+\log d}{n}\right\}.$$

Equation (L.7) proves the desired lower bound in this regime, with an explicit constant $1/2048$.

For the remaining parameters, embed a scalar experiment in the first coordinate. Its unknown target is $\boldsymbol{w}^* = \pm\boldsymbol{e}_1$. Its sample Hessian is $\boldsymbol{e}_1\boldsymbol{e}_1^\top$ with probability $p_0 = \min\{\tau, 1/(n+1)\}$ and zero otherwise.

On the all-zero sample the two targets are indistinguishable, and their average squared error in the first coordinate is at least one for every estimator. Thus some target has expected risk at least

$$\frac{p_0}{2}(1-p_0)^n \geq \frac{1}{4e}\min\{\tau, 1/n\}. \tag{L.8}$$

This bounds all estimators, including those not required to interpolate. The resource expression in (L.1) is at most a universal constant times $\min\{\tau, 1/n\}$ whenever $n < 16$, or $n\tau < 16$, or $d \leq 2$. For example use its trace bound when $n\tau \leq 1$, its bound $\log 17/n$ when $1 < n\tau < 16$, and its bound $(1+\log 2)/n$ when $d \leq 2$. If $n < 16$, also use $\min\{1, \tau\} \leq 16\min\{\tau, 1/n\}$, completing the lower bound for all admissible parameters. □

Upper bounds and the free-estimator law. The minimum-norm empirical solution interpolates because $\boldsymbol{w}^*$ is feasible. Its residual relative to $\boldsymbol{w}^*$ is $-P_{\ker H_S}\boldsymbol{w}^*$, and both its output and residual have norm at most one. Corollary D.2, passed to the instancewise limit using Lemma C.9 and bounded risk, therefore proves the upper bound in (L.1).

Ridge with center zero and $\lambda = 1/n$ has expected risk at most $1/(2n)$ by Proposition K.1. Its residual is $-\lambda(H_S + \lambda I)^{-1}\boldsymbol{w}^*$, a Euclidean contraction of $\boldsymbol{w}^*$, so its population risk is also pathwise at most $\tau/2$. Equation (L.8) proves the matching free-estimator lower bound. Both upper estimators have output norm at most one, whereas the lower bounds allowed unrestricted norms. This also proves the last assertion of Theorem L.1. □

L.4 An exactly rank-one sampling law

Proof of Theorem L.2. For $n \geq 16$ and $d \geq 3$, let $m = \min\{d-1, \lfloor\sqrt{n}\rfloor\}$ and $D = m+1$. Set $p = 1 - m^{-1/n}$ and $s = 1 - mp$. Then $mp \leq \log n/(2\sqrt{n}) \leq 1/2$, so $s \geq 1/2$. For hidden $J \in [D]$, take $\boldsymbol{w}^*_J = \boldsymbol{e}_J$ and draw the unit feature

$$\boldsymbol{x} = \boldsymbol{b} = D^{-1/2}\mathbf{1} \quad \text{with probability } s, \qquad \boldsymbol{x} = \boldsymbol{e}_j \quad \text{with probability } p \quad (j \neq J). \tag{L.9}$$

The common-feature label is $1/\sqrt{D}$ and every gate label is zero. All sample Hessians have rank exactly one and trace exactly one. The population loss is

$$F_J(\boldsymbol{w}) = \frac{s}{2D}\left(\sum_{j=1}^{D} w_j - 1\right)^2 + \frac{p}{2}\sum_{j\neq J} w_j^2.$$

For any observed transcript with gate set V, the likelihood under each candidate $J \in U = [D] \setminus V$ is $s^{N_c}p^{n-N_c}$, where N_c counts common observations. Thus a uniform prior on J again gives an exactly uniform posterior on U, also after conditioning on the learner's randomization.

Let L be the number of absent non-target gates. Its mean is one. The missing indicators are pairwise negatively correlated, since $(1-2p)^n \leq (1-p)^{2n} = m^{-2}$, so $\mathbb{E}L^2 \leq 2$ and $\mathbb{P}(1 \leq L \leq 8) \geq 7/18$. The probability of seeing no common feature is $(mp)^n \leq 2^{-n}$. Therefore $\mathbb{P}(N_c \geq 1,\ 2 \leq |U| \leq 9) \geq 7/18 - 2^{-n} \geq 1/4$. On this event, every interpolator has $\boldsymbol{w}_V = 0$ and $\sum_U w_j = 1$. Its posterior population risk is consequently at least $p(|U|-1)/(2|U|^2) \geq 4p/81$. The average over targets is at least $p/81$, so one fixed instance attains that risk. The calibration gives $p \geq (\log m)/(2n)$ and $\log m \geq \frac{1}{5}\min\{\log(1+n), 1+\log d\}$. The latter inequality follows by considering $m = d-1$ and $m = \lfloor\sqrt{n}\rfloor$ separately. In the second case, use $m \geq \sqrt{n}/2$ and $n \geq 16$. This proves the lower bound. The remaining bounded ranges of n or d follow from the scalar two-target construction in (L.8) with $\tau = 1$. Corollary D.2 at $\tau = 1$, passed to the instancewise limit as above, proves the minimum-norm upper bound in (L.3). Proposition K.1 and the same scalar lower bound give the free-estimator rate. □

