BENIGN MISFITTING BEYOND ONE SPIKE

Anonymous authors
Paper under double-blind review

ABSTRACT

Benign misfitting occurs when accurate prediction requires larger empirical training loss than the zero predictor's. We show this phenomenon occurs for noiseless overparameterized Gaussian regression with signal spikes of unequal strengths and independent training and test pairs from one distribution. Specifically, we show that there exists a range of sample sizes in which every sufficiently accurate linear predictor in the span of the training data must misfit. So good generalization here cannot be explained by good training fit. A reflection argument extends this limitation to all learned linear predictors when no signal spike orientation is preferred in advance. We then characterize when a single epoch of zero-initialized stochastic gradient descent can reach predictors that generalize well, despite this requiring actually ascending in empirical training loss. An explicit pair of examples is given to separate the geometry forcing misfitting from whether such misfitting predictors are learnable by a single epoch of constant-rate SGD.

1 WHY UNEQUAL SIGNAL STRENGTHS MATTER

Must a predictor fit its training examples well to generalize well? Interpolation fits very well — matching every training label exactly — and interpolating predictors can generalize well in overparameterized regression (Bartlett et al., 2020; Tsigler & Bartlett, 2023). Benign misfitting is a different possibility. A predictor can generalize well yet fit its training data worse than the zero predictor does. In the problems studied here, the worse training fit is *necessary* for better generalization, even though both errors use squared loss and the iid training and test data come from the same distribution. A single pass of stochastic gradient descent (SGD) can nevertheless reach such a predictor, ending with much *larger* training loss than it started with.

Both findings were established in a stylized model (Ranade & Sahai, 2026). There, the label depends on one high-variance signal direction, called a covariance spike, and is independent of the coordinates along a large number d_0 of other low-variance directions, called nuisance directions. The test inputs they considered are Gaussian, but their training inputs are fixed, with equal spike coordinates and mutually orthogonal nuisance components. Their results concerned predictors formed as linear combinations of the training inputs, which we call *span predictors*.

Such a simple model invites the suspicion that the counterintuitive behavior is merely an artifact of the simplification. With one signal spike, the minimum-norm interpolator (MNI), the shortest predictor vector that fits every training label, is just a scaled-down version of the span predictor with smallest test risk. Several relevant signal directions with unequal variances can remove that rescaling shortcut.[1] Still, scalar inflation of predictors can improve test risk in more general anisotropic models (Freeman, 2025), and negative ridge can improve generalization with just one tuning parameter (Kobak et al., 2020). Does generalizing well still require worse training fit, and can a single SGD pass learn? What about learning algorithms that leave the span?

Even for unequal signal spike strengths, there is a range of training-data sizes n (8) in which accurate span predictors exist, but every span predictor that generalizes significantly better than the zero predictor must have much larger training loss than the zero predictor (Theorem 1). Because a span predictor is a linear combination of the training inputs, it necessarily retains nuisance components from the training inputs used to construct it. These components interact coherently with the training

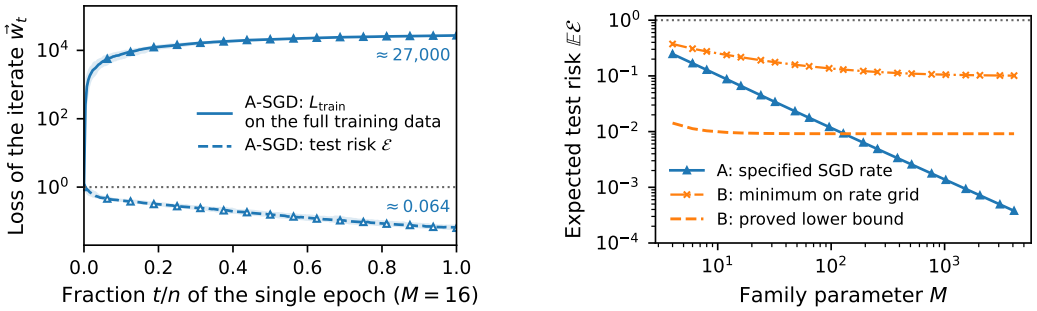
[1] Suppose a learned predictor had half the required amplitude for one such relevant direction, but a quarter of what is required for another relevant direction. No common multiplier would give both directions their required amplitudes (see also Zhao, 2026, Section 3.2).

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.

inputs they came from but only noncoherently with independent test inputs. The high training loss is a consequence of that coherent gain, a factor of about d_0/n (Section 2.1).

The true linear predictor fits every training label, but the training data cannot tell it apart from its reflection across the training span — the reflection fits every label just as well, and the correspondingly reflected signal eigenspaces assign the same density to every training point. We use this reflection observation to show that the necessity of misfitting extends to all learned linear predictors (Theorem 2) as long as no directions are a priori preferred to be signal directions. Essentially, moving perpendicular to the span changes no training prediction, but the amount that test risk improves in one world is more than offset by how much worse it gets in the other equally plausible mirror world.

We consider inputs in $\mathbb{R}^d$ with p signal spikes and $d_0 = d - p$ nuisance directions, where the label has unit variance and depends linearly on the signal subspace alone. A single epoch of zero-initialized SGD can learn such multispike models while ending with much larger training loss than it started with (Figure 1(a)), even at ordinary scales such as $n = 400$ training points and $d = 40{,}002$ (Section 5). Theorem 3 characterizes when some constant learning rate gives vanishing expected test risk uniformly over such labels. However, unlike the single-spike case, such a learning rate need not always exist even when an accurate span predictor does. Section 5 gives two closely related problem families, A and B, whose best span predictors have the same leading test risk and training loss. They differ only in the strengths of signal spike directions that happen to be irrelevant to the label, yet only A admits vanishing expected test risk under a constant-rate SGD pass (Figure 1(b)). However, a negative-ridge predictor tuned from the dimensions alone learns both (Appendix F.2), so B's failure to learn is specific to the constant-rate SGD pass.



(a) Single epoch in family A, $n = M^2 = 256$

(b) Expected risk of SGD in A and B

Figure 1: A single SGD pass can reduce test risk while increasing training loss, and directions irrelevant to the label can decide whether its expected test risk vanishes. Families A and B differ only in the strengths of signal spikes irrelevant to the label (Section 5). (a) Medians and pointwise 10th–90th percentiles across 64 data sets for test risk $\mathcal{E}(\vec{w}_t; \vec{\beta})$ and average training loss $L_{\text{train}}(\vec{w}_t; \vec{\beta})$ for family A, with every iterate scored on the same full training set. The first update alone raises the median training loss to about 260 (Appendix F.5). Theorem 1 proves that accurate span prediction *requires* large training loss in this family as M grows. (b) Exact finite-sample expected risk at A's specified rate, B's minimum over tested rates, and B's analytic lower bound (42) for every positive constant rate.

A smaller learning rate reduces error created in strong directions but can leave weaker label-bearing directions unlearned when the pass ends. Training longer on the same data meets different obstacles. Zero-initialized gradient descent (GD) and SGD stay in the training span. If repeated passes drive their training loss to zero, they approach the span's only interpolator, the MNI (see also Muthukumar et al., 2021, Section 2.1). An accurate span predictor in the range (8) must instead have more training loss than the zero predictor, so no path whose training loss never rises above its initial value can reach it. Starting from zero, updates that never enlarge the residual on the example they use keep test risk near that of the zero predictor, however many epochs they take (Proposition E.1 in Appendix E).

Our SGD proofs use the exact Gaussian least-mean-squares (LMS) dynamics of Bordelon & Pehlevan (2022, Theorem 1 and eq. (6)), in the classical setting of Sayed (2008, Chapters 22–23). Earlier "benign underfitting" examples show that SGD can generalize despite unfavorable training behavior (Koren et al., 2022). Here the converse requires poor training fit from *every* accurate span predic-

tor. Ranade & Sahai (2026, Section 7.1 and Appendix E) discuss this relationship in detail. By Theorem 1, no span predictor that fits the training data better than the zero predictor can generalize significantly better than it. So good generalization here cannot be explained by a small training loss plus a uniform-convergence or norm bound on the train–test gap (Nagarajan & Kolter, 2019; Koehler et al., 2021). Appendix G explains these relationships and the precise SGD results we use.

2 ACCURATE SPAN PREDICTORS MUST MISFIT

The training data are n independent input–label pairs $(\vec{x}_i, y_i)$, with $\vec{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Only coordinates in a p-dimensional signal subspace can contribute to the label. The remaining $d_0 = d-p$ orthogonal directions are nuisance directions, whose coordinates are independent of the label. The learner observes only the pairs $(\vec{x}_i, y_i)$, and we use coordinates aligned with the signal directions only to describe the distribution.

Let $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \dots, \gamma_p)$ contain the signal variances, which we call the spike strengths, ordered as $\gamma_1 \geq \cdots \geq \gamma_p > 1$, so every spike is stronger than every nuisance direction. With equal strengths, the input covariance is the bi-level covariance of Muthukumar et al. (2021, Definition 5), a standard model for benign overfitting. The vector $\vec{\beta} \in \mathbb{R}^p$, with $\|\vec{\beta}\| = 1$, gives the label coefficients in unit-variance signal coordinates. Each input–label pair $(\vec{x}_i, y_i)$ is generated by

$$\vec{x}_i = (\boldsymbol{\Gamma}^{1/2}\vec{s}_i, \vec{n}_i), \qquad y_i = \vec{\beta}^\top \vec{s}_i, \qquad \vec{s}_i \sim \mathcal{N}(\vec{0}, \boldsymbol{I}_p), \quad \vec{n}_i \sim \mathcal{N}(\vec{0}, \boldsymbol{I}_{d_0}). \tag{1}$$

The two Gaussian blocks in (1) are independent. The coordinates of $\vec{s}_i$ have unit variance, while the corresponding coordinates of $\vec{x}_i$ have variances γ_j. A signal coordinate j is irrelevant to the label of this particular target when $\beta_j = 0$. Norms are Euclidean, $\vec{e}_j$ is the jth coordinate unit vector, $\|\cdot\|_{\mathrm{op}}$ is the operator norm, and $\boldsymbol{A} \preceq \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

Let $\boldsymbol{X} = [\vec{x}_1, \dots, \vec{x}_n]$ and $\vec{y} = (y_1, \dots, y_n)^\top$. The standardized signal matrix $\boldsymbol{S} \in \mathbb{R}^{n\times p}$ has rows $\vec{s}_i^\top$, and the nuisance matrix is $\boldsymbol{N} = [\vec{n}_1, \dots, \vec{n}_n] \in \mathbb{R}^{d_0 \times n}$. Thus the blocks of $\boldsymbol{X}$ are $\boldsymbol{\Gamma}^{1/2}\boldsymbol{S}^\top$ and $\boldsymbol{N}$, and $\vec{y} = \boldsymbol{S}\vec{\beta} = \boldsymbol{X}^\top \vec{u}_{\vec{\beta}}$ for the true predictor $\vec{u}_{\vec{\beta}} = (\boldsymbol{\Gamma}^{-1/2}\vec{\beta}, \vec{0})$.

A linear predictor $\vec{u} \in \mathbb{R}^d$ predicts $\vec{u}^\top \vec{x}$. Write its signal and nuisance parts as $\vec{u} = (\vec{u}^{\mathrm{sig}}, \vec{u}^{\mathrm{nui}})$. The generalization error is the test risk $\mathcal{E}$, and the training error is the training loss L_{train}:

$$\mathcal{E}(\vec{u}; \vec{\beta}) := \mathbb{E}_{\mathrm{test}}(y - \vec{u}^\top \vec{x})^2 = \|\boldsymbol{\Gamma}^{1/2}\vec{u}^{\mathrm{sig}} - \vec{\beta}\|^2 + \|\vec{u}^{\mathrm{nui}}\|^2, \tag{2}$$

$$L_{\mathrm{train}}(\vec{u}; \vec{\beta}) := \frac{1}{n}\|\boldsymbol{X}^\top \vec{u} - \vec{y}\|^2 = \frac{1}{n}\|\boldsymbol{S}(\boldsymbol{\Gamma}^{1/2}\vec{u}^{\mathrm{sig}} - \vec{\beta}) + \boldsymbol{N}^\top \vec{u}^{\mathrm{nui}}\|^2. \tag{3}$$

In (2), hold $\vec{u}$ fixed and average over a new pair $(\vec{x}, y)$ drawn independently of the training data. The independent, zero-mean signal and nuisance blocks make their cross term vanish. We write $\mathbb{E}\mathcal{E}(\vec{u}; \vec{\beta})$ when we also average over the training data and any independent learner randomness. The zero predictor has test risk one because $\mathbb{E}y^2 = \|\vec{\beta}\|^2 = 1$. Its training loss is $L_0(\vec{\beta}) := L_{\mathrm{train}}(\vec{0}; \vec{\beta}) = \|\boldsymbol{S}\vec{\beta}\|^2/n$. For each target $\vec{\beta}$, $L_0(\vec{\beta}) \to 1$ in probability by the law of large numbers.

2.1 HOW THE TWO LOSSES DIFFER

For a span predictor $\vec{u} = \boldsymbol{X}\vec{\alpha}$, the coefficients $\vec{\alpha} \in \mathbb{R}^n$ determine both its signal and nuisance parts. Let $\boldsymbol{W} = \boldsymbol{N}^\top \boldsymbol{N}$ be the nuisance Gram matrix, whose entries are the inner products of the nuisance vectors. The signal part of $\vec{u}$ is $\boldsymbol{\Gamma}^{1/2}\boldsymbol{S}^\top\vec{\alpha}$. Taking its inner product with the fresh test input's signal part $\boldsymbol{\Gamma}^{1/2}\vec{s}$ introduces a second square-root factor, giving the signal response $\vec{a}$ and the prediction

$$\vec{a} = \boldsymbol{\Gamma}\boldsymbol{S}^\top\vec{\alpha}, \qquad \vec{u}^\top\vec{x} = \vec{a}^\top\vec{s} + (\boldsymbol{N}\vec{\alpha})^\top\vec{n}. \tag{4}$$

Accurate prediction requires the learned coefficients $\vec{a}$ of the standardized signal coordinates to be close to the true coefficients $\vec{\beta}$. Substituting $\vec{u} = \boldsymbol{X}\vec{\alpha}$ into (2) gives its generalization error

$$\mathcal{E}(\boldsymbol{X}\vec{\alpha}; \vec{\beta}) = \underbrace{\|\vec{a} - \vec{\beta}\|^2}_{\text{signal error}} + \underbrace{\vec{\alpha}^\top \boldsymbol{W}\vec{\alpha}}_{\text{nuisance variance}}. \tag{5}$$

On the training data, $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{W} + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S}^\top$. Thus, n times the training error is

$$nL_{\mathrm{train}}(\boldsymbol{X}\vec{\alpha}; \vec{\beta}) = \|\boldsymbol{W}\vec{\alpha} + \boldsymbol{S}(\vec{a} - \vec{\beta})\|^2. \tag{6}$$

On fresh test data, the nuisance components that the predictor retains from the training inputs contribute the variance $\vec{\alpha}^\top \boldsymbol{W}\vec{\alpha}$ in (5). Evaluating the same predictor on the training inputs gives the nuisance prediction vector $\boldsymbol{W}\vec{\alpha}$ in (6). At $\vec{a} = \vec{\beta}$, training loss is therefore $\vec{\alpha}^\top \boldsymbol{W}^2\vec{\alpha}/n$, about d_0/n times test risk when $\boldsymbol{W}/d_0$ is close to the identity. The converse (10) bounds how much signal error can cancel this inherently amplified contribution.