L.5 A sample-rank ceiling for arbitrary interpolators

Proof of Theorem L.3. The case $r = 1$ follows from Theorem L.2: its lower bound is a subclass of rank-one quadratic observations, while its upper bound follows from $\operatorname{tr} H \leq 1$ and hence also

applies to all such quadratics. For $r \geq 2$, first suppose $n \geq 16$ and $d \geq 3$. Set $B = n(r-1)$ and $m = \lfloor \min\{d-1, \sqrt{B}, e^{n/4}\} \rfloor$. Define $p = 1 - m^{-1/n}$ and $\mu = mp$. Then $m \geq 2$ and

$$\mu \leq \frac{m \log m}{n} \leq \frac{\sqrt{B} \log B}{2n} \leq \frac{B}{n} = r - 1.$$

Use the hidden-target family (L.5) with $c = 1/2$, but replace the independent mask by the following exchangeable mask on the m non-target coordinates. Let A be the adjacent-integer random variable with mean μ: if $a = \lfloor \mu \rfloor$ and $t = \mu - a$, then $\mathbb{P}(A = a+1) = t$ and $\mathbb{P}(A = a) = 1 - t$. Conditional on A, choose a uniform A-subset of those coordinates. The mask has coordinate marginals p, so the population loss is unchanged. Its sample Hessian satisfies

$$0 \preceq H_{\boldsymbol{Z}} = \tfrac{1}{2}\boldsymbol{b}\boldsymbol{b}^\top + \tfrac{1}{2} \operatorname{diag} \boldsymbol{Z} \preceq I, \qquad \operatorname{rank} H_{\boldsymbol{Z}} \leq 1 + A \leq r.$$

Write $\pi_a = \mathbb{P}(A = a)$. For every candidate $J \in U$, the likelihood of an observed mask with support W is $\pi_{|W|} / \binom{m}{|W|}$, independently of that candidate. The posterior on U remains uniform. The deterministic shared term imposes $\sum_U w_j = 1$, and the observed gates impose $\boldsymbol{w}_V = 0$ on every interpolator. Consequently the posterior risk bound $p(K-1)/(4K^2)$ from the hidden-target lemma is unchanged.

The occupancy estimate also survives dependence within each mask. The adjacent-integer law minimizes variance among integer-valued random variables of mean μ. Indeed, for every integer-valued X, $\mathbb{E}[(X-a)(X-a-1)] \geq 0$. If $\mathbb{E}X = \mu$, this gives $\operatorname{Var} X \geq t(1-t) = \operatorname{Var} A$. Apply this to a Binomial(m, p) variable to obtain $\operatorname{Var} A \leq mp(1-p)$. For distinct non-target coordinates j, k it follows that

$$\mathbb{P}(Z_j = Z_k = 1) = \frac{\mathbb{E}[A(A-1)]}{m(m-1)} \leq p^2, \qquad \mathbb{P}(Z_j = Z_k = 0) \leq (1-p)^2.$$

Masks are independent across observations. Each non-target coordinate is therefore missing with probability $(1-p)^n = 1/m$, and two such missing indicators have nonpositive covariance. Once again $\mathbb{E}L = 1$, $\mathbb{E}L^2 \leq 2$, and $\mathbb{P}(1 \leq L \leq 8) \geq 7/18$. The same posterior-risk calculation gives an average risk of at least $p/128$, attained by at least one fixed instance.

The parameter comparison in the proof of Theorem L.1, now with $B = n(r-1)$, gives

$$p \geq \frac{1}{16} \min\left\{1, \frac{\log(1+B)}{n}, \frac{1 + \log d}{n}\right\}.$$

For $r \geq 2$, $nr/2 \leq B \leq nr$, so this is a universal constant times the expression in (L.4). If $n < 16$ or $d \leq 2$, that expression is at most a universal constant times $1/n$, and the scalar two-target lower bound with $\tau = 1$ applies. This covers every integer n, d, r.

For the upper bound, the minimum-norm interpolator is the limiting empirical projection, to which the sample-rank resource upper bound in Corollary D.5 applies. Ridge gives the free-estimator $O(1/n)$ bound, and the rank-one scalar two-target family supplies its matching lower bound. Both upper estimators have norm at most one, whereas none of these lower bounds restricted the output norm. □

L.6 Relationship to prior interpolation and proper-learning bounds

Coupon-collector constructions and logarithmic penalties for restricted regression learners are established ideas. In particular, Høgsgaard et al. (2026), Theorem 3.12, give a logarithmic proper-learning sample-complexity penalty for specially constructed regression classes under cutoff loss. Their proof traces the coupon-collector strategy to Auer & Ortner (2007). The extension to every proper learner in binary classification is due to Bousquet et al. (2020a).

The result here concerns bounded Euclidean norms, globally smooth convex quadratic losses, and a sharp dimension/trace law within the linear parameter class. Removing only the empirical interpolation constraint allows a better parameter estimator in that same linear class. The general proper-learning lower bound for a specially restricted nonconvex hypothesis family does not directly yield this convex-class statement. We do not claim that the general idea of an interpolation price or the coupon-collector proof method is new.

The hidden-index family also makes the information restriction explicit. Its feature/Hessian law and target vary jointly with J, but each pair is fixed before sampling. A learner given that hidden index can simply return its zero-risk target. More generally, the result does not constrain arbitrary nonlinear test-time predictors. It constrains parameter estimators that continue to satisfy every observed interpolation equation.

M THE INFORMATION REQUIRED TO CERTIFY CONTINUED OPTIMIZATION

We prove Theorem 5.1 when the certificate has access to complete regression observations and hence their full loss functions. Theorem M.1 establishes the certification frontier by constructing an upper certificate and matching it with fixed indistinguishable populations. Theorem M.2 extends a single endpoint coverage event to all future iterates and prefix averages. Corollary M.3 then characterizes the validation budget for reliable acceptance.

M.1 EXPERIMENT AND CERTIFICATION FRONTIER

We consider bounded-feature noiseless regression, $f(\boldsymbol{w}; z) = (\boldsymbol{a}_z^\top \boldsymbol{w} - \boldsymbol{a}_z^\top \boldsymbol{w}^*)^2/2$, with $\|\boldsymbol{a}_z\| \le 1$, an unknown fixed target $\boldsymbol{w}^*$, and a known deterministic initialization $\boldsymbol{w}_1$ satisfying $\|\boldsymbol{w}_1 - \boldsymbol{w}^*\| \le 1$. The class is denoted by $\mathcal{Q}_d^{(1)}$ and includes zero features. Training observes n complete feature-response pairs $S \sim D^n$. Independently of S, the validation sample $V = (Z'_j)_{j=1}^m \sim D^m$ contains m fresh pairs, used only to certify the training-selected output. In particular, validation does not change that output.

Put $H_S = \nabla^2 F_S$. Both prescribed algorithms select the observable endpoint

$$\boldsymbol{w}_\infty(S) = \boldsymbol{w}_1 - H_S^+ \nabla F_S(\boldsymbol{w}_1), \qquad R_D(S) = F_D(\boldsymbol{w}_\infty(S)). \tag{M.1}$$

By Lemma C.9, its residual in analysis coordinates is $P_{\ker H_S}(\boldsymbol{w}_1 - \boldsymbol{w}^*)$. Computing (M.1) requires no knowledge of $\boldsymbol{w}^*$ or the population Hessian. A certificate $U \in [0, 1/2]$ may use S, V, the entire training transcript, and independent randomness. Fix $\delta = \beta = 1/16$. Uniform coverage (honesty) and the optimal zero-risk resolution are defined by

$$\sup_{D \in \mathcal{Q}_d^{(1)}} \mathbb{P}_D\{R_D(S) > U\} \le \delta, \tag{M.2}$$

$$W(n, m, d) = \inf_{U \text{ satisfying (M.2)}} \inf \left\{ a \ge 0 : \sup_{D \in \mathcal{Q}_d^{(1)}} \mathbb{P}_D\{R_D(S) = 0,\ U > a\} \le \beta \right\}. \tag{M.3}$$

All probabilities include algorithm and certificate randomness. The precision event is joint and unconditional, rather than conditioned on a favorable training event. The deterministic cap $R_D(S) \le 1/2$ makes the restriction $U \in [0, 1/2]$ harmless.