For this section and Section 3, assume $1 \le p \le n \le d_0$. Then $\boldsymbol{N}$ and $\boldsymbol{S}$ have full column rank almost surely. In particular, $\boldsymbol{W}$ is positive definite ($\boldsymbol{W} \succ 0$), so the test risk in (5) is a strictly convex quadratic in $\vec{\alpha}$. Let $\vec{u}_*(\vec{\beta})$ be the unique span predictor with smallest test risk. This predictor describes what the training span permits when the true distribution is known.

Theorem 1 describes the losses of the best span predictor $\vec{u}_*(\vec{\beta})$ using two quantities:

$$\tau = \frac{d_0}{n}, \qquad B(\vec{\beta})^2 = \|\boldsymbol{\Gamma}^{-1}\vec{\beta}\|^2 = \sum_{j=1}^{p} \frac{\beta_j^2}{\gamma_j^2}. \tag{7}$$

Here τ is the factor d_0/n above, and $B(\vec{\beta})$ favors target components in weaker directions. Theorem 1 keeps the training inputs $\boldsymbol{X}$ fixed while comparing all unit targets, with labels $\vec{y} = \boldsymbol{S}\vec{\beta}$ changing accordingly. Dimensions and strengths may vary with n along the sequence of problems.

Theorem 1 (Accurate span predictors exist and must misfit)**.** *Along a family of problems* (1)*, suppose*

$$n \to \infty, \qquad \frac{p}{n} \to 0, \qquad \frac{n}{d_0} \to 0, \qquad \frac{d_0}{\gamma_p^2} \ll n \ll \frac{d_0}{\gamma_1}. \tag{8}$$

Here $a_n \ll b_n$ means $a_n/b_n \to 0$. The best span predictor satisfies

$$\sup_{\|\vec{\beta}\|=1} \left| \frac{\mathcal{E}(\vec{u}_*(\vec{\beta}); \vec{\beta})}{\tau B(\vec{\beta})^2} - 1 \right| \xrightarrow{\mathbb{P}} 0, \qquad \sup_{\|\vec{\beta}\|=1} \left| \frac{L_{\text{train}}(\vec{u}_*(\vec{\beta}); \vec{\beta})}{\tau^2 B(\vec{\beta})^2} - 1 \right| \xrightarrow{\mathbb{P}} 0. \tag{9}$$

Furthermore, fix $g \in (0,1)$. There is a deterministic $\epsilon_n(g) \to 0$ and an event determined only by the input matrix $\boldsymbol{X}$ whose probability tends to one. On this event, simultaneously for every unit target $\vec{\beta}$ and every span predictor $\vec{u}$ with $\mathcal{E}(\vec{u}; \vec{\beta}) \le g$, both the training loss and its ratio to the zero predictor's training loss $L_0(\vec{\beta})$ obey

$$\min\left\{ L_{\text{train}}(\vec{u}; \vec{\beta}), \frac{L_{\text{train}}(\vec{u}; \vec{\beta})}{L_0(\vec{\beta})} \right\} \ge [1 - \epsilon_n(g)](1 - \sqrt{g})^2(\tau/\gamma_1)^2. \tag{10}$$

The sample-size condition in (8) is equivalently $\gamma_1 \ll \tau \ll \gamma_p^2$. The right inequality gives vanishing best-span test risk in (9), since $\tau B(\vec{\beta})^2 \le \tau/\gamma_p^2 \to 0$. The left makes the training-loss bound (10) diverge. Even halving the zero predictor's test risk, by taking $g = 1/2$, requires this poor fit. The *converse* (10) rules out accurate span predictors with small training loss, including the MNI.

Proof of the converse. Suppose $\mathcal{E}(\boldsymbol{X}\vec{\alpha}; \vec{\beta}) \le g$, with $g \in (0,1)$ as in the theorem. Equation (5) gives $\|\vec{a} - \vec{\beta}\| \le \sqrt{g}$, so $\|\vec{a}\| \ge 1 - \sqrt{g}$. Producing this response requires a sufficiently large coefficient vector $\vec{\alpha}$. Applying (4) and the definition of the operator norm gives

$$\|\boldsymbol{S}\|_{\text{op}}\|\vec{\alpha}\| \ge \|\boldsymbol{S}^\top\vec{\alpha}\| = \|\boldsymbol{\Gamma}^{-1}\vec{a}\| \ge \frac{\|\vec{a}\|}{\gamma_1} \ge \frac{1 - \sqrt{g}}{\gamma_1}. \tag{11}$$

Write $[x]_+ = \max\{x, 0\}$. Starting from (6), the reverse triangle inequality limits how much the signal error can cancel the nuisance contribution. Together with (11), it gives

$$\sqrt{nL_{\text{train}}} \ge [\lambda_{\min}(\boldsymbol{W})\|\vec{\alpha}\| - \|\boldsymbol{S}\|_{\text{op}}\sqrt{g}]_+ \tag{12}$$

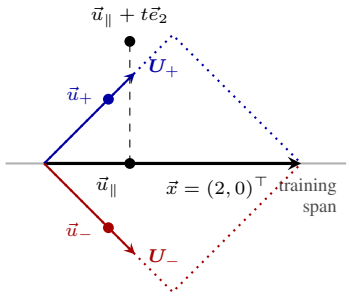
$$\ge \left[\lambda_{\min}(\boldsymbol{W}) \left(\frac{1 - \sqrt{g}}{\|\boldsymbol{S}\|_{\text{op}}\gamma_1} \right) - \|\boldsymbol{S}\|_{\text{op}}\sqrt{g} \right]_+ . \tag{13}$$

Gaussian singular-value bounds (Davidson & Szarek, 2001, Theorem II.13) give one event of probability tending to one on which both $\boldsymbol{W}/d_0$ and $\boldsymbol{S}^\top\boldsymbol{S}/n$ approach their identity matrices in operator

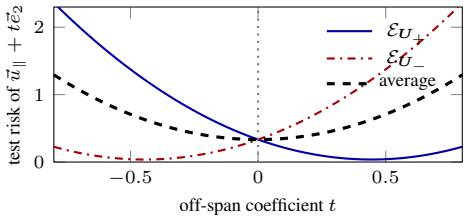
norm. On this event, $\lambda_{\min}(\boldsymbol{W}) = (1 - o(1))d_0$ and $\|\boldsymbol{S}\|_{\mathrm{op}} = (1 + o(1))\sqrt{n}$. Dividing (13) by $\sqrt{n}$ makes its first term $(1 - o(1))(d_0/n)((1 - \sqrt{g})/\gamma_1)$ and its second term $(1 + o(1))\sqrt{g}$. The first term diverges since $(n/d_0)\gamma_1 \to 0$ by (8). The same event gives $L_0(\vec{\beta}) = 1 + o(1)$ uniformly over unit targets $\vec{\beta}$. Squaring, with $d_0/n = \tau$, proves (10). Appendix A.1 gives the finite probability bounds, and Appendices A.2–A.3 prove (9). □

3 Unknown spike directions make the span sufficient

Theorem 1 applies only to span predictors. For an arbitrary learned linear predictor $\vec{u}$, write $\vec{u} = \vec{u}_{\parallel} + \vec{u}_{\perp}$, where $\vec{u}_{\parallel}$ lies in the training span and $\vec{u}_{\perp}$ is perpendicular to it. This decomposition uses the observed training inputs, not the signal–nuisance split. Since $\vec{u}_{\perp}^{\top}\vec{x}_i = 0$ for every training input, removing this *off-span* component preserves every training prediction. Figure 2 shows why an off-span gain in one frame does not improve average or worst-case risk over reflected frames.



(a) Two worlds behind one observation



(b) Exact risk along the off-span line

Figure 2: An off-span move that helps one compatible model hurts its mirror more. With $\gamma = 2$ and $\beta = 1$, the colored directions $\boldsymbol{U}_{\pm} = (1, \pm 1)^{\top}/\sqrt{2}$ both give label $y = 1$ at the observed input $\vec{x} = (2, 0)^{\top}$ and give that input the same Gaussian density, because it has the same projection length on both (the dotted right triangles are congruent). The colored dots are their zero-risk predictors $\vec{u}_{\pm} = \boldsymbol{U}_{\pm}/\sqrt{2}$. The black dot $\vec{u}_{\parallel} = (2/3, 0)^{\top}$ minimizes test risk within the training span in both worlds. Moving to $\vec{u}_{\parallel} + t\vec{e}_2$ leaves the training prediction unchanged. The two test risks plotted in (b), $\frac{1}{3} + \frac{3}{2}t^2 \mp \frac{4}{3}t$, have opposite linear terms, so their average and their maximum are both minimized at $t = 0$.

The two frames in Figure 2, the two possible *worlds*, give the observed input the same label and density. Modeling the pair as equally plausible makes every nonzero off-span move increase average risk. The pair's maximum risk also increases. We now extend this reflected-pair comparison to several signal directions and the full training span.

Keep the strengths $\boldsymbol{\Gamma}$ and target coefficients $\vec{\beta}$ fixed while varying the signal directions. A *spike frame* $\boldsymbol{U} \in \mathbb{R}^{d \times p}$ records these directions in its orthonormal columns. An input's component along column j has variance γ_j. Its component along any unit direction perpendicular to all columns of $\boldsymbol{U}$ has variance one. The coefficient on column j required to predict the label is $\beta_j/\sqrt{\gamma_j}$. Thus the covariance $\boldsymbol{\Sigma}_{\boldsymbol{U}}$, zero-risk predictor $\vec{u}_{\boldsymbol{U},\vec{\beta}}$, and test risk $\mathcal{E}_{\boldsymbol{U}}$ are

$$\boldsymbol{\Sigma}_{\boldsymbol{U}} = \boldsymbol{I}_d + \boldsymbol{U}(\boldsymbol{\Gamma} - \boldsymbol{I}_p)\boldsymbol{U}^{\top}, \qquad \vec{u}_{\boldsymbol{U},\vec{\beta}} = \boldsymbol{U}\boldsymbol{\Gamma}^{-1/2}\vec{\beta}, \tag{14}$$

$$\mathcal{E}_{\boldsymbol{U}}(\vec{u}; \vec{\beta}) = (\vec{u} - \vec{u}_{\boldsymbol{U},\vec{\beta}})^{\top}\boldsymbol{\Sigma}_{\boldsymbol{U}}(\vec{u} - \vec{u}_{\boldsymbol{U},\vec{\beta}}). \tag{15}$$

At frame $\boldsymbol{U}$, the label is $y = \vec{u}_{\boldsymbol{U},\vec{\beta}}^{\top}\vec{x}$. Section 2's aligned coordinates give the frame $\boldsymbol{U}_0 = (\boldsymbol{I}_p, \boldsymbol{0})^{\top}$. An orthogonal matrix $\boldsymbol{H}$ rotates this frame to $\boldsymbol{U} = \boldsymbol{H}\boldsymbol{U}_0$ and rotates each input in (1), while leaving its label unchanged. The learner observes the rotated training data $D = (\boldsymbol{X}, \vec{y})$, not $\boldsymbol{H}$.

The learner does not know the true $\boldsymbol{U}$ and knows nothing about it a priori. Every frame gives $\boldsymbol{X}$ positive Gaussian density, but only some give the observed labels. Call a frame $(\vec{\beta}, \boldsymbol{\Gamma})$-*compatible* with observed training data D when its zero-risk predictor assigns each training input $\vec{x}_i$ the observed label y_i. The compatible set is

$$\mathcal{U}_{\vec{\beta},\boldsymbol{\Gamma},D} = \{\widetilde{\boldsymbol{U}} : \widetilde{\boldsymbol{U}}^{\top}\widetilde{\boldsymbol{U}} = \boldsymbol{I}_p,\ \boldsymbol{X}^{\top}\widetilde{\boldsymbol{U}}\boldsymbol{\Gamma}^{-1/2}\vec{\beta} = \vec{y}\}. \tag{16}$$

The true generating frame belongs to $\mathcal{U}_{\vec{\beta},\mathbf{\Gamma},D}$. Keeping $(\vec{\beta},\mathbf{\Gamma})$ fixed specifies the distributions being compared, not information supplied to the learner. Two compatible frames can give a training input different densities.

For the average comparison, we model the absence of a preferred orientation by the rotation-invariant probability law on frames (Eaton, 1989). Conditioning this prior and the data model on D gives the posterior law $\Pi_{\vec{\beta},\mathbf{\Gamma},D}$, supported on the compatible set $\mathcal{U}_{\vec{\beta},\mathbf{\Gamma},D}$. The worst-case comparison uses no prior and compares only frames that explain every observed input equally well. With $f_{\widetilde{\boldsymbol{U}}}$ the Gaussian input density at frame $\widetilde{\boldsymbol{U}}$, two compatible frames $\boldsymbol{U}$ and $\boldsymbol{V}$ belong to the same *equal-density class* $\mathcal{C}$ when $f_{\boldsymbol{U}}(\vec{x}_i) = f_{\boldsymbol{V}}(\vec{x}_i)$ for every i. For a predictor $\vec{u}$, define its posterior-average test risk and its worst-case test risk over an equal-density class $\mathcal{C}$ by

$$\mathcal{R}^{\mathrm{post}}_{\vec{\beta},\mathbf{\Gamma},D}(\vec{u}) = \int \mathcal{E}_{\widetilde{\boldsymbol{U}}}(\vec{u};\vec{\beta})\,\Pi_{\vec{\beta},\mathbf{\Gamma},D}(d\widetilde{\boldsymbol{U}}), \tag{17}$$

$$\mathcal{R}^{\mathrm{wc}}_{\vec{\beta},\mathbf{\Gamma},D}(\vec{u};\mathcal{C}) = \sup_{\widetilde{\boldsymbol{U}}\in\mathcal{C}} \mathcal{E}_{\widetilde{\boldsymbol{U}}}(\vec{u};\vec{\beta}). \tag{18}$$

Since $1 \le p \le n \le d_0$, $\boldsymbol{X}$ has full column rank almost surely. The orthogonal projector $\boldsymbol{P}_X$ onto the training span and the reflection $\boldsymbol{J}_X$ across it give explicit formulas for $\vec{u}_\parallel$ and $\vec{u}_\perp$:

$$\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top, \quad \boldsymbol{J}_X = 2\boldsymbol{P}_X - \boldsymbol{I}_d, \quad \vec{u}_\parallel = \boldsymbol{P}_X\vec{u}, \quad \vec{u}_\perp = (\boldsymbol{I}_d - \boldsymbol{P}_X)\vec{u}. \tag{19}$$

Theorem 2 (Projection suffices under either frame criterion)**.** *Fix strengths $\mathbf{\Gamma}$ and a unit target $\vec{\beta}$, with $1 \le p \le n \le d_0$. On one event of probability one for a frame drawn from the rotation-invariant law and training data then drawn at that frame, every predictor $\vec{u}$ decomposed as in* (19) *satisfies*

$$\mathcal{R}^{\mathrm{post}}_{\vec{\beta},\mathbf{\Gamma},D}(\vec{u}) = \mathcal{R}^{\mathrm{post}}_{\vec{\beta},\mathbf{\Gamma},D}(\vec{u}_\parallel) + \int \vec{u}_\perp^\top \mathbf{\Sigma}_{\widetilde{\boldsymbol{U}}}\vec{u}_\perp\,\Pi_{\vec{\beta},\mathbf{\Gamma},D}(d\widetilde{\boldsymbol{U}}). \tag{20}$$

At every training data set with full-rank $\boldsymbol{X}$ and for every equal-density class $\mathcal{C}$, every such $\vec{u}$ satisfies

$$\mathcal{R}^{\mathrm{wc}}_{\vec{\beta},\mathbf{\Gamma},D}(\vec{u};\mathcal{C}) \ge \mathcal{R}^{\mathrm{wc}}_{\vec{\beta},\mathbf{\Gamma},D}(\vec{u}_\parallel;\mathcal{C}) + \|\vec{u}_\perp\|^2. \tag{21}$$

Also $\boldsymbol{X}^\top\vec{u} = \boldsymbol{X}^\top\vec{u}_\parallel$ and $\|\vec{u}\|^2 = \|\vec{u}_\parallel\|^2 + \|\vec{u}_\perp\|^2$. The integral in (20) *is at least $\|\vec{u}_\perp\|^2$, because every $\mathbf{\Sigma}_{\widetilde{\boldsymbol{U}}} \succeq \boldsymbol{I}_d$. Under either criterion, every attainable risk and training-loss pair is matched or improved by a span predictor.*

Proof. The reflection $\boldsymbol{J}_X$ in (19) is symmetric and orthogonal and fixes every column of $\boldsymbol{X}$. Hence $\boldsymbol{X}^\top\boldsymbol{J}_X\boldsymbol{U} = \boldsymbol{X}^\top\boldsymbol{U}$, so (16) gives the same training labels at $\boldsymbol{U}$ and $\boldsymbol{J}_X\boldsymbol{U}$. Substituting $\boldsymbol{J}_X\boldsymbol{U}$ into (14) and using $\boldsymbol{J}_X^2 = \boldsymbol{I}_d$ gives its covariance and zero-risk predictor,

$$\mathbf{\Sigma}_{\boldsymbol{J}_X\boldsymbol{U}} = \boldsymbol{J}_X\mathbf{\Sigma}_{\boldsymbol{U}}\boldsymbol{J}_X, \qquad \vec{u}_{\boldsymbol{J}_X\boldsymbol{U},\vec{\beta}} = \boldsymbol{J}_X\vec{u}_{\boldsymbol{U},\vec{\beta}}. \tag{22}$$

The covariance identity in (22) preserves the Gaussian density at each $\vec{x}_i$ (Appendix B). For fixed inputs, the frame prior is invariant too, so the posterior law is invariant under this reflection.