Theorem M.1 (Sharp fixed-confidence certification frontier)**.** *There are absolute constants $c, C > 0$ such that, for $n \ge 2^{16}$, $d \ge \lceil \sqrt{n} \rceil$, and every integer $m \ge 0$,*

$$c \min\left\{ \frac{\log n}{n}, \frac{1}{m} \right\} \le W(n, m, d) \le C \min\left\{ \frac{\log n}{n}, \frac{1}{m} \right\}. \tag{M.4}$$

Here $1/m = \infty$ at $m = 0$. The upper certificate is bounded by the displayed scale almost surely whenever $R_D(S) = 0$. For every honest certificate, some fixed admissible instance satisfies

$$\mathbb{P}_D\left\{ R_D(S) = 0,\ U \ge c \min\left\{ \frac{\log n}{n}, \frac{1}{m} \right\} \right\} \ge \frac{1}{8}. \tag{M.5}$$

The lower-bound law, target, and initialization are fixed before sampling. The theorem applies to both GD and actual reused-data SGD.

Proof. For the upper bound, write $H = \mathbb{E}H_z$ and $b_0 = 1 - e^{-1}$. Since $\operatorname{tr} H \le 1$, Lemma E.2, together with $F_S(\boldsymbol{w}_\infty) = 0$ and the unit residual bound, shows that

$$U_{\mathrm{tr}} = \min\left\{ \frac{1}{2}, \frac{9}{4n} \log\big(32(1 + 2n)\big) \right\}$$

fails to bound $R_D(S)$ with probability at most $1/32$. For $m > 0$ set

$$\widehat{R}_V = \frac{1}{m}\sum_{j=1}^{m} f(\boldsymbol{w}_\infty; Z'_j), \qquad U_{\mathrm{val}} = \frac{1}{b_0}\left(\widehat{R}_V + \frac{\log 32}{2m}\right), \qquad U = \min\{U_{\mathrm{tr}}, U_{\mathrm{val}}\}.$$

Use $U = U_{\mathrm{tr}}$ if $m = 0$. Conditional on S, the variables $X_j = 2f(\boldsymbol{w}_\infty; Z'_j) \in [0,1]$ have mean $\mu = 2R_D(S)$. The inequality $e^{-x} \le 1 - b_0 x$ gives $\mathbb{E}[\exp(b_0 m\mu - \sum_j X_j) \mid S] \le 1$. The exponential Markov inequality and a union bound prove honesty. If $R_D(S) = 0$, all validation losses vanish almost surely conditional on S, proving the upper assertion in (M.4). The main-text relaxation (5.3) is valid with sufficiently large universal C_1 and $C_2 = 3$, since $1/b_0 < 3$ and $(\log 32)/(2b_0) < 3$.

For the lower bound, first suppose $0 \le m \le n$. Set

$$M_* = \lfloor \sqrt{n}/8 \rfloor, \qquad t = \min\{\log M_*/n, (\log 2)/m\},$$
$$M = \lfloor e^{nt} \rfloor, \qquad p = 1 - M^{-1/n},$$

omitting the second term at $m = 0$. Then $2 \le M \le M_*$, $(1-p)^n = 1/M$, $p \le 1/128$, $np^2 \le 1/64$, and $Mp \le 1/64$. In an active subspace $E \cong \mathbb{R}^{M+1} \subseteq \mathbb{R}^d$ put

$$s = 1/M, \quad \boldsymbol{b} = M^{-1/2}\mathbf{1}, \quad \boldsymbol{v} = (\sqrt{s}, -\boldsymbol{b}),$$
$$\boldsymbol{u} = \frac{(1, \sqrt{s}\,\boldsymbol{b})}{2(1+s)}, \qquad \boldsymbol{w}_1 = 0, \quad \boldsymbol{w}^* = -\boldsymbol{u}.$$

The base law P draws $H_z = \boldsymbol{v}\boldsymbol{v}^\top/4$ with probability $1 - Mp$ and $H_z = \boldsymbol{e}_j\boldsymbol{e}_j^\top/3$ with probability p for each $j \in [M]$. Every sample Hessian has rank exactly one and operator norm at most $3/8$, and $\|\boldsymbol{u}\| < 1/2$. Realize the shared atom by $\boldsymbol{x} = \boldsymbol{v}/2$ and gate j by $\boldsymbol{x} = \boldsymbol{e}_j/\sqrt{3}$, with $y = \boldsymbol{x}^\top \boldsymbol{w}^*$. All laws below use these same feature representatives, target, and initialization, so the experiment identities also hold for the complete feature-response pairs. The gate categories are identifiable from the full observed loss functions. For each fixed j, define the i.i.d. alternative law $P_j = P(\cdot \mid \text{one draw is not gate } j)$.

Let K be the number of missing gates. Define G as the event that exactly one gate is missing and at least one shared constraint is observed. On G, if the unique missing gate is j, then

$$(\ker H_S) \cap E = \mathrm{span}\{\boldsymbol{e}_0 + \boldsymbol{e}_j\}, \qquad \boldsymbol{w}_\infty - \boldsymbol{w}^* = \frac{1}{4}(\boldsymbol{e}_0 + \boldsymbol{e}_j),$$
$$R_P(S) = \frac{p}{96} =: \kappa, \qquad R_{P_j}(S) = 0. \tag{M.6}$$

These identities follow because $\boldsymbol{u} - \boldsymbol{e}_0/2$ is a multiple of $\boldsymbol{v}$, which is perpendicular to the displayed kernel. Unused ambient coordinates have zero target and initialization and do not contribute.