To compare the risks at $\boldsymbol{U}$ and $\boldsymbol{J}_X\boldsymbol{U}$, substitute (22) into (15). Since $\boldsymbol{J}_X\vec{u} = \vec{u}_\parallel - \vec{u}_\perp$,

$$\mathcal{E}_{\boldsymbol{J}_X\boldsymbol{U}}(\vec{u};\vec{\beta}) = (\boldsymbol{J}_X\vec{u} - \vec{u}_{\boldsymbol{U},\vec{\beta}})^\top\mathbf{\Sigma}_{\boldsymbol{U}}(\boldsymbol{J}_X\vec{u} - \vec{u}_{\boldsymbol{U},\vec{\beta}}) = \mathcal{E}_{\boldsymbol{U}}(\vec{u}_\parallel - \vec{u}_\perp;\vec{\beta}). \tag{23}$$

Reflecting the frame changes the sign of the off-span component in the risk. Write $\vec{\delta} = \vec{u}_\parallel - \vec{u}_{\boldsymbol{U},\vec{\beta}}$. Expanding $\mathcal{E}_{\boldsymbol{U}}(\vec{u}_\parallel \pm \vec{u}_\perp;\vec{\beta}) = (\vec{\delta} \pm \vec{u}_\perp)^\top\mathbf{\Sigma}_{\boldsymbol{U}}(\vec{\delta} \pm \vec{u}_\perp)$ by (15) and averaging the two signs gives

$$\begin{aligned}\frac{\mathcal{E}_{\boldsymbol{U}}(\vec{u};\vec{\beta}) + \mathcal{E}_{\boldsymbol{J}_X\boldsymbol{U}}(\vec{u};\vec{\beta})}{2} &= \vec{\delta}^\top\mathbf{\Sigma}_{\boldsymbol{U}}\vec{\delta} + \vec{u}_\perp^\top\mathbf{\Sigma}_{\boldsymbol{U}}\vec{u}_\perp + \underbrace{\vec{u}_\perp^\top\mathbf{\Sigma}_{\boldsymbol{U}}\vec{\delta} - \vec{u}_\perp^\top\mathbf{\Sigma}_{\boldsymbol{U}}\vec{\delta}}_{\text{the cross terms cancel}} \\ &= \mathcal{E}_{\boldsymbol{U}}(\vec{u}_\parallel;\vec{\beta}) + \vec{u}_\perp^\top\mathbf{\Sigma}_{\boldsymbol{U}}\vec{u}_\perp.\end{aligned} \tag{24}$$

Posterior invariance gives the two risks on the left of (24) the same posterior integral, so integrating (24) proves (20). Reflection preserves the labels and the density at each $\vec{x}_i$, so $\boldsymbol{U} \in \mathcal{C}$ exactly when $\boldsymbol{J}_X\boldsymbol{U} \in \mathcal{C}$. For each $\boldsymbol{U} \in \mathcal{C}$, the worst risk over $\mathcal{C}$ is at least the average of $\mathcal{E}_{\boldsymbol{U}}(\vec{u};\vec{\beta})$ and $\mathcal{E}_{\boldsymbol{J}_X\boldsymbol{U}}(\vec{u};\vec{\beta})$. By (24) and $\mathbf{\Sigma}_{\boldsymbol{U}} \succeq \boldsymbol{I}_d$, that average is at least $\mathcal{E}_{\boldsymbol{U}}(\vec{u}_\parallel;\vec{\beta}) + \|\vec{u}_\perp\|^2$. Taking the supremum over $\boldsymbol{U} \in \mathcal{C}$ proves (21). Since $\boldsymbol{X}^\top\vec{u}_\perp = \vec{0}$, projection keeps the training predictions and training loss and lowers $\|\vec{u}\|^2$ by $\|\vec{u}_\perp\|^2$. Appendix B proves the posterior is reflection-invariant. □

These comparisons extend the misfitting requirement to learned linear predictors outside the span. Projection uses only $\boldsymbol{X}$ and preserves training loss. Under (8), any learning rule whose test risk (averaged over hidden frames and training data) tends to zero must therefore have training loss diverging in probability. For the worst case, the same $\vec{u}_{\|}$ satisfies (21) in every class, including the class $\mathcal{C}_{\mathrm{gen}}$ of the generating frame, which the learner does not know. So a predictor whose worst-case test risk over $\mathcal{C}_{\mathrm{gen}}$ is at most $g \in (0,1)$ obeys the converse (10) on the event of Theorem 1. Appendix B.1 proves both conclusions.

Because they concern the returned linear predictor, these conclusions also cover randomly initialized SGD and Adam (Kingma & Ba, 2015) in this model, even when their resulting predictors leave the span.

4 When does a constant-rate SGD pass learn? Two ratios decide

Accurate span predictors exist in the range (8) of Theorem 1. We now characterize when one constant-rate SGD pass achieves vanishing expected test risk. Use each independent training pair $(\vec{x}_i, y_i)$ once, in an order fixed before seeing the data, with one positive learning rate. Starting at zero, the update is

$$\vec{w}_0 = \vec{0}, \qquad \vec{w}_{t+1} = \vec{w}_t + \eta(y_{t+1} - \vec{w}_t^{\top}\vec{x}_{t+1})\vec{x}_{t+1}, \qquad 0 \le t < n. \tag{25}$$

Update (25) is a gradient step on $\frac{1}{2}(y_{t+1} - \vec{w}_t^{\top}\vec{x}_{t+1})^2$ for a fresh pair, independent of the earlier pairs that determine $\vec{w}_t$. Each update adds a multiple of the next input, so $\vec{w}_t$ stays in the span of the inputs already used. For the existence claim in Theorem 3, a rate may be supplied using the strengths and dimensions, but it is chosen before drawing the training data and does not use the hidden directions or target coefficients. Here p, d_0, n may be arbitrary positive integers.

Write $\vec{w}_t = (\vec{w}_t^{\mathrm{sig}}, \vec{w}_t^{\mathrm{nui}})$ in the aligned coordinates. For signal coordinate j, the iterate puts coefficient $\sqrt{\gamma_j} w_{t,j}^{\mathrm{sig}}$ on the standardized feature s_j, whose true coefficient is β_j. The standardized signal residual $\vec{r}_t = \vec{\beta} - \boldsymbol{\Gamma}^{1/2}\vec{w}_t^{\mathrm{sig}}$ therefore lists the errors in these coefficients. With the nuisance part $\vec{v}_t = \vec{w}_t^{\mathrm{nui}}$, equation (2) gives $\mathcal{E}(\vec{w}_t; \vec{\beta}) = \|\vec{r}_t\|^2 + \|\vec{v}_t\|^2$. Averaging over the training pairs used by time t, define the expected test-risk contribution $T_{t,j}$ of signal coordinate j, their total T_t, and the expected nuisance variance N_t by

$$T_{t,j} = \mathbb{E} r_{t,j}^2, \qquad T_t = \sum_{j=1}^{p} T_{t,j}, \qquad N_t = \mathbb{E}\|\vec{v}_t\|^2, \qquad \mathbb{E}\mathcal{E}(\vec{w}_t; \vec{\beta}) = T_t + N_t. \tag{26}$$

Specializing Bordelon & Pehlevan (2022, Theorem 1 and eq. (6)) to one fresh Gaussian example per update, letting $\rho_j = (1-\eta\gamma_j)^2 + \eta^2\gamma_j^2$ and $\mu = (1-\eta)^2 + \eta^2(d_0+1)$ be the signal and nuisance multipliers, we get the recurrences

$$T_{t+1,j} = \rho_j T_{t,j} + \eta^2\gamma_j^2(T_t + N_t), \qquad T_{0,j} = \beta_j^2, \tag{27}$$

$$N_{t+1} = \mu N_t + \eta^2 d_0 T_t, \qquad N_0 = 0. \tag{28}$$

Every multiplier in (27) and (28) is nonnegative at every rate. In (27), the first term carries forward existing error, and the second is new error from stepping along a random input, proportional to γ_j^2 even when $\beta_j = 0$. Taking conditional means one update at a time (Appendix D) gives the mean signal residual

$$\mathbb{E} r_{t,j} = (1-\eta\gamma_j)^t \beta_j. \tag{29}$$

So a coordinate with $\beta_j = 0$ has zero mean residual (29), yet $T_{1,j} = \eta^2\gamma_j^2$ from $T_0 = 1$ and $N_0 = 0$. Appendix C obtains (27) and (28) from the cited Gaussian moment identity.

To state the one-pass learning criterion, define two ratios:

$$R_{\mathrm{wk}} = \frac{n\gamma_p^2}{d_0}, \qquad q_n = \frac{\operatorname{tr}\boldsymbol{\Gamma}}{n\gamma_p} = \frac{1}{n}\sum_{j=1}^{p}\frac{\gamma_j}{\gamma_p}. \tag{30}$$

The condition $R_{\mathrm{wk}} \to \infty$ is the same as $d_0/\gamma_p^2 \ll n$ in (8). The trace ratio q_n adds the strengths of all signal directions, including those irrelevant to the label, relative to $n\gamma_p$. At one spike it is just

$1/n$. Keeping the weakest strength γ_p fixed, strengthening a signal direction irrelevant to the label increases q_n without changing the target.

Here consistency means that expected test risk tends to zero. For the floor (32), write $H_j = \sum_{i:\gamma_i \geq \gamma_j} \gamma_i$ for the total strength of signal directions at least as strong as direction j.

Theorem 3 (Two ratios characterize constant-rate consistency)**.** *Consider a family of Gaussian problems* (1) *in which the deterministic constant rate η_n for the SGD pass of length n is chosen without seeing the realized training data or the target $\vec{\beta}$. Write $\vec{w}_n^{(\eta_n)}(\vec{\beta})$ for the final iterate of* (25) *at rate η_n and target $\vec{\beta}$. Then*

$$\left[\exists\,(\eta_n > 0): \sup_{\|\vec{\beta}\|=1} \mathbb{E}\mathcal{E}(\vec{w}_n^{(\eta_n)}(\vec{\beta}); \vec{\beta}) \to 0\right] \iff R_{\mathrm{wk}} \to \infty \text{ and } q_n \to 0. \tag{31}$$

For any positive strengths, any n, and every deterministic positive constant rate, the final iterate also satisfies

$$\mathbb{E}\mathcal{E}(\vec{w}_n; \vec{\beta}) \geq \sum_{j=1}^{p} \beta_j^2 \left(\frac{H_j}{H_j + 2\gamma_j}\right)^{2n}. \tag{32}$$

In (31) one rate serves every target, and the supremum is outside the expectation, unlike Theorem 1's simultaneous guarantee on one data set. The floor (32) bounds expected risk, which a few runs with very large risk can dominate (Appendix F.6). In family B of Section 5, the median final test risk of simulated runs is also far above the best-span risk at every tested rate (Appendix F.7).

Why the two conditions suffice. Write $\eta = \eta_n$ for the rate at sample size n. Appendix D.2 proves that the following limits suffice by bounding the errors accumulated in (27)–(28):

$$n\eta\gamma_p \to \infty, \qquad n\eta^2(d_0 + 2) \to 0, \qquad \eta \operatorname{tr} \mathbf{\Gamma} \to 0. \tag{33}$$

The first asks for progress in the weakest direction within the n updates. The second keeps the expected nuisance variance N_t, which starts at zero, small over those updates only, so it is not a stability condition and allows $\eta_n d_0 \to \infty$. The third controls the error created in the signal directions. Together they ask for $1/(n\gamma_p) \ll \eta_n \ll \min\{1/\sqrt{n(d_0+2)}, 1/\operatorname{tr}\mathbf{\Gamma}\}$. The lower end divided by the two upper ends gives $\sqrt{(1 + 2/d_0)/R_{\mathrm{wk}}}$ and q_n, which must both vanish for such a rate to exist. When $R_{\mathrm{wk}} \to \infty$ and $q_n \to 0$, the geometric mean of the lower end and the smaller upper end works (Appendix D.2 gives an explicit choice). □

Why the signal trace gives a lower bound. The lower bound comes from a conflict between learning a weaker direction and controlling the other, stronger signal directions. Measure error in the original signal coefficients rather than the standardized ones. Coefficient j has error $\beta_j/\sqrt{\gamma_j} - w_{t,j}^{\mathrm{sig}} = r_{t,j}/\sqrt{\gamma_j}$, so the expected squared error is $V_t = \sum_j T_{t,j}/\gamma_j = \mathbb{E}\|\mathbf{\Gamma}^{-1/2}\vec{\beta} - \vec{w}_t^{\mathrm{sig}}\|^2$. Dividing each $T_{t,j}$ by γ_j makes the first-order decrease exactly $-2\eta T_t$, while the source terms $\eta^2\gamma_j^2(T_t + N_t)$ of (27), divided by γ_j, sum to $\eta^2 \operatorname{tr}\mathbf{\Gamma}(T_t + N_t)$. Dividing (27) by γ_j, summing, and using $N_t \geq 0$ and $\gamma_j \geq \gamma_p$ gives

$$V_{t+1} - V_t = -2\eta T_t + 2\eta^2 \sum_j \gamma_j T_{t,j} + \eta^2 \operatorname{tr}\mathbf{\Gamma}(T_t + N_t) \tag{34}$$

$$\geq \eta[\eta(\operatorname{tr}\mathbf{\Gamma} + 2\gamma_p) - 2]T_t. \tag{35}$$

First take the unit target $\vec{\beta} = \vec{e}_p$. If $\eta \geq 2/(\operatorname{tr}\mathbf{\Gamma} + 2\gamma_p)$, then V_t is nondecreasing by (35), so $\mathbb{E}\mathcal{E}(\vec{w}_n; \vec{e}_p) \geq T_n \geq \gamma_p V_n \geq \gamma_p V_0 = 1$. Below this threshold, Jensen's inequality and (29) give

$$\mathbb{E}\mathcal{E}(\vec{w}_n; \vec{e}_p) \geq (1 - \eta\gamma_p)^{2n} \geq \left(\frac{\operatorname{tr}\mathbf{\Gamma}}{\operatorname{tr}\mathbf{\Gamma} + 2\gamma_p}\right)^{2n}. \tag{36}$$

Appendix D.1 extends this bound to every target by comparing with smaller problems and using linearity in $\beta_1^2, \ldots, \beta_p^2$, which proves (32).