To bound the probability of G, write $B_r = (M)_r/M^r$, where $(M)_r$ is the falling factorial, and interpret terms with $r > M$ as zero. For $1 \le r \le 4$,

$$\mathbb{E}(K)_r = (M)_r(1-rp)^n \le B_r, \qquad \mathbb{E}(K)_3 \ge B_3 e^{-5np^2} \ge (1 - 5/64)B_3.$$

The second bound uses $\log(1-3p) - 3\log(1-p) \ge -9p^2/[2(1-3p)] \ge -5p^2$ for $p \le 1/128$. For every integer $K \ge 0$,

$$K - (K)_2 + \tfrac{1}{2}(K)_3 - \tfrac{1}{6}(K)_4 = -\frac{K(K-2)(K-3)(K-4)}{6} \le \mathbf{1}_{\{K=1\}}.$$

Also $1 - B_2 + B_3/2 - B_4/6 = 1/3 + 1/(2M) - 5/(6M^2) + 1/M^3 \ge 1/3$. Subtracting the probability $(Mp)^n \le 1/64$ of missing the shared constraint yields

$$g := P^n(G) \ge \frac{1}{3} - \frac{5}{128} - \frac{1}{64} > \frac{1}{4}. \tag{M.7}$$

The categorical laws give the following identities for the full ordered training sample:

$$\frac{dP_j^n}{dP^n}(S) = M\mathbf{1}_{\{j \text{ missing}\}}, \qquad \frac{dQ_n}{dP^n}(S) = K(S), \qquad Q_n = \frac{1}{M}\sum_{j=1}^{M} P_j^n.$$

Thus Q_n and P^n have equal restrictions to G, and there the mixture index is the unique missing gate. Under P, validation avoids that gate with probability $q_m = (1-p)^m \geq 1/2$. Conditional on this event and on S, its full law is P_j^m. Any transcript, including the infinite SGD index path, and any further post-processing use the same conditional probability kernel for these identical observed losses. Hence

$$q_m \frac{1}{M} \sum_j \mathbb{P}_{P_j}\{G, U < \kappa\} = \mathbb{P}_P\{G, \text{ validation avoids its missing gate, } U < \kappa\} \leq \delta. \quad \text{(M.8)}$$

The last event violates honesty under P by (M.6). Since $M^{-1}\sum_j P_j^n(G) = g$,

$$\frac{1}{M} \sum_j \mathbb{P}_{P_j}\{G, U \geq \kappa\} \geq g - \delta/q_m > 1/8.$$

Choose one fixed j attaining this average. Every P_j^n is an i.i.d. experiment, and averaging identifies j before any sample is drawn. The outer mixture is used only in the proof. Under P_j, G implies zero actual risk. Finally, $\lfloor x \rfloor \geq \sqrt{x}$ for $x \geq 2$ and $1 - e^{-x} \geq x/2$ for $0 \leq x \leq 1/4$ give $\kappa \geq t/384 \geq c \min\{(\log n)/n, 1/m\}$, using $\log M_* \geq (\log n)/4$ for our range of n.

For $m > n$, a scalar pair of populations suffices. Take $\boldsymbol{w}_1 = 0$, $\boldsymbol{w}^* = -\boldsymbol{e}_1$, and $p_0 = 1 - e^{-(\log 2)/(n+m)}$. Under D_0 every feature is zero. Under D_1 the feature is $\boldsymbol{e}_1$ with probability p_0 and zero otherwise, with response $y = \boldsymbol{x}^\top \boldsymbol{w}^*$. Both are admissible fixed laws with the same target. The all-zero observation event has probability $1/2$ under D_1 and produces the full D_0 experiment, including its constant training path. On it $R_{D_1} = p_0/2$, whereas $R_{D_0} = 0$. Honesty therefore forces $\mathbb{P}_{D_0}\{U \geq p_0/2\} \geq 1 - 2\delta = 7/8$. Since $p_0/2 \geq (\log 2)/(8m)$, this proves (M.5) also for $m > n$. That event has probability greater than β, so every resolution strictly below its threshold fails (M.3). This proves the lower bound in (M.4). □

M.2 From the endpoint to every future output

Theorem M.2 (One endpoint certificate controls every future tail)**.** *Let U be any honest nonnegative endpoint certificate, and set $d_t = \|\boldsymbol{w}_t - \boldsymbol{w}_\infty\|$ and $\bar{d}_T = T^{-1}\sum_{t=1}^T d_t$. For GD and actual reused-data SGD, with probability at least $1-\delta$, simultaneously for every integer $T \geq 1$,*

$$\sup_{t \geq T} F_D(\boldsymbol{w}_t) \leq (\sqrt{U} + d_T/\sqrt{2})^2, \qquad \sup_{k \geq T} F_D(\bar{\boldsymbol{w}}_k) \leq (\sqrt{U} + \bar{d}_T/\sqrt{2})^2. \quad \text{(M.9)}$$

Both bounds are observable and nonincreasing in T, and converge to U almost surely. The horizon may be selected after inspecting the validation data and trajectory.