At $\vec{\beta} = \vec{e}_p$, $H_p = \operatorname{tr}\mathbf{\Gamma}$, so the floor (32) is at least e^{-4/q_n} by $(1 + x)^{-1} \geq e^{-x}$. Thus (31) requires $q_n \to 0$. Appendix D.3 proves the remaining necessity $R_{\mathrm{wk}} \to \infty$ by expanding the nuisance recurrence (28). That step uses $\gamma_p > 1$. □

The learning conditions of Theorem 3 and the misfitting conditions of Theorem 1 hold together when

$$\max\left\{\frac{d_0}{\gamma_p^2}, \frac{\operatorname{tr}\mathbf{\Gamma}}{\gamma_p}\right\} \ll n \ll \frac{d_0}{\gamma_1}. \tag{37}$$

Condition (37) implies all assumptions of Theorems 1 and 3 (Appendix D.4). At a rate satisfying (33), Markov's inequality puts the final test risk below any fixed $g \in (0,1)$ with probability tending to one. Theorem 1 therefore gives

$$\frac{L_{\text{train}}(\vec{w}_n;\vec{\beta})}{L_{\text{train}}(\vec{w}_0;\vec{\beta})} \xrightarrow{\mathbb{P}} \infty. \tag{38}$$

Both losses in (38) use the same full training data. Thus SGD ends with much larger training loss than it started with, even though its test risk tends to zero (Figure 1(a)).

Learning within one pass forces large steps. In the range (8), any constant-rate pass with expected test risk at most a fixed $g \in (0,1)$ has $\eta_n d_0 \to \infty$, so each update multiplies the residual on the example it uses by $1 - \eta_n \|\vec{x}_i\|^2$, which tends to $-\infty$ in probability uniformly over the examples (Appendix E.2). In that range (8), steps from zero that never enlarge the residual on their example cannot learn over any number of epochs (Proposition E.1 in Appendix E).

5 Irrelevant strong directions can block constant-rate SGD

To isolate the effect of signal directions irrelevant to the label, problem families A and B share, at each integer $M \geq 2$, the dimensions, target, and variances of the two directions used by the label:

$$\begin{aligned} &n = M^2, \qquad d_0 = M^7, \qquad p = M+1, \qquad \vec{\beta} = (\vec{e}_1 + \vec{e}_{M+1})/\sqrt{2}, \\ &\mathbf{\Gamma}_A = \operatorname{diag}(M^4, M^3 \boldsymbol{I}_M), \qquad \mathbf{\Gamma}_B = \operatorname{diag}(M^4 \boldsymbol{I}_M, M^3). \end{aligned} \tag{39}$$

The first and last coordinates each contribute half the label variance. Coordinates $2, \ldots, M$ do not contribute to the label, but their variances increase from M^3 in A to M^4 in B. The learning ratios (30) are $R_{\text{wk}} = M$ in both families, with $q_n = 2/M$ in A and $q_n = 1 + M^{-2}$ in B.

Corollary 4 (Matching leading risks and losses do not ensure SGD success)**.** *For both families in* (39)*, the best span predictor $\vec{u}_*$ (the span predictor with smallest test risk) satisfies*

$$\frac{\mathcal{E}(\vec{u}_*;\vec{\beta})}{(M^{-1}+M^{-3})/2} \xrightarrow{\mathbb{P}} 1, \qquad \frac{L_{\text{train}}(\vec{u}_*;\vec{\beta})}{(M^4+M^2)/2} \xrightarrow{\mathbb{P}} 1. \tag{40}$$

For either family, let $\vec{u}_{\text{int}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\vec{y}$ be its minimum-norm interpolator. Even the best scalar multiple, chosen with knowledge of the realized training data and the true distribution, obeys

$$Z_M := \inf_{c\in\mathbb{R}} \mathcal{E}(c\vec{u}_{\text{int}};\vec{\beta}) \xrightarrow{\mathbb{P}} \frac{1}{2}, \qquad \mathbb{E}Z_M \longrightarrow \frac{1}{2}. \tag{41}$$

In A, at the rate $\eta_M = \log M/(2M^5)$, the final SGD iterate has vanishing expected test risk, so by Theorem 1 its training loss diverges in probability. In B, at every deterministic positive constant rate, the final SGD iterate satisfies

$$\mathbb{E}\mathcal{E}(\vec{w}_n;\vec{\beta}) \geq \frac{1}{2}\left(\frac{M^2+1}{M^2+3}\right)^{2M^2}. \tag{42}$$

The lower bound in (42) *tends to $e^{-4}/2 > 0$.*

A's rate satisfies (33), with $n\eta_M\gamma_p = \frac{1}{2}\log M \to \infty$ and $\eta_M \operatorname{tr}\mathbf{\Gamma}_A = (\log M)/M \to 0$. In B, the weakest direction carries half the label variance. The unit-target argument (35)–(36), weighted by this factor as in the proof of (32), therefore requires both $n\eta\gamma_p \to \infty$ and $\eta < 2/(\operatorname{tr}\mathbf{\Gamma}_B + 2\gamma_p)$ for vanishing expected risk. But the latter gives $n\eta\gamma_p < 2M^2/(M^2+3) < 2$. Appendix F proves the corollary. The nuisance block makes accurate span predictors misfit in both families. In both families of (39), guarded negative ridge uses only the training data and the supplied n and d_0, yet has expected test risk $O(M^{-1})$ (Appendix F.2). So the signal trace decides whether a constant-rate SGD pass achieves vanishing expected test risk, not whether another estimator can do so. Accurate misfitting predictors, and one SGD pass that reaches them, do not need the dimensions of (39). With $n = 400$, $d = 40{,}002$, spike strengths 30 and 25, and $\vec{\beta} = (1,1)/\sqrt{2}$, the minimum-norm interpolator has median test risk 0.62, the best span predictor 0.12 at training loss 9 times the zero predictor's, and one SGD pass at the explicit rate (D.8) reaches 0.20 at 4 times (Appendix F.8).

AI USE STATEMENT

We used generative AI tools for tasks that require disclosure and for tasks with recommended disclosure. Under required disclosure: the tools helped formulate the explicit pair of families in Section 5 as a specialization of Theorems 1 and 3 to sharply illustrate the separation that we wanted to show, helped simplify the sufficiency argument for Theorem 3 to the bound (D.2), proposed the scalar-multiple clause (41) of Corollary 4 with its proof and the within-pass computation behind Figure 1(a), assisted in writing the proofs in Appendices A–H from the authors' derivations and longer internal working paper, and implemented the numerical computations, including the explicit simulations of Appendix F.8. The Gaussian samples in Figure 1 were drawn by ordinary pseudo-random routines in code whose implementation the tools assisted; no data values were produced by a language model. We did not use generative AI tools to clean or reformat data or to interpret results beyond the exact calculations reported. Under recommended disclosure: Antigravity was our primary editor; the tools (Claude Code and Codex) drafted and edited parts of the text, produced the figures from fixed numerical archives, summarized and located related literature, and formatted references. We have reviewed all AI-assisted work. The authors checked each theorem statement and each proof line by line against their own derivations, reran the finite numerical checks independently of the tools, and verified every citation against the cited source. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

We also give permission as authors for reviewers to use generative AI tools to check the correctness of the mathematics in this paper and, if they want, to help run their own numerical tests. (Of course, we hope that reviewers will use their human judgement to understand the significance and importance of this work.)

ETHICS STATEMENT

This work is theoretical. It involves no human subjects, personal data, or deployed systems, and we are not aware of ethical concerns specific to it.

REPRODUCIBILITY STATEMENT

The main-text arguments and Appendices A–E together prove the claims in Sections 2–4, including the finite events behind every limiting statement. Appendix F proves Corollary 4 and the guarded-ridge guarantee, and the guide before Appendix A shows where each claim is completed. Appendix F gives the exact numerical protocols for both panels of Figure 1, including the seed, the draw order, the rate grid, and the precision checks, and Appendices F.6 and F.7 describe the computations behind Figures 4 and 5. The supplementary material contains the code that computes the exact expected risks, samples the Gaussian Gram quantities, simulates the passes of Figures 4 and 5, and renders the figures, together with archives of every sampled predictor's risk and losses and of every searched learning rate for Figures 1 and 3. For Figures 4 and 5, the archives record the exact expected risks and the plotted percentiles, together with every run's final test risk and training loss for Figure 5, and the code regenerates every run from its seed. Appendix F.8 gives the protocol for Tables 3 and 4, and the supplement contains their explicit-simulation script and archive.

REFERENCES

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.

Blake Bordelon and Cengiz Pehlevan. Learning curves for SGD on structured features. In *International Conference on Learning Representations*, 2022. arXiv:2106.02713.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-distribution generalization in kernel regression. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Dennis Chemnitz and Maximilian Engel. Characterizing dynamical stability of stochastic gradient descent in overparameterized learning. *Journal of Machine Learning Research*, 26(134):1–46, 2025. URL https://www.jmlr.org/papers/v26/24-1547.html.

Kenneth R. Davidson and Stanisław J. Szarek. Local operator theory, random matrices and Banach spaces. In William B. Johnson and Joram Lindenstrauss (eds.), *Handbook of the Geometry of Banach Spaces*, volume 1, pp. 317–366. North-Holland, 2001.

Morris L. Eaton. *Group Invariance Applications in Statistics*, volume 1 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, 1989.

Jake Freeman. Shrinkage to infinity: Reducing test error by inflating the minimum norm interpolator in linear models. arXiv:2510.19206, 2025. Version 2, revised 30 April 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Tomer Koren, Roi Livni, Yishay Mansour, and Uri Sherman. Benign underfitting of stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19605–19617, 2022.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, 1982.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vítor H. Nascimento and Ali H. Sayed. On the learning mechanism of adaptive filters. *IEEE Transactions on Signal Processing*, 48(6):1609–1625, 2000. doi: 10.1109/78.845919.

Gireeja Ranade and Anant Sahai. Control capacity. *IEEE Transactions on Information Theory*, 65 (1):235–254, 2019. doi: 10.1109/TIT.2018.2868929.

Gireeja Ranade and Anant Sahai. The Fourth Quadrant: A Stylized View of Benign Misfitting. arXiv:2608.01032v1, 2026. URL https://arxiv.org/abs/2608.01032v1.

Ali H. Sayed. *Adaptive Filters*. Wiley–IEEE Press, 2008.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of SGD for least-squares in the interpolation regime. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21581–21591, 2021. arXiv:2102.03183.

Justin Werfel, Xiaohui Xie, and H. Sebastian Seung. Learning curves for stochastic gradient descent in linear feedforward networks. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2004.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Last iterate risk bounds of SGD with decaying stepsize for overparameterized linear regression. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24280–24314. PMLR, 2022a. arXiv:2110.06198.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. The power and limitation of pretraining-finetuning for linear regression under covariate shift. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

Jingfeng Wu, Peter L. Bartlett, Sham M. Kakade, Jason D. Lee, and Bin Yu. Risk comparisons in linear regression: Implicit regularization dominates explicit regularization (extended abstract). In *Proceedings of the Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6849–6851, 2026. Full version: arXiv:2509.17251.

Peng Zhao. Beyond negative-ridge endpoints: Mixed-sign spectral regularization via negative-shifted gradient descent. arXiv:2607.22474v1, July 2026. URL https://arxiv.org/abs/2607.22474. Version 1, submitted 24 July 2026.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Benign overfitting of constant-stepsize SGD for linear regression. *Journal of Machine Learning Research*, 24 (326):1–58, 2023.

GUIDE TO THE PROOFS

The main text derives the deterministic converse bound (13), the reflected-pair risk identity (24), and, from the Gaussian recurrences, the weakest-unit-target case of the signal-trace lower bound (32) and the necessity of $q_n \to 0$. The appendices supply the remaining probability and conditioning arguments, the best-span calculation, the rest of the constant-rate characterization, and the proofs for the paired families. Table 1 lists, for each main-text conclusion, what remains to be proved and where.

Table 1: Where each main-text conclusion is completed.

Main-text conclusion	What remains to be proved	Where
Theorem 1, converse (10)	One input event of explicit probability, valid for all unit targets and accurate span predictors at once, with deterministic errors tending to zero	Appendix A.1
Theorem 1, best-span limits (9)	The minimizing span predictor, then relative bounds on both losses, uniform over targets	Appendices A.2 and A.3
Theorem 2	Reflection preserves each input density and the posterior law, although the reflection depends on the observed inputs	Appendix B
Learning rules outside the span (Section 3)	Training-loss divergence from vanishing average risk, and the transfer within the generating frame's equal-density class	Appendix B.1
Recurrences (27)–(28) and mean residual (29)	Applicability of the cited Gaussian identity, its conversion, and a first-moment calculation	Appendices C.1–C.2 and the start of Appendix D
Theorem 3, floor (32)	Extension from the weakest unit target to every target	Appendix D.1
Theorem 3, equivalence (31)	A finite risk bound with an explicit rate, and necessity of $R_{\mathrm{wk}} \to \infty$	Appendices D.2 and D.3
Joint window (37) and (38)	Every assumption of Theorems 1 and 3, and the risk and converse events on the same data	Appendix D.4
Fitting, one-pass step size, and Proposition E.1	Convergence to the interpolator on fixed data, $\eta_n d_0 \to \infty$ along the window, and the small-step obstruction	Appendix E
Corollary 4	Substitution into Theorems 1 and 3, and the optimization over scalar multiples of the realized interpolator	Appendices F and F.1
Guarded negative ridge (Section 5)	An estimator using only permitted information, with an expected-risk bound that also controls exceptional data sets	Appendix F.2

Appendices F.4 and F.5 give the numerical protocols for Figures 1 and 3, and Appendices F.6 and F.7 give those for Figures 4 and 5. Appendix F.8 reports three finite examples at ordinary scale. No proof uses them. Appendix G relates the results to earlier SGD and interpolation work and evaluates each family's pass under the other's input law, and Appendix H gives an additional bound for rates chosen from earlier examples. Neither is a step in the proofs of Theorems 1–3 or Corollary 4.

A FINITE GEOMETRY AND THE BEST SPAN PREDICTOR

Section 2 derives the deterministic lower bound (13) and uses concentration to obtain the converse (10). Appendix A.1 supplies the finite probability bounds and makes the simultaneous guarantee over targets and predictors explicit. The best-span limits (9) remain to be proved. Appendix A.2 finds the unique minimizing span predictor on fixed training data, and Appendix A.3 proves both limits uniformly over unit targets.

A.1 A FINITE BOUND APPLIES TO EVERY ACCURATE SPAN PREDICTOR

The bound (12) involves $\lambda_{\min}(\boldsymbol{W})$ for the nuisance Gram matrix $\boldsymbol{W}$ and $\|\boldsymbol{S}\|_{\mathrm{op}}$ for the standardized signal matrix $\boldsymbol{S}$. We need a lower bound on the first and an upper bound on the second, holding together with explicit probability. For a failure probability $\delta \in (0,1)$, define

$$h_\delta = \log(2/\delta), \qquad \Lambda_\delta = (\sqrt{d_0} - \sqrt{n} - \sqrt{2h_\delta})_+^2, \qquad s_\delta = \sqrt{n} + \sqrt{p} + \sqrt{2h_\delta}. \tag{A.1}$$

The Gaussian singular-value bounds of Davidson & Szarek (2001, Theorem II.13) imply

$$\mathbb{P}\{\boldsymbol{W} \succeq \Lambda_\delta \boldsymbol{I}_n,\ \|\boldsymbol{S}\|_{\mathrm{op}} \le s_\delta\} \ge 1 - \delta. \tag{A.2}$$

Each of the two failures has probability at most $e^{-h_\delta} = \delta/2$. The event (A.2) uses no labels. Let the rank event be the event that $\boldsymbol{S}$ and $\boldsymbol{N}$ have full column rank. It has probability one, because conditional on the preceding columns, each independent Gaussian column has a density and so lies in their proper span with probability zero. On it, $\boldsymbol{S}\vec{\beta} \neq \vec{0}$ for every unit target, so $L_0(\vec{\beta}) > 0$ in (A.4) below.

On the event (A.2), suppose $\mathcal{E}(\boldsymbol{X}\vec{\alpha}; \vec{\beta}) \le g$ for $g \ge 0$. Any bound $\|\boldsymbol{\Gamma}^{-1}\vec{a}\| \ge b \ge 0$ forces $\|\vec{\alpha}\| \ge b/\|\boldsymbol{S}\|_{\mathrm{op}}$ by (11). Substituting this coefficient bound into (12) gives

$$L_{\mathrm{train}}(\boldsymbol{X}\vec{\alpha}; \vec{\beta}) \ge \frac{1}{n}\left[\frac{\Lambda_\delta}{s_\delta} b - s_\delta\sqrt{g}\right]_+^2, \tag{A.3}$$

$$\frac{L_{\mathrm{train}}(\boldsymbol{X}\vec{\alpha}; \vec{\beta})}{L_0(\vec{\beta})} \ge \left[\frac{\Lambda_\delta}{s_\delta^2} b - \sqrt{g}\right]_+^2. \tag{A.4}$$

For (A.4), divide the square-root form of (A.3) by $\sqrt{L_0(\vec{\beta})} = \|\boldsymbol{S}\vec{\beta}\|/\sqrt{n}$, then use $\|\boldsymbol{S}\vec{\beta}\| \le s_\delta$. The bracket is a positive part and hence nonnegative, so replacing the divisor $\|\boldsymbol{S}\vec{\beta}\|$ by its upper bound s_δ can only decrease the right side. Two valid lower bounds b on $\|\boldsymbol{\Gamma}^{-1}\vec{a}\|$, for the signal response $\vec{a} = \boldsymbol{\Gamma}\boldsymbol{S}^\top\vec{\alpha}$ of (4), are

$$b = \frac{(1-\sqrt{g})_+}{\gamma_1}, \qquad b = \left(B(\vec{\beta}) - \frac{\sqrt{g}}{\gamma_p}\right)_+. \tag{A.5}$$

The first follows from (11). The second follows from the reverse triangle inequality $\|\boldsymbol{\Gamma}^{-1}\vec{a}\| \ge B(\vec{\beta}) - \|\boldsymbol{\Gamma}^{-1}\|_{\mathrm{op}}\|\vec{a} - \vec{\beta}\|$, with $\|\boldsymbol{\Gamma}^{-1}\vec{\beta}\| = B(\vec{\beta})$ from (7), $\|\boldsymbol{\Gamma}^{-1}\|_{\mathrm{op}} = 1/\gamma_p$, and $\|\vec{a} - \vec{\beta}\| \le \sqrt{g}$ from (5). The first bound leads to the converse (10). The second keeps the target's dependence on the signal strengths, which Appendix A.3 needs for the sharp training-loss scale in (9). The event (A.2) and the rank event depend on neither g nor the target. Consequently, on their intersection, (A.3) and (A.4) hold, with either bound, simultaneously for every $g \ge 0$, every unit target, and every span predictor with $\mathcal{E}(\boldsymbol{X}\vec{\alpha}; \vec{\beta}) \le g$.

The first response bound in (A.5) does not depend on the target. For the transfers in Appendix B.1, denote the resulting raw and normalized training-loss bounds by

$$\ell_n(g,\delta) = \frac{1}{n}\left[\frac{\Lambda_\delta}{s_\delta}\frac{(1-\sqrt{g})_+}{\gamma_1} - s_\delta\sqrt{g}\right]_+^2, \tag{A.6}$$

$$\widetilde{\ell}_n(g,\delta) = \left[\frac{\Lambda_\delta}{s_\delta^2}\frac{(1-\sqrt{g})_+}{\gamma_1} - \sqrt{g}\right]_+^2. \tag{A.7}$$

For $n \ge 2$, take $\delta = n^{-2}$. Under (8), $\Lambda_\delta/d_0 \to 1$ and $s_\delta/\sqrt{n} \to 1$. For each fixed $g \in (0,1)$, divide (A.6) and (A.7) by $(1-\sqrt{g})^2(\tau/\gamma_1)^2$. Both deterministic ratios approach one, because $\Lambda_\delta/d_0 \to 1$, $s_\delta/\sqrt{n} \to 1$, and the subtracted terms are of relative order $\gamma_1/\tau \to 0$. Define $1 - \epsilon_n(g)$ as the smaller ratio. This is the deterministic $\epsilon_n(g) \to 0$ of (10), with $\tau = d_0/n$ from (7).

A.2 THE SMALLEST NUISANCE VARIANCE FOR A GIVEN SIGNAL RESPONSE

By (5), the test risk of a span predictor $\boldsymbol{X}\vec{\alpha}$ depends on $\vec{\alpha}$ only through its signal response $\vec{a} = \boldsymbol{\Gamma}\boldsymbol{S}^\top\vec{\alpha}$ and its nuisance variance $\vec{\alpha}^\top\boldsymbol{W}\vec{\alpha}$. We therefore minimize in two stages. First, fix $\vec{a}$ and find the smallest nuisance variance among coefficient vectors with that response. Second, minimize the resulting test risk over $\vec{a}$. On the rank event, define the Gram matrix $\boldsymbol{Q}$ of the columns of $\boldsymbol{S}$ in the inner product given by $\boldsymbol{W}^{-1}$, and the response matrix $\boldsymbol{M}_{\mathrm{resp}}$, by

$$\boldsymbol{Q} = \boldsymbol{S}^\top\boldsymbol{W}^{-1}\boldsymbol{S}, \qquad \boldsymbol{M}_{\mathrm{resp}} = \boldsymbol{\Gamma}^{-1}\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}^{-1}. \tag{A.8}$$

Both matrices in (A.8) are positive definite because $\boldsymbol{W} \succ 0$ and $\boldsymbol{S}$ has full column rank. The matrix $\boldsymbol{M}_{\mathrm{resp}}$ will express the least nuisance variance as a quadratic in the desired response $\vec{a}$. The coefficient vector

$$\vec{\alpha}(\vec{a}) = \boldsymbol{W}^{-1}\boldsymbol{S}\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}^{-1}\vec{a} \tag{A.9}$$

has that response because $\boldsymbol{\Gamma}\boldsymbol{S}^\top\vec{\alpha}(\vec{a}) = \vec{a}$. Any other coefficient vector with response $\vec{a}$ is $\vec{\alpha}(\vec{a}) + \vec{h}$ with $\boldsymbol{S}^\top\vec{h} = \vec{0}$. Its nuisance variance expands as

$$(\vec{\alpha}(\vec{a}) + \vec{h})^\top\boldsymbol{W}(\vec{\alpha}(\vec{a}) + \vec{h}) = \underbrace{\vec{\alpha}(\vec{a})^\top\boldsymbol{W}\vec{\alpha}(\vec{a})}_{=\vec{a}^\top\boldsymbol{M}_{\mathrm{resp}}\vec{a}} + 2\underbrace{\vec{\alpha}(\vec{a})^\top\boldsymbol{W}\vec{h}}_{=(\boldsymbol{\Gamma}^{-1}\vec{a})^\top\boldsymbol{Q}^{-1}\boldsymbol{S}^\top\vec{h}=0} + \vec{h}^\top\boldsymbol{W}\vec{h}, \tag{A.10}$$

where the first underbraced identity uses $\boldsymbol{S}^\top\boldsymbol{W}^{-1}\boldsymbol{S} = \boldsymbol{Q}$ from (A.8) and the second uses $\boldsymbol{S}^\top\vec{h} = \vec{0}$. Because $\boldsymbol{W} \succ 0$, the last term is positive unless $\vec{h} = \vec{0}$, so (A.9) is the unique minimizer. This is the equality-constrained quadratic calculation, also described by Boyd & Vandenberghe (2004, Section 10.1).

After minimizing nuisance variance, the remaining test risk is $\|\vec{a} - \vec{\beta}\|^2 + \vec{a}^\top\boldsymbol{M}_{\mathrm{resp}}\vec{a}$. Setting its gradient to zero gives $(\boldsymbol{I}_p + \boldsymbol{M}_{\mathrm{resp}})\vec{a} = \vec{\beta}$, so

$$\vec{a}_* = (\boldsymbol{I}_p + \boldsymbol{M}_{\mathrm{resp}})^{-1}\vec{\beta}, \qquad \vec{u}_* = \boldsymbol{X}\vec{\alpha}(\vec{a}_*), \tag{A.11}$$

$$\mathcal{E}(\vec{u}_*; \vec{\beta}) = \vec{\beta}^\top\boldsymbol{M}_{\mathrm{resp}}(\boldsymbol{I}_p + \boldsymbol{M}_{\mathrm{resp}})^{-1}\vec{\beta} = \vec{\beta}^\top(\boldsymbol{I}_p + \boldsymbol{\Gamma}\boldsymbol{Q}\boldsymbol{\Gamma})^{-1}\vec{\beta}. \tag{A.12}$$

Substituting $\vec{a}_*$ gives the middle form in (A.12), and the last form uses $\boldsymbol{M}_{\mathrm{resp}}^{-1} = \boldsymbol{\Gamma}\boldsymbol{Q}\boldsymbol{\Gamma}$. The positive-definite response quadratic ensures uniqueness.

A.3 CONCENTRATION GIVES THE BEST-SPAN RISK AND TRAINING LOSS

The leading risk $\tau B(\vec{\beta})^2$ in (9) tends to zero, so an additive $o(1)$ error would not prove the ratios there. We need relative matrix bounds that hold on one event for every target, however different the signal strengths are. For each target $\vec{\beta}$, abbreviate $B = B(\vec{\beta})$ in this subsection. The next event traps the nuisance Gram matrix $\boldsymbol{W}$ and the signal Gram matrix $\boldsymbol{S}^\top\boldsymbol{S}$ between scalar multiples of the identity. For $n \geq 2$, set $z_n = \sqrt{2\log(4n^2)}$, $u_W = (\sqrt{n} + z_n)/\sqrt{d_0}$, and $u_S = (\sqrt{p} + z_n)/\sqrt{n}$. When $u_W, u_S < 1$, the upper and lower singular-value tails of $\boldsymbol{N}$ and of $\boldsymbol{S}$ (Davidson & Szarek, 2001, Theorem II.13), each at most $e^{-z_n^2/2} = 1/(4n^2)$, give one event of probability at least $1 - n^{-2}$ on which

$$(1 - u_W)^2 d_0 \boldsymbol{I}_n \preceq \boldsymbol{W} \preceq (1 + u_W)^2 d_0 \boldsymbol{I}_n, \tag{A.13}$$

$$(1 - u_S)^2 n \boldsymbol{I}_p \preceq \boldsymbol{S}^\top\boldsymbol{S} \preceq (1 + u_S)^2 n \boldsymbol{I}_p. \tag{A.14}$$

The conditions $u_W, u_S < 1$ hold eventually under (8). Since $\boldsymbol{Q} = \boldsymbol{S}^\top\boldsymbol{W}^{-1}\boldsymbol{S}$ by (A.8), (A.13) and (A.14) give $c_-\boldsymbol{I}_p \preceq \tau\boldsymbol{Q} \preceq c_+\boldsymbol{I}_p$, where $c_- = (1-u_S)^2/(1+u_W)^2$ and $c_+ = (1+u_S)^2/(1-u_W)^2$ both tend to one. Set $e_n = \max\{1 - c_-, c_+ - 1\}$ to use one error tolerance for both sides. It is deterministic and tends to zero. Inverting the bounds on $\tau\boldsymbol{Q}$ reverses their order. Multiplying on the left and right by $\boldsymbol{\Gamma}^{-1}$ in $\boldsymbol{M}_{\mathrm{resp}} = \boldsymbol{\Gamma}^{-1}\boldsymbol{Q}^{-1}\boldsymbol{\Gamma}^{-1}$ preserves that order, with no need for $\boldsymbol{Q}$ and $\boldsymbol{\Gamma}$ to commute. Write $R_{\mathrm{wk}} = \gamma_p^2/\tau$ for the weakest-spike ratio of (30), which diverges under (8). For all sufficiently large n, $e_n < 1$ and

$$\frac{\tau}{1 + e_n}\boldsymbol{\Gamma}^{-2} \preceq \boldsymbol{M}_{\mathrm{resp}} \preceq \frac{\tau}{1 - e_n}\boldsymbol{\Gamma}^{-2}, \qquad \|\boldsymbol{M}_{\mathrm{resp}}\|_{\mathrm{op}} \leq \frac{\tau}{(1 - e_n)\gamma_p^2} = \frac{1}{(1 - e_n)R_{\mathrm{wk}}} \to 0. \tag{A.15}$$

By Appendix A.2, $\vec{a}^\top \boldsymbol{M}_{\mathrm{resp}} \vec{a}$ is the least nuisance variance with which a span predictor produces the signal response $\vec{a}$. Thus (A.15) says that a unit response in signal direction j costs about τ/γ_j^2 in nuisance variance, within the factors $(1 \pm e_n)^{-1}$. Producing exactly the target response $\vec{a} = \vec{\beta}$ therefore costs the least nuisance variance $\vec{\beta}^\top \boldsymbol{M}_{\mathrm{resp}} \vec{\beta} = (1 + o(1))\tau B(\vec{\beta})^2$, uniformly over unit targets. The best span predictor accepts some signal error to reduce this nuisance variance, and (A.16) below shows that the resulting test risk is still $(1 + o(1))\tau B(\vec{\beta})^2$, the leading best-span risk in (9).

Test-risk ratio. For each eigenvalue $m > 0$ of $\boldsymbol{M}_{\mathrm{resp}}$, $m/(1 + \|\boldsymbol{M}_{\mathrm{resp}}\|_{\mathrm{op}}) \le m/(1+m) \le m$. Apply this inequality to (A.12), then apply both parts of (A.15):

$$\frac{1}{(1+e_n)\,[1 + 1/((1-e_n)R_{\mathrm{wk}})]} \le \frac{\mathcal{E}(\vec{u}_*; \vec{\beta})}{\tau B^2} \le \frac{1}{1-e_n}. \tag{A.16}$$

Both bounds are deterministic, do not depend on the target, and tend to one. This proves the test-risk part of (9).