Proof. Every individual training loss vanishes at $\boldsymbol{w}_\infty$, so a reused step satisfies $\boldsymbol{w}_{t+1} - \boldsymbol{w}_\infty = (I - H_{i_t}/4)(\boldsymbol{w}_t - \boldsymbol{w}_\infty)$. For GD, H_{i_t} is replaced by H_S. Thus d_t and $\bar{d}_T$ are pathwise nonincreasing. The triangle inequality for the population quadratic seminorm gives, for every $\boldsymbol{w}$,

$$\sqrt{F_D(\boldsymbol{w})} \leq \sqrt{F_D(\boldsymbol{w}_\infty)} + \|\boldsymbol{w} - \boldsymbol{w}_\infty\|/\sqrt{2}.$$

On the single event $R_D(S) \leq U$, this proves the first bound. The second follows from $\|\bar{\boldsymbol{w}}_k - \boldsymbol{w}_\infty\| \leq \bar{d}_k \leq \bar{d}_T$. Lemma C.9 gives convergence for both algorithms, including the realized SGD trajectory. Since all horizons share the same coverage event, no union bound over horizons is needed. □

The statistical floor does not give a uniform time for reducing the observable optimization remainders. Conversely, a valid upper bound for an entire future tail also bounds the limiting risk by continuity, so it inherits Theorem M.1.

M.3 A sound acceptance rule with a positive-risk margin

Corollary M.3 (Fresh information for a separated acceptance test)**.** *Under the conditions of Theorem M.1, fix $0 < \varepsilon \leq c_0(\log n)/n$ for a sufficiently small absolute c_0. An acceptance rule $\phi \in \{0,1\}$ must satisfy $\sup_D \mathbb{P}_D\{\phi = 1, R_D(S) > \varepsilon\} \leq \delta$ and $\sup_D \mathbb{P}_D\{\phi = 0, R_D(S) \leq \varepsilon/64\} \leq \delta$. The smallest sufficient validation budget is $\Theta(1/\varepsilon)$. The lower bound already holds with the second condition weakened to $\sup_D \mathbb{P}_D\{\phi = 0, R_D(S) = 0\} \leq \beta$.*

Proof. Given an acceptance rule, set $U = \varepsilon$ on acceptance and $U = 1/2$ on rejection. Soundness gives honesty, while zero-risk completeness gives $W(n, m, d) \le \varepsilon$. The lower bound in (M.4) implies $m \ge c/\varepsilon$ in the stated range of ε. For the upper bound, recall $b_0 = 1 - e^{-1}$ and accept when the certificate constructed above is at most ε, using $m \ge 2 \log 32/\varepsilon$. Soundness follows from honesty. Conditional on any S with $R_D(S) \le \varepsilon/64$, rejection implies $\widehat{R}_V > b_0\varepsilon/2$. For $X_j = 2f(\boldsymbol{w}_\infty; Z'_j)$, $\mathbb{E}[X_j \mid S] \le \varepsilon/32$ and $e^x \le 1 + (e-1)x$ on $[0, 1]$. Therefore the conditional rejection probability is at most

$$\exp\{-m\varepsilon[b_0 - (e-1)/32]\} \le e^{-m\varepsilon/2} \le 1/32 < \delta.$$

Integrating over S proves completeness. The relaxed certificate in (5.3), with $C_2 = 3$, gives the same acceptance guarantee when $m \ge 12/\varepsilon$: rejection implies $\widehat{R}_V > \varepsilon/3 - 1/m \ge \varepsilon/4$, so conditional Markov bounds its probability by $1/16$ whenever $R_D(S) \le \varepsilon/64$. No distinction is required for risks in $(\varepsilon/64, \varepsilon]$. □

These guarantees quantify the information needed to certify the prescribed interpolator. Training alone can still produce a predictor with $O(1/n)$ expected risk through stopping at $T = \Theta(n)$ or ridge regularization. The future-tail certificate also permits nonmonotone population risk during continued training.