Training-loss lower bound. Set $g = \mathcal{E}(\vec{u}_*; \vec{\beta})$. This data-dependent value is allowed because (12) and (A.5) hold deterministically for every $g \ge \mathcal{E}(\boldsymbol{X}\vec{\alpha}; \vec{\beta})$. Use the second response bound $b = B(1 - \sqrt{g}/(\gamma_p B))_+$ of (A.5). Substitute it into (12) for $\vec{u}_*$ and divide by $\sqrt{n}\,\tau B$. The bounds (A.13) and (A.14) then give

$$\frac{\sqrt{L_{\mathrm{train}}(\vec{u}_*; \vec{\beta})}}{\tau B} \ge \Big[\frac{(1-u_W)^2}{1+u_S}\Big(1 - \underbrace{\frac{\sqrt{g}}{\gamma_p B}}_{\text{response correction}}\Big)_+ - (1+u_S)\underbrace{\frac{\sqrt{g}}{\tau B}}_{\text{cancellation correction}}\Big]_+. \tag{A.17}$$

The upper bound in (A.16) gives $g \le \tau B^2/(1-e_n)$, so the two corrections obey

$$\frac{\sqrt{g}}{\gamma_p B} \le \frac{1}{\sqrt{(1-e_n)R_{\mathrm{wk}}}} \to 0, \qquad \frac{\sqrt{g}}{\tau B} \le \frac{1}{\sqrt{(1-e_n)\tau}} \to 0. \tag{A.18}$$

The response correction vanishes because $R_{\mathrm{wk}} \to \infty$, and the cancellation correction because $\gamma_1 \ll \tau$ with $\gamma_1 > 1$ forces $\tau \to \infty$. The lower bound in (A.17) therefore tends to one uniformly over unit targets.

Training-loss upper bound. For the upper bound, (6) and the triangle inequality give $\sqrt{nL_{\mathrm{train}}} \le \sqrt{\lambda_{\max}(\boldsymbol{W})}\sqrt{\vec{\alpha}^\top \boldsymbol{W}\vec{\alpha}} + \|\boldsymbol{S}\|_{\mathrm{op}}\|\vec{a} - \vec{\beta}\|$. Apply Cauchy–Schwarz to these two terms and use (5) to obtain

$$nL_{\mathrm{train}}(\boldsymbol{X}\vec{\alpha}; \vec{\beta}) \le [\lambda_{\max}(\boldsymbol{W}) + \|\boldsymbol{S}\|_{\mathrm{op}}^2]\mathcal{E}(\boldsymbol{X}\vec{\alpha}; \vec{\beta}). \tag{A.19}$$

For the best span predictor, substitute $\vec{\alpha} = \vec{\alpha}(\vec{a}_*)$ in (A.19) and divide by $n\tau^2 B^2$. The factor $[\lambda_{\max}(\boldsymbol{W}) + \|\boldsymbol{S}\|_{\mathrm{op}}^2]/(n\tau)$ tends to one, and so does $\mathcal{E}(\vec{u}_*; \vec{\beta})/(\tau B^2)$. This upper bound and the preceding lower bound prove the training-loss limit in (9). The normalized loss $L_{\mathrm{train}}(\vec{u}_*; \vec{\beta})/L_0(\vec{\beta})$ has the same leading term, because on the event of (A.13)–(A.14) every unit target has $(1-u_S)^2 \le L_0(\vec{\beta}) = \|\boldsymbol{S}\vec{\beta}\|^2/n \le (1+u_S)^2$. With Appendix A.1, this completes the proof of Theorem 1.

B Reflection and transfer beyond the training span

Section 3 proves the pair-risk identity (24). The posterior identity (20) also needs the posterior law to be invariant under the reflection $\boldsymbol{J}_X$, which depends on the observed inputs. We prove this invariance and check that the reflection preserves each equal-density class. Appendix B.1 then proves the two training-loss consequences stated after Theorem 2.

Keep $(\vec{\beta}, \boldsymbol{\Gamma})$ fixed, as in the compatibility definition (16). The covariance in (14) has determinant $\det \boldsymbol{\Gamma}$ and inverse $\boldsymbol{I}_d + \boldsymbol{U}(\boldsymbol{\Gamma}^{-1} - \boldsymbol{I}_p)\boldsymbol{U}^\top$. The Gaussian input density $f_{\boldsymbol{U}}$ at frame $\boldsymbol{U}$, which defines the equal-density classes of Section 3, is therefore

$$f_{\boldsymbol{U}}(\vec{x}) = \frac{\exp\{-\frac{1}{2}\,[\|\vec{x}\|^2 + (\boldsymbol{U}^\top\vec{x})^\top(\boldsymbol{\Gamma}^{-1} - \boldsymbol{I}_p)(\boldsymbol{U}^\top\vec{x})]\}}{(2\pi)^{d/2}\sqrt{\det\boldsymbol{\Gamma}}}. \tag{B.1}$$

Compatibility fixes the scalar $\vec{\beta}^\top \mathbf{\Gamma}^{-1/2} \boldsymbol{U}^\top \vec{x}_i = y_i$, not the entire vector $\boldsymbol{U}^\top \vec{x}_i$ in (B.1). Thus different compatible frames can give an input different densities. Reflection across the training span preserves all of these projected vectors, which is the stronger property used below.

For training data with full-rank $\boldsymbol{X}$, $\boldsymbol{J}_X$ in (19) is symmetric and orthogonal, and fixes every column of $\boldsymbol{X}$. Therefore

$$\boldsymbol{X}^\top \boldsymbol{J}_X \boldsymbol{U} = \boldsymbol{X}^\top \boldsymbol{U}, \qquad \det \mathbf{\Sigma}_{\boldsymbol{J}_X \boldsymbol{U}} = \det \mathbf{\Sigma}_{\boldsymbol{U}}, \qquad \boldsymbol{X}^\top \mathbf{\Sigma}_{\boldsymbol{J}_X \boldsymbol{U}}^{-1} \boldsymbol{X} = \boldsymbol{X}^\top \mathbf{\Sigma}_{\boldsymbol{U}}^{-1} \boldsymbol{X}. \tag{B.2}$$

The first identity in (B.2) preserves the labels, and its row i says $(\boldsymbol{J}_X \boldsymbol{U})^\top \vec{x}_i = \boldsymbol{U}^\top \vec{x}_i$. This preserves the exponent in (B.1), whose denominator is the same for every frame, so $f_{\boldsymbol{J}_X \boldsymbol{U}}(\vec{x}_i) = f_{\boldsymbol{U}}(\vec{x}_i)$ for every i. Conditional on the frame, the training inputs are independent, so their densities multiply to give equal likelihoods for the full input matrix.

Let μ_{Haar} be the rotation-invariant (Haar) law on frames used for the posterior in (17), and write $\vec{y}_{\boldsymbol{U}}(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{U} \mathbf{\Gamma}^{-1/2} \vec{\beta}$ for the labels that frame $\boldsymbol{U}$ assigns. The labels are a deterministic function of the frame and the inputs. For every bounded measurable h, Fubini's theorem puts the frame integral inside. In that inner integral, $\boldsymbol{X}$ is fixed and has full rank almost surely, so $\boldsymbol{J}_X$ is a fixed orthogonal matrix, and substituting $\boldsymbol{U} \mapsto \boldsymbol{J}_X \boldsymbol{U}$ there gives

$$\begin{aligned} \mathbb{E}[h(\boldsymbol{U}, D)] &= \iint h\big(\boldsymbol{U}, (\boldsymbol{X}, \vec{y}_{\boldsymbol{U}}(\boldsymbol{X}))\big) \prod_{i=1}^{n} f_{\boldsymbol{U}}(\vec{x}_i)\, \mu_{\mathrm{Haar}}(d\boldsymbol{U})\, d\boldsymbol{X} \\ &= \iint h\big(\boldsymbol{J}_X \boldsymbol{U}, (\boldsymbol{X}, \vec{y}_{\boldsymbol{U}}(\boldsymbol{X}))\big) \prod_{i=1}^{n} f_{\boldsymbol{U}}(\vec{x}_i)\, \mu_{\mathrm{Haar}}(d\boldsymbol{U})\, d\boldsymbol{X} = \mathbb{E}[h(\boldsymbol{J}_X \boldsymbol{U}, D)]. \end{aligned} \tag{B.3}$$

The substitution preserves μ_{Haar}, which is invariant under every orthogonal matrix, and it preserves the label vector and every density factor by (B.2). Thus $(\boldsymbol{U}, D)$ and $(\boldsymbol{J}_X \boldsymbol{U}, D)$ have the same law, so the posterior law $\Pi_{\vec{\beta}, \mathbf{\Gamma}, D}$ of (17) is invariant under $\boldsymbol{J}_X$ for almost every D. For every such D, integrating (24) against $\Pi_{\vec{\beta}, \mathbf{\Gamma}, D}$ proves (20) for every predictor $\vec{u}$ at once. In particular, a set of frames and its reflected set have equal posterior probability, although the posterior need not be uniform over the compatible set.

For the worst-case criterion, fix an equal-density class $\mathcal{C}$. For $\boldsymbol{U} \in \mathcal{C}$, $\boldsymbol{J}_X \boldsymbol{U}$ is compatible by the first identity in (B.2) and gives each $\vec{x}_i$ the same density, so $\boldsymbol{J}_X \boldsymbol{U} \in \mathcal{C}$. For any $\boldsymbol{U} \in \mathcal{C}$, the worst risk over $\mathcal{C}$ is at least the pair average (24). Taking the supremum over $\boldsymbol{U} \in \mathcal{C}$ proves (21). This argument uses no prior or posterior.

B.1 Transferring the misfitting requirement to learning rules

Let $A(D)$ be a Borel-measurable learning rule returning one linear predictor, possibly using an independent random seed. Define its projection by $A^{\mathrm{sp}}(D) = \boldsymbol{P}_X A(D)$ when $\boldsymbol{X}$ has full column rank, and set it to zero on the null event where $\boldsymbol{X}$ is rank deficient. The projector is continuous on the full-rank domain, so this is a measurable rule. Projection uses only $\boldsymbol{X}$. The original and projected rules give identical predictions on the training inputs, for every target.

Averaged risk forces training-loss divergence. For fixed target coefficients $\vec{\beta}$, let $\mathbb{P}_{\mathrm{Haar}}$ and $\mathbb{E}_{\mathrm{Haar}}$ average over the rotation-invariant frame law (Haar measure), the training data drawn at that frame, and any independent learner randomness. Integrating (20) gives

$$\mathbb{E}_{\mathrm{Haar}} \mathcal{E}_{\boldsymbol{U}}(A(D); \vec{\beta}) = \mathbb{E}_{\mathrm{Haar}} \mathcal{E}_{\boldsymbol{U}}(A^{\mathrm{sp}}(D); \vec{\beta}) + \mathbb{E}_{\mathrm{Haar}} \left[\vec{u}_{\perp, A}^\top \mathbf{\Sigma}_{\boldsymbol{U}} \vec{u}_{\perp, A} \right], \tag{B.4}$$

where $\vec{u}_{\perp, A} = (\boldsymbol{I}_d - \boldsymbol{P}_X) A(D)$ is the off-span component of the returned predictor. The equality in (B.4) is valid for infinite expectations by nonnegativity and Tonelli's theorem.

Suppose the original rule satisfies $\mathbb{E}_{\mathrm{Haar}} \mathcal{E}_{\boldsymbol{U}}(A(D); \vec{\beta}) \le r_n < \infty$. By (B.4), its projected rule has no larger average test risk. Markov's inequality therefore bounds the probability that the projected rule's true test risk exceeds $g > 0$ by r_n / g. The event (A.2) is stated through $\boldsymbol{W}$ and $\boldsymbol{S}$, which are defined in the aligned coordinates. At each fixed frame $\boldsymbol{U} = \boldsymbol{H} \boldsymbol{U}_0$, applying $\boldsymbol{H}^\top$ returns the data to the law (1), so (A.2) still fails with probability at most δ. So with $\mathbb{P}_{\mathrm{Haar}}$-probability at least $1 - \delta - r_n / g$, the event (A.2) holds and the projected rule has true test risk $\mathcal{E}_{\boldsymbol{U}}(A^{\mathrm{sp}}(D); \vec{\beta}) \le g$. On

this intersection, the projected rule's training loss is at least $\ell_n(g,\delta)$ of (A.6), and its ratio to $L_0(\vec{\beta})$ is at least $\widetilde{\ell}_n(g,\delta)$ of (A.7). Projection preserves training loss, so

$$\mathbb{P}_{\mathrm{Haar}}\{L_{\mathrm{train}}(A(D);\vec{\beta}) < \ell_n(g,\delta)\} \le \delta + r_n/g, \tag{B.5}$$

$$\mathbb{P}_{\mathrm{Haar}}\{L_{\mathrm{train}}(A(D);\vec{\beta})/L_0(\vec{\beta}) < \widetilde{\ell}_n(g,\delta)\} \le \delta + r_n/g. \tag{B.6}$$

Choose any fixed $g \in (0,1)$ and $\delta = n^{-2}$ for $n \ge 2$. Under (8), Appendix A.1 shows that $\ell_n(g,n^{-2})$ and $\widetilde{\ell}_n(g,n^{-2})$ are $(1+o(1))(1-\sqrt{g})^2(\tau/\gamma_1)^2$, which diverges because $\gamma_1 \ll \tau$. If $r_n \to 0$, the right sides of (B.5) and (B.6) vanish, which proves the training-loss divergence stated after Theorem 2.

Worst-case risk in the generating class forces the converse. The worst-case transfer needs no averaging. Let $\mathcal{C}_{\mathrm{gen}}$ be the equal-density class containing the generating frame $\boldsymbol{H}\boldsymbol{U}_0$. If $\mathcal{R}^{\mathrm{wc}}_{\vec{\beta},\boldsymbol{\Gamma},D}(\vec{u};\mathcal{C}_{\mathrm{gen}}) \le g$ for a fixed $g \in (0,1)$, then (21) for $\mathcal{C}_{\mathrm{gen}}$ gives $\mathcal{R}^{\mathrm{wc}}_{\vec{\beta},\boldsymbol{\Gamma},D}(\vec{u}_\parallel;\mathcal{C}_{\mathrm{gen}}) \le g$ for $\vec{u}_\parallel = \boldsymbol{P}_X\vec{u}$. The generating frame belongs to $\mathcal{C}_{\mathrm{gen}}$, so $\vec{u}_\parallel$ has test risk at most g there. After rotating to the aligned coordinates, on the event of Theorem 1 its training loss obeys (10). The original predictor $\vec{u}$ has exactly the same training loss. This transfer uses membership of the generating frame, and the event of Theorem 1 concerns that frame's own signal and nuisance blocks, not those of every compatible frame. This is a conditional worst-case conclusion, not an ex ante minimax claim.

Equivariant rules. Call a rule orthogonally equivariant when rotating every training input rotates its returned predictor in the same way, $A(\boldsymbol{H}\boldsymbol{X},\vec{y}) = \boldsymbol{H}A(\boldsymbol{X},\vec{y})$ for every orthogonal $\boldsymbol{H}$. For such a rule, the joint law of training loss and test risk is the same at every frame $\boldsymbol{H}\boldsymbol{U}_0$. So (B.5) and (B.6) also hold at every fixed frame. We do not make that fixed-frame conclusion for an arbitrary non-equivariant rule.