N A FLUCTUATION BRIDGE FOR ARBITRARY QUADRATIC OBSERVATIONS

At the diagnostic horizon, the remaining finite-time gap is, for all sufficiently large n,

$$c\frac{\log\log n}{n} \le \mathcal{R}_{\mathsf{A}}(n, n, \lceil n \log n \rceil; 1) \le C\frac{\log n}{n}, \qquad \mathsf{A} \in \{\mathrm{GD}, \mathrm{SGD}\}. \tag{N.1}$$

Here $c, C > 0$ are absolute, both algorithms use step $1/4$, and the supremum is over arbitrary i.i.d. quadratic observations with the stated dimension and population-trace budget. The diagnostic construction in Appendix C proves the lower bound. The upper bound follows from Corollary D.2. The comparison below shows that the two algorithm-specific risks differ by only $O(1/n)$ at this horizon. It leaves their common sharp finite-time order unresolved.

The conditional-mean identity transfers GD lower bounds to SGD. The following additive comparison also transfers upper bounds when its remainder is small enough. It requires no rank-one assumption, commutation, or empirical spectral gap. Throughout this appendix, we work in target-zero error coordinates with initialization $\boldsymbol{w}_1 = \boldsymbol{u}$ and prefix average $\bar{\boldsymbol{w}}_T = T^{-1}\sum_{t=1}^{T} \boldsymbol{w}_t$, using the indexing of Section 2. For a constant step $0 < \eta \le 1$, write $g_{T,\eta}(H_S) = T^{-1}\sum_{j=0}^{T-1}(I - \eta H_S)^j$.

Theorem N.1 (Conditional PSD fluctuation bridge)**.** *Fix sample Hessians $0 \preceq H_i \preceq I$, a population Hessian $0 \preceq H \preceq I$, and $\|\boldsymbol{u}\| \le 1$. Put $H_S = n^{-1}\sum_i H_i$, $\Pi = P_{\operatorname{ran} H_S}$, $Z = \lambda^+_{\max}(H - 2H_S)$, and $F(\boldsymbol{w}) = \boldsymbol{w}^\top H\boldsymbol{w}/2$. For GD and fixed-dataset SGD using the same initialization and step $0 < \eta \le 1$, every $T \ge 1$ satisfies*

$$\begin{aligned} 0 \le \Delta_T(S) &:= \mathbb{E}_{\mathrm{alg}}[F(\bar{\boldsymbol{w}}_T^{\mathrm{SGD}}) \mid S] - F(g_{T,\eta}(H_S)\boldsymbol{u}) \\ &\le \frac{2 + nZ}{2(2-\eta)T}\|\Pi\boldsymbol{u}\|^2. \end{aligned} \tag{N.2}$$

At $\eta = 1/4$, the upper bound is $(4 + 2nZ)\|\Pi\boldsymbol{u}\|^2/(7T)$.

Proof. Operator bound. Set $P = I - \eta H_S$ and $M_s = \sum_{j=0}^{s-1} P^j$. Positivity gives $\ker H_S = \bigcap_i \ker H_i$, so $\operatorname{ran} H_i \subseteq \operatorname{ran} H_S$. Moreover, $H_i \preceq nH_S$ implies $H_i H_S^+ H_i \preceq nH_i$. Indeed, $(H_S^+)^{1/2} H_i (H_S^+)^{1/2} \preceq n\Pi$ bounds the squared operator norm of $(H_S^+)^{1/2} H_i^{1/2}$ by n. Its other Gram matrix is therefore at most nI, and congruence by $H_i^{1/2}$ gives the required bound. For every positive empirical eigenvalue a, $M_s(a) \le \min\{s, (\eta a)^{-1}\}$ and $aM_s(a)^2 \le s/\eta$. Since $H \preceq 2H_S + ZI$, these scalar inequalities give

$$H_i M_s H M_s H_i \preceq \frac{s}{\eta}\big(2H_i^2 + ZH_i H_S^+ H_i\big) \preceq \frac{s}{\eta}(2 + nZ)H_i. \tag{N.3}$$

The inverse bound is used only on $\operatorname{ran} H_S$, so it also applies when H_S is singular. The contraction assumption gives $H_i^2 \preceq H_i$.