Which sets of frames does the worst-case comparison cover? The argument used only that the set of frames is nonempty and closed under $\boldsymbol{U} \mapsto \boldsymbol{J}_X\boldsymbol{U}$. The bound (21) therefore holds with the supremum taken over any such set of compatible frames. Examples are the compatible frames whose joint input likelihood $f_{\widetilde{\boldsymbol{U}}}(\boldsymbol{X}) = \prod_{i=1}^n f_{\widetilde{\boldsymbol{U}}}(\vec{x}_i)$ takes a given value, lies in a band $[\kappa_1,\kappa_2]$, or is at least κ_1. Filtering frames by input likelihood never removes the reflected alternative. The entire compatible set (16) is also reflection-closed, so the bound holds for the worst case over all compatible frames. The transfer of the converse (10) holds for any such set that contains the generating frame, such as a likelihood band containing the generating frame's likelihood. Both comparisons in Theorem 2 also hold under an upper bound on predictor norm, because projection does not increase the norm: $\|\vec{u}_\parallel\| \le \|\vec{u}\|$.

C The cited Gaussian SGD result in this paper's notation

The only identity imported from Bordelon & Pehlevan (2022) is their coefficient-error recurrence, their equation (6), with mini-batch size $m = 1$. Appendix C.1 quotes its source and checks its hypotheses, and Appendix C.2 converts it into the recurrences (27) and (28). A reader who accepts the cited identity can go directly to Appendix C.2.

C.1 The imported identity and its hypotheses

The quoted result. The passages below are quoted verbatim from Sections 3 and 4.1 of Bordelon & Pehlevan (2022), arXiv:2106.02713v5 (CC BY 4.0). Their equation numbers carry the prefix BP, so their equation (3) appears as (BP3). Appendix numbers inside the quotations are theirs, and [. . .] marks an omitted passage. Their Section 3 defines the SGD update and the test loss.

> At each time step t, the weights are updated by taking a stochastic gradient step on a fresh mini-batch of m examples
>
> $$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \frac{\eta}{m}\sum_{\mu=1}^{m}\boldsymbol{\psi}_{t,\mu}\left(\boldsymbol{w}_t\cdot\boldsymbol{\psi}_{t,\mu} - y_{t,\mu}\right), \tag{BP3}$$

where each of the vectors $\boldsymbol{\psi}_{t,\mu}$ are sampled independently and $y_{t,\mu} = \boldsymbol{w}^* \cdot \boldsymbol{\psi}_{t,\mu}$. [...] At each timestep, the test-loss, or generalization error, has the form

$$L_t = \left\langle \left(\boldsymbol{w}_t \cdot \boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{w}^* \cdot \boldsymbol{\psi}(\boldsymbol{x}) - y_\perp(\boldsymbol{x})\right)^2 \right\rangle_{\boldsymbol{x}}$$
$$= (\boldsymbol{w}_t - \boldsymbol{w}^*)^\top \boldsymbol{\Sigma} (\boldsymbol{w}_t - \boldsymbol{w}^*) + \left\langle y_\perp(\boldsymbol{x})^2 \right\rangle_{\boldsymbol{x}}, \tag{BP4}$$

which quantifies exactly the test error of the vector $\boldsymbol{w}_t$.

Their Section 4.1 states Theorem 1 and sketches its proof.

Theorem 1. *Suppose the features $\boldsymbol{\psi}$ follow a Gaussian distribution $\boldsymbol{\psi} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and the target function is learnable in these features $y = \boldsymbol{w}^* \cdot \boldsymbol{\psi}$. After t steps of SGD with minibatch size m and learning rate η, the expected (over possible sample sequences $\mathcal{D}_t$) test loss $\langle L_t \rangle_{\mathcal{D}_t}$ has the form*

$$\langle L_t \rangle_{\mathcal{D}_t} = \boldsymbol{\lambda}^\top \boldsymbol{A}^t \boldsymbol{v}^2, \quad \boldsymbol{A} = (\boldsymbol{I} - \eta \operatorname{diag}(\boldsymbol{\lambda}))^2 + \frac{\eta^2}{m} \operatorname{diag}\left(\boldsymbol{\lambda}^2\right) + \frac{\eta^2}{m} \boldsymbol{\lambda}\boldsymbol{\lambda}^\top \tag{BP5}$$

where $\boldsymbol{\lambda}$ is a vector containing the eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{v}^2$ is a vector containing elements $(\boldsymbol{v}^2)_k = v_k^2 = (\boldsymbol{u}_k \cdot \boldsymbol{w}^*)^2$ for eigenvectors $\boldsymbol{u}_k$ of $\boldsymbol{\Sigma}$. The function $\operatorname{diag}(\cdot)$ constructs a diagonal matrix with the argument vector placed along the diagonal.

Proof. See Appendix B for the full derivation. We will provide a brief sketch of the proof here. The strategy of the proof relies on the fact that $\langle L_t \rangle = \operatorname{Tr} \boldsymbol{\Sigma}\, \boldsymbol{C}_t$ where $\boldsymbol{C}_t = \left\langle (\boldsymbol{w}_t - \boldsymbol{w}^*)(\boldsymbol{w}_t - \boldsymbol{w}^*)^\top \right\rangle_{\mathcal{D}_t}$. We derive the following recursion relation for this error matrix

$$\boldsymbol{C}_{t+1} = (\boldsymbol{I} - \eta\boldsymbol{\Sigma})\boldsymbol{C}_t(\boldsymbol{I} - \eta\boldsymbol{\Sigma}) + \frac{\eta^2}{m}\left[\boldsymbol{\Sigma}\boldsymbol{C}_t\boldsymbol{\Sigma} + \boldsymbol{\Sigma} \operatorname{Tr}\left(\boldsymbol{\Sigma}\boldsymbol{C}_t\right)\right] \tag{BP6}$$

The loss only depends on $c_{k,t} = \boldsymbol{u}_k^\top \boldsymbol{C}_t \boldsymbol{u}_k$. Solving the recurrence, $\boldsymbol{c}_t = \boldsymbol{A}^t \boldsymbol{v}^2$ and using $\langle L_t \rangle = \sum_k \lambda_k \boldsymbol{u}_k^\top \boldsymbol{C}_t \boldsymbol{u}_k = \sum_k c_{k,t}\lambda_k = \boldsymbol{\lambda}^\top \boldsymbol{A}^t \boldsymbol{v}^2$, we obtain the desired result.

Their Appendix B proves (BP6) in expanded form, as their equation (24), using the Wick–Isserlis theorem for Gaussian fourth moments.

Dictionary. Table 2 translates their symbols. Here $\vec{w}_\star = \vec{u}_{\boldsymbol{U}_0,\vec{\beta}} = \vec{u}_{\vec{\beta}}$ is the zero-risk predictor of (14) in the aligned coordinates, the true predictor of Section 2. Under the rank assumptions of Section 2, it differs from the best span predictor $\vec{u}_*$, whose risk is positive. Their v_k and N are unrelated to our $\vec{v}_t$ and N_t.

Table 2: Symbols of Bordelon & Pehlevan (2022) and their values here, for the pass (25).

Bordelon–Pehlevan	This paper	Reason
features $\boldsymbol{\psi}(\boldsymbol{x})$, dimension N	input $\vec{x}$, dimension d	the feature map is the identity
$\boldsymbol{\Sigma}$	$\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\Gamma}, \boldsymbol{I}_{d_0})$	covariance of $\vec{x}$ in (1)
$\boldsymbol{w}^*$ and $y_\perp$	$\vec{w}_\star = (\boldsymbol{\Gamma}^{-1/2}\vec{\beta}, \vec{0})$ and 0	$y = \vec{\beta}^\top \vec{s} = \vec{w}_\star^\top \vec{x}$
m and η	1 and the constant rate η	one pair per update
$(\boldsymbol{\psi}_{t,1}, y_{t,1})$	$(\vec{x}_{t+1}, y_{t+1})$	(BP3) becomes (25)
$\langle\cdot\rangle_{\mathcal{D}_t}$	$\mathbb{E}$ over the first t pairs	
L_t	$\mathcal{E}(\vec{w}_t; \vec{\beta})$	(BP4) with $y_\perp = 0$ and (2)
$\boldsymbol{C}_t$	$\boldsymbol{\Omega}_t = \mathbb{E}[(\vec{w}_t - \vec{w}_\star)(\vec{w}_t - \vec{w}_\star)^\top]$	$\vec{w}_0 = \vec{0}$ gives $\boldsymbol{\Omega}_0 = \vec{w}_\star \vec{w}_\star^\top$
$\boldsymbol{\lambda}$	$(\gamma_1, \dots, \gamma_p, 1, \dots, 1)$	d_0 unit nuisance eigenvalues
$\boldsymbol{u}_k$	$\vec{e}_k$	$\boldsymbol{\Sigma}$ is diagonal
v_k^2	β_k^2/γ_k if $k \le p$, and 0 if $k > p$	$v_k = \vec{e}_k^\top \vec{w}_\star$
$c_{k,t}$	$(\boldsymbol{\Omega}_t)_{kk}$	$\vec{e}_k^\top \boldsymbol{\Omega}_t \vec{e}_k$

The hypotheses hold. Under (1), the two blocks of $\vec{x}$ are independent, so $\vec{x} \sim \mathcal{N}(\vec{0}, \mathbf{\Sigma})$. The label is $y = \vec{w}_\star^\top \vec{x}$, so the target is learnable and $y_\perp = 0$. The pass (25) uses each independent pair once, so the pair used by update $t+1$ is independent of $\vec{w}_t$, which depends only on the first t pairs. This is (BP3) with a fresh mini-batch of size $m = 1$, at a constant rate. Hence (BP6) holds with $m = 1$ for $0 \le t < n$. It is an exact identity at every finite t, d and η, not a convergence guarantee, and applying it requires no small-rate or stability condition. This is why it covers the large one-pass rates used here. All moments involved are finite because $\vec{w}_t$ is a polynomial in finitely many Gaussian coordinates.

C.2 Deriving the signal and nuisance recurrences

Let $\mathbf{\Omega}_t = \mathbb{E}[(\vec{w}_t - \vec{w}_\star)(\vec{w}_t - \vec{w}_\star)^\top]$ be the second-moment matrix of the coefficient error, where $\vec{w}_\star = (\mathbf{\Gamma}^{-1/2}\vec{\beta}, \vec{0})$ is the zero-risk predictor and $\mathbf{\Sigma} = \operatorname{diag}(\mathbf{\Gamma}, \boldsymbol{I}_{d_0})$ is the input covariance in aligned coordinates (Table 2). With $m = 1$ and the dictionary, (BP6) becomes

$$\mathbf{\Omega}_{t+1} = (\boldsymbol{I} - \eta\mathbf{\Sigma})\mathbf{\Omega}_t(\boldsymbol{I} - \eta\mathbf{\Sigma}) + \eta^2\left[\mathbf{\Sigma}\mathbf{\Omega}_t\mathbf{\Sigma} + \mathbf{\Sigma}\operatorname{tr}(\mathbf{\Sigma}\mathbf{\Omega}_t)\right], \qquad \mathbf{\Omega}_0 = \vec{w}_\star\vec{w}_\star^\top. \tag{C.1}$$

Since $\mathbf{\Sigma}$ is diagonal with entries λ_k, the (k,k) entry of (C.1) is

$$\begin{aligned}(\mathbf{\Omega}_{t+1})_{kk} &= (1-\eta\lambda_k)^2(\mathbf{\Omega}_t)_{kk} + \eta^2\left[\lambda_k^2(\mathbf{\Omega}_t)_{kk} + \lambda_k\operatorname{tr}(\mathbf{\Sigma}\mathbf{\Omega}_t)\right] \\ &= (1 - 2\eta\lambda_k + 2\eta^2\lambda_k^2)(\mathbf{\Omega}_t)_{kk} + \eta^2\lambda_k\operatorname{tr}(\mathbf{\Sigma}\mathbf{\Omega}_t),\end{aligned} \tag{C.2}$$

where $\operatorname{tr}(\mathbf{\Sigma}\mathbf{\Omega}_t) = \sum_i \lambda_i(\mathbf{\Omega}_t)_{ii}$. Because $\mathbf{\Sigma}$ is diagonal, (C.2) involves only the diagonal entries of $\mathbf{\Omega}_t$.

The diagonal entries of $\mathbf{\Omega}_t$ are expected squared coefficient errors. Multiplying a signal entry by its input variance γ_j gives that coordinate's expected contribution to test risk, and summing the nuisance entries, whose input variances are one, gives N_t. The signal residual is $\vec{r}_t = \vec{\beta} - \mathbf{\Gamma}^{1/2}\vec{w}_t^{\,\mathrm{sig}} = \mathbf{\Gamma}^{1/2}(\vec{w}_\star - \vec{w}_t)^{\mathrm{sig}}$, and the nuisance part is $\vec{v}_t = \vec{w}_t^{\,\mathrm{nui}} = -(\vec{w}_\star - \vec{w}_t)^{\mathrm{nui}}$. Hence (26) gives

$$T_{t,j} = \gamma_j(\mathbf{\Omega}_t)_{jj}, \qquad N_t = \sum_{k=p+1}^{d}(\mathbf{\Omega}_t)_{kk}, \qquad \operatorname{tr}(\mathbf{\Sigma}\mathbf{\Omega}_t) = T_t + N_t = \mathbb{E}\mathcal{E}(\vec{w}_t;\vec{\beta}). \tag{C.3}$$

For a signal coordinate $j \le p$, $\lambda_j = \gamma_j$. Multiplying (C.2) at $k = j$ by γ_j gives $T_{t+1,j} = \rho_j T_{t,j} + \eta^2\gamma_j^2(T_t + N_t)$. Here $\rho_j = 1 - 2\eta\gamma_j + 2\eta^2\gamma_j^2 = (1-\eta\gamma_j)^2 + \eta^2\gamma_j^2$ is the signal multiplier defined before (27), and $T_{0,j} = \gamma_j(\vec{w}_\star)_j^2 = \beta_j^2$. This is (27). For the d_0 nuisance coordinates, $\lambda_k = 1$. Summing (C.2) over them gives

$$N_{t+1} = (1 - 2\eta + 2\eta^2)N_t + \eta^2 d_0(T_t + N_t) = \mu N_t + \eta^2 d_0 T_t,$$

where $\mu = 1 - 2\eta + \eta^2(d_0 + 2) = (1-\eta)^2 + \eta^2(d_0+1)$ is the nuisance multiplier defined before (27), and $N_0 = \|\vec{w}_\star^{\,\mathrm{nui}}\|^2 = 0$. This is (28).

As a check, because T_t contains $T_{t,j}$, the complete coefficient of $T_{t,j}$ in (27) is $\rho_j + \eta^2\gamma_j^2 = 1 - 2\eta\gamma_j + 3\eta^2\gamma_j^2 = \mathbb{E}(1 - \eta\gamma_j s_j^2)^2$, the scalar Gaussian LMS factor with fourth moment $\mathbb{E}s_j^4 = 3$.

At a fixed rate, (27)–(28) are linear and start from $T_{0,j} = \beta_j^2$ and $N_0 = 0$, so the expected risk is linear in $\beta_1^2, \ldots, \beta_p^2$, as used in Appendix D.1. The off-diagonal entries of $\mathbf{\Omega}_0$, which carry the products $\beta_i\beta_j$ with $i \ne j$, never enter these recurrences. This decomposition holds for expected risk and need not hold for the realized risk on one data set.

As a check against the source, (C.2) is their equation (25) at $m = 1$. Stacked over k, it reads $\boldsymbol{c}_{t+1} = \boldsymbol{A}\boldsymbol{c}_t$ with $\boldsymbol{A}$ from (BP5) at $m = 1$, which is how their Theorem 1 follows.