Martingale expansion. Let $\mathcal{F}_t$ contain S and the indices preceding update t. The vectors $\boldsymbol{\zeta}_t = -\eta(H_{i_t} - H_S)\boldsymbol{w}_t$ are conditional martingale differences. Expanding $\boldsymbol{w}_{t+1} = P\boldsymbol{w}_t + \boldsymbol{\zeta}_t$ yields

$$\bar{\boldsymbol{w}}_T^{\mathrm{SGD}} - g_{T,\eta}(H_S)\boldsymbol{u} = \frac{1}{T}\sum_{t=1}^{T-1} M_{T-t}\boldsymbol{\zeta}_t.$$

The coefficients are deterministic conditional on S, so distinct-time cross terms have zero expectation. For any fixed PSD matrix B,

$$\mathbb{E}[\boldsymbol{\zeta}_t^\top B\boldsymbol{\zeta}_t \mid \mathcal{F}_t, S] = \eta^2\boldsymbol{w}_t^\top\left(\frac{1}{n}\sum_i H_iBH_i - H_SBH_S\right)\boldsymbol{w}_t \le \frac{\eta^2}{n}\sum_i \boldsymbol{w}_t^\top H_iBH_i\boldsymbol{w}_t.$$

Using (N.3) with $s = T - t$ therefore gives

$$\begin{aligned} 0 \le \Delta_T(S) &= \frac{1}{2T^2}\sum_{t=1}^{T-1}\mathbb{E}_{\mathrm{alg}}[\boldsymbol{\zeta}_t^\top M_{T-t}HM_{T-t}\boldsymbol{\zeta}_t \mid S] \\ &\le \frac{\eta(2+nZ)}{2T}\sum_{t=1}^{T-1}\mathbb{E}_{\mathrm{alg}}[\boldsymbol{w}_t^\top H_S\boldsymbol{w}_t \mid S]. \end{aligned}$$

Energy estimate. To bound the sum, subtract the invariant empirical-kernel component and set $\boldsymbol{e}_t = \boldsymbol{w}_t - (I-\Pi)\boldsymbol{u} \in \operatorname{ran} H_S$. Every update satisfies $\|(I-\eta H_i)\boldsymbol{e}_t\|^2 \le \|\boldsymbol{e}_t\|^2 - \eta(2-\eta)\boldsymbol{e}_t^\top H_i\boldsymbol{e}_t$. Conditional expectation and telescoping give $\sum_{t\ge1}\mathbb{E}_{\mathrm{alg}}[\boldsymbol{w}_t^\top H_S\boldsymbol{w}_t \mid S] \le \|\Pi\boldsymbol{u}\|^2/[\eta(2-\eta)]$. Substituting this estimate gives (N.2). For $T = 1$, both outputs equal $\boldsymbol{u}$ and the sums are empty. □

Corollary N.2 (Expected risk and diagnostic equivalence)**.** *For i.i.d. observations, deterministic initialization, and step $\eta = 1/4$, consider the quadratic class with $\operatorname{tr} H \le \tau$. Define $D_{n,d,\tau} := \min\{n, (28/3)(1+\log d), 9\log(1+2n\tau)\}$ and set $b_{n,d,T,\tau} := (4+2D_{n,d,\tau})/(7T)$. Let $R_{\mathsf{A}}(D,\boldsymbol{u},T) = \mathbb{E}F(\bar{\boldsymbol{w}}_T^{\mathsf{A}})$. Then*

$$0 \le R_{\mathrm{SGD}}(D,\boldsymbol{u},T) - R_{\mathrm{GD}}(D,\boldsymbol{u},T) \le b_{n,d,T,\tau}, \tag{N.4}$$

$$0 \le \mathcal{R}_{\mathrm{SGD}}(n,d,T;\tau) - \mathcal{R}_{\mathrm{GD}}(n,d,T;\tau) \le b_{n,d,T,\tau}. \tag{N.5}$$

In particular, for $T_n = \lceil n\log n\rceil$ and all sufficiently large n,

$$0 \le \mathcal{R}_{\mathrm{SGD}}(n,n,T_n;1) - \mathcal{R}_{\mathrm{GD}}(n,n,T_n;1) \le C/n, \tag{N.6}$$

$$1 \le \frac{\mathcal{R}_{\mathrm{SGD}}(n,n,T_n;1)}{\mathcal{R}_{\mathrm{GD}}(n,n,T_n;1)} \le 1 + \frac{C'}{\log\log n}. \tag{N.7}$$

Proof. Theorem D.1 gives $n\mathbb{E}Z \le D_{n,d,\tau}$. Averaging (N.2) and using $\|\Pi\boldsymbol{u}\|^2 \le 1$ yields (N.4). Taking suprema over the same admissible class gives (N.5), without requiring either supremum to be attained or both to share an extremizer. At the diagnostic horizon, $D_{n,n,1} = O(\log n)$ gives (N.6). The diagnostic construction in Appendix C gives $\mathcal{R}_{\mathrm{GD}}(n,n,T_n;1) \ge c\log\log n/n$. Dividing the additive bound by this lower bound establishes (N.7). □

The ratio in (N.7) compares class suprema and need not bound the risk ratio on an individual instance. Determining the sharp order within (N.1) remains open. The bridge reduces this upper-bound problem to GD but does not determine its answer.