D Completing the constant-rate characterization

Section 4 uses the recurrences (27)–(28), derived in Appendix C, and the mean residual (29), derived below. Given these, Section 4 proves the lower bound (32) for the weakest unit target and the necessity of $q_n \to 0$ in (31). Appendix D.1 extends (32) to every target. We complete (31) in two steps. Appendix D.2 proves a finite risk bound and constructs a rate when $R_{\mathrm{wk}} \to \infty$ and $q_n \to 0$. Appendix D.3 proves that consistency requires $R_{\mathrm{wk}} \to \infty$. Neither step assumes the geometric window (8). Appendix D.4 then verifies the joint window and derives the training-loss conclusion (38).

The mean residual (29) **needs only first moments.** Let $\mathcal{F}_t$ contain the first t training pairs. With the signal residual $\vec{r}_t = \vec{\beta} - \mathbf{\Gamma}^{1/2}\vec{w}_t^{\,\mathrm{sig}}$ and nuisance part $\vec{v}_t = \vec{w}_t^{\,\mathrm{nui}}$ of Section 4, write the next input as $\vec{x}_{t+1} = (\mathbf{\Gamma}^{1/2}\vec{s}, \vec{n})$. Its residual is $\vec{r}_t^{\,\top}\vec{s} - \vec{v}_t^{\,\top}\vec{n}$, so the signal block of (25) gives

$$\vec{r}_{t+1} = \vec{r}_t - \eta\mathbf{\Gamma}\vec{s}(\vec{s}^{\,\top}\vec{r}_t - \vec{n}^{\,\top}\vec{v}_t).$$

The next pair is independent of $\mathcal{F}_t$, with $\mathbb{E}\vec{s}\vec{s}^{\,\top} = \boldsymbol{I}_p$ and $\mathbb{E}\vec{s}\vec{n}^{\,\top} = \mathbf{0}$, so

$$\mathbb{E}[\vec{r}_{t+1} \mid \mathcal{F}_t] = (\boldsymbol{I}_p - \eta\mathbf{\Gamma})\vec{r}_t.$$

Starting at $\vec{r}_0 = \vec{\beta}$ proves (29).

D.1 The trace floor for every target

Section 4 proves the floor (32) for the unit target $\vec{e}_p$ in the weakest direction, where $H_p = \operatorname{tr}\mathbf{\Gamma}$. We extend it to every unit target at the same fixed rate η. For the unit target $\vec{e}_j$, consider the smaller Gaussian problem that retains the signal coordinates with $\gamma_i \geq \gamma_j$ and the nuisance block. Both problems start from $T_{0,j} = 1$, with every other $T_{0,i}$ and N_0 zero. Removing coordinates drops only nonnegative terms from the sums T_t and $T_t + N_t$ in (27)–(28), and every multiplier there is nonnegative. So, by induction at the same rate, each retained $T_{t,i}$ and N_t is no larger in the smaller problem than in the original. The smaller problem has weakest strength γ_j and signal trace H_j, so the argument of Section 4 at its weakest direction gives it expected risk at least $[H_j/(H_j + 2\gamma_j)]^{2n}$. The original problem's expected risk for $\vec{e}_j$ is at least as large.

At this fixed rate, (27)–(28) are linear in the initial vector $(\beta_1^2, \ldots, \beta_p^2, 0)$, which is the sum of the unit-target initial vectors weighted by β_j^2. Thus

$$\mathbb{E}\mathcal{E}(\vec{w}_n^{(\eta)}(\vec{\beta}); \vec{\beta}) = \sum_{j=1}^{p} \beta_j^2\, \mathbb{E}\mathcal{E}(\vec{w}_n^{(\eta)}(\vec{e}_j); \vec{e}_j), \tag{D.1}$$

and the unit-target bounds prove (32). Appendix C.2 explains why no products $\beta_i\beta_j$ with $i \neq j$ enter. Identity (D.1) holds for expected risk and need not hold for the realized risk on one data set.

D.2 Bounding accumulated error during the pass

The following finite bound proves the sufficiency of (33).

Lemma D.1 (A finite expected-risk bound for one pass)**.** *Let η be a deterministic constant rate with $0 < \eta\gamma_1 \leq 1/2$. Write $\phi = \eta \operatorname{tr}\mathbf{\Gamma}$, $C_n = \exp\{n\eta^2(d_0 + 2)\}$, and $\psi_n = C_n n\eta^2 d_0$. If $\phi(1 + \psi_n) \leq 1/2$, then for every predetermined unit target $\vec{\beta}$ the final iterate of* (25) *satisfies*

$$\mathbb{E}\mathcal{E}(\vec{w}_n; \vec{\beta}) \leq e^{-n\eta\gamma_p} + 2\phi(1 + \psi_n) + 2\psi_n. \tag{D.2}$$

The three terms in (D.2) bound, respectively, the initial signal error not yet removed, the signal error created during the pass, and the nuisance error.

Proof. The nuisance recurrence (28) is driven by the expected signal error T_t, and the signal recurrence (27) by the expected total error $T_t + N_t$. We first bound the largest nuisance error during the pass by the largest signal error, then substitute that bound into the signal recurrence to bound both. Write $T_{\max} = \max_{0\leq t\leq n} T_t$ and $N_{\max} = \max_{0\leq t\leq n} N_t$. Expanding (28) and (27) over earlier updates gives, for $1 \leq t \leq n$,

$$N_t = \eta^2 d_0 \sum_{i=0}^{t-1} \mu^{t-1-i} T_i, \qquad T_{t,j} = \rho_j^t \beta_j^2 + \eta^2\gamma_j^2 \sum_{i=0}^{t-1} \rho_j^{t-1-i}(T_i + N_i). \tag{D.3}$$

The nuisance multiplier $\mu = (1 - \eta)^2 + \eta^2(d_0 + 1)$ of (28) is positive and at most $1 + \eta^2(d_0 + 2)$, so $1 + x \leq e^x$ gives $0 < \mu^r \leq C_n$ for $0 \leq r \leq n$. The nuisance sum uses only these first n powers of μ, so this argument also permits $\mu > 1$. The signal weights ρ_j^{t-1-i} form a geometric sequence with ratio ρ_j, the signal multiplier of (27). Because $0 < \eta\gamma_j \leq 1/2$,

$$1 - \rho_j = 2\eta\gamma_j(1 - \eta\gamma_j) > 0, \qquad \frac{\eta^2\gamma_j^2}{1 - \rho_j} = \frac{\eta\gamma_j}{2(1 - \eta\gamma_j)} \leq \eta\gamma_j. \tag{D.4}$$

Bounding each source by its maximum, using $\mu^r \le C_n$ in the nuisance sum and (D.4) in the signal sum, gives

$$N_t \le C_n t\eta^2 d_0 T_{\max} \le \psi_n T_{\max}, \tag{D.5}$$

$$T_{t,j} \le \rho_j^t \beta_j^2 + \eta\gamma_j (T_{\max} + N_{\max}). \tag{D.6}$$

Take the maximum over t in (D.5), which also holds at $t = 0$ because $N_0 = 0$, and sum (D.6) over j using $\rho_j^t \le 1$ and $\sum_j \beta_j^2 = 1$ before maximizing:

$$N_{\max} \le \psi_n T_{\max}, \qquad T_{\max} \le 1 + \phi(T_{\max} + N_{\max}). \tag{D.7}$$

Substituting the first bound in (D.7) into the second gives $[1 - \phi(1 + \psi_n)]T_{\max} \le 1$. Since $\phi(1 + \psi_n) \le 1/2$, this gives $T_{\max} \le 2$ and $N_{\max} \le 2\psi_n$. Because $0 < \eta\gamma_j \le 1/2$, $\rho_j \le 1 - \eta\gamma_j \le e^{-\eta\gamma_p}$, so summing (D.6) over j at $t = n$ gives $T_n \le e^{-n\eta\gamma_p} + \phi(T_{\max} + N_{\max}) \le e^{-n\eta\gamma_p} + 2\phi(1 + \psi_n)$. Adding $N_n \le 2\psi_n$ and using $\mathbb{E}\mathcal{E}(\vec{w}_n; \vec{\beta}) = T_n + N_n$ from (26) proves (D.2). □

Under (33), $\phi \to 0$, $C_n \to 1$, and $\psi_n \to 0$. Thus $\eta\gamma_1 \le 1/2$ and $\phi(1 + \psi_n) \le 1/2$ hold eventually, and every term in (D.2) vanishes. The bound is uniform over predetermined unit targets because it uses only $\sum_j \beta_j^2 = 1$.

For an explicit rate, suppose $R_{\text{wk}} \to \infty$ and $q_n \to 0$. Combine the two vanishing ratios in $\xi_n = R_{\text{wk}}^{-1} + q_n$, and let the progress $n\eta_n\gamma_p = \vartheta_n$ diverge slowly enough that the other two conditions in (33) still vanish. Choose

$$\xi_n = R_{\text{wk}}^{-1} + q_n, \qquad \vartheta_n = \tfrac{1}{2}\log(1/\xi_n), \qquad \eta_n = \frac{\vartheta_n}{n\gamma_p} \quad (0 < \xi_n < 1). \tag{D.8}$$

Then

$$n\eta_n\gamma_p = \vartheta_n \to \infty,$$
$$n\eta_n^2(d_0 + 2) \le 3\vartheta_n^2\xi_n = 3\vartheta_n^2 e^{-2\vartheta_n} \to 0,$$
$$\eta_n \operatorname{tr}\mathbf{\Gamma} = \vartheta_n q_n \le \vartheta_n\xi_n = \vartheta_n e^{-2\vartheta_n} \to 0,$$

so (33) holds. The factor 3 uses $d_0 + 2 \le 3d_0$. At the finitely many indices where $\xi_n \ge 1$, any positive rate may be used, since only the limits matter.

Rates satisfying (33) make $\eta\gamma_1$ small, even though $\eta\|\vec{x}_i\|^2$ can be large. For example, the rate $\eta_M = \log M/(2M^5)$ used for family A in Corollary 4 satisfies $\eta_M\gamma_1 = (\log M)/(2M) \to 0$ but $\eta_M d_0 = \frac{1}{2}M^2 \log M \to \infty$. The initial-error term $\rho_j^t\beta_j^2$ in (D.3) therefore decays, and Lemma D.1 separately bounds the error created during the pass. Meanwhile, with probability tending to one, $\eta_M\|\vec{x}_i\|^2 > 2$ for every training input, so an update on a pair with nonzero residual reverses and enlarges that residual (Appendix E.2).

D.3 Consistency also requires the weakest-spike ratio to diverge

Learning the weakest direction requires $n\eta\gamma_p \to \infty$. If the weakest-spike ratio $R_{\text{wk}} = n\gamma_p^2/d_0$ of (30) stayed bounded along a subsequence, every rate with $n\eta\gamma_p \to \infty$ would make the expected nuisance variance N_n diverge. Use the target $\vec{\beta} = \vec{e}_p$. Suppose its expected risk tends to zero. When $\eta \ge 2/(\operatorname{tr}\mathbf{\Gamma} + 2\gamma_p)$, (35) makes the expected squared error $V_t = \sum_j T_{t,j}/\gamma_j$ in the original signal coefficients nondecreasing, so $\mathbb{E}\mathcal{E}(\vec{w}_n; \vec{e}_p) \ge T_n \ge \gamma_p V_n \ge \gamma_p V_0 = 1$, as in the proof of (32). Consistency therefore requires, eventually,

$$0 < a := \eta\gamma_p < \frac{2\gamma_p}{\operatorname{tr}\mathbf{\Gamma} + 2\gamma_p} \le \frac{2}{3}. \tag{D.9}$$

On the range (D.9), $-\log(1 - a) \le 3a$. Jensen and (29) yield

$$\mathbb{E}\mathcal{E}(\vec{w}_n; \vec{e}_p) \ge (1 - a)^{2n} \ge e^{-6na}. \tag{D.10}$$

Thus consistency forces $na = n\eta\gamma_p \to \infty$.

If R_{wk} did not diverge, choose a subsequence with $R_{\text{wk}} \le C < \infty$. On this subsequence, $\eta d_0 = na\,\gamma_p/R_{\text{wk}} \ge na/C \to \infty$, using $\gamma_p > 1$. Since $\mu - 1 = \eta[\eta(d_0 + 2) - 2]$, we have $\mu \ge 1$

eventually. Thus nuisance error created earlier in the pass is not damped before the pass ends. Use the nuisance sum in (D.3), drop the factors $\mu^{n-1-i} \geq 1$, use $T_i \geq T_{i,p} \geq (1-a)^{2i}$ from Jensen and (29), and sum the finite geometric series:

$$N_n \geq d_0 \eta^2 \sum_{i=0}^{n-1} (1-a)^{2i} = \frac{na}{R_{\mathrm{wk}}(2-a)} [1-(1-a)^{2n}] \tag{D.11}$$

$$\geq \frac{na}{2C} [1-(1-a)^{2n}] \longrightarrow \infty. \tag{D.12}$$

The last limit uses $(1-a)^{2n} \leq e^{-2na} \to 0$. It contradicts $N_n \leq \mathbb{E}\mathcal{E}(\vec{w}_n; \vec{e}_p) \to 0$. This proves necessity of $R_{\mathrm{wk}} \to \infty$.

Together with the necessity of $q_n \to 0$ proved in Section 4, this proves the necessary direction of (31). Lemma D.1 and the explicit rate (D.8) prove the sufficient direction. Section 4 and Appendix D.1 prove (32), which completes the proof of Theorem 3.

Rate requirements at every signal strength. If the eigenspace of strength γ_j carries a fraction of label variance bounded away from zero, vanishing expected test risk requires $n\eta_n\gamma_j \to \infty$ and $\eta_n < 2/(H_j + 2\gamma_j)$ eventually. Both follow by applying the large-rate step (35) and the Jensen step (36) to the smaller problem of Appendix D.1 that retains the signal coordinates with $\gamma_i \geq \gamma_j$, whose weakest strength is γ_j and whose signal trace is H_j. A smaller rate reduces the error created in the stronger directions only by making less progress in the weaker direction during the n updates.

D.4 The joint window supplies every geometric assumption

The left side of (37) gives $R_{\mathrm{wk}} \to \infty$ and $q_n \to 0$. Since $q_n \geq p/n \geq 1/n$, we have $n \to \infty$ and $p/n \to 0$. The two sides imply $\gamma_p \leq \gamma_1 \ll \tau \ll \gamma_p^2$, hence $\gamma_p \to \infty$ and $\tau \to \infty$. Therefore $n/d_0 = 1/\tau \to 0$, and every assumption of Theorem 1 holds. Fix $g \in (0,1)$ and a rate satisfying (33). Lemma D.1 and Markov's inequality give $\sup_{\|\vec{\beta}\|=1} \mathbb{P}\{\mathcal{E}(\vec{w}_n; \vec{\beta}) > g\} \to 0$. The final iterate lies in the training span. For each predetermined target, on the intersection of $\{\mathcal{E}(\vec{w}_n; \vec{\beta}) \leq g\}$ with the input-only event of Theorem 1, the converse (10) applies to $\vec{w}_n$. Since $L_{\mathrm{train}}(\vec{w}_0; \vec{\beta}) = L_0(\vec{\beta})$, its normalized part proves (38) and its raw part proves $L_{\mathrm{train}}(\vec{w}_n; \vec{\beta}) \xrightarrow{\mathbb{P}} \infty$. The same training data are used to compare the initial and final iterates' training losses.

E Fitting and replaying the training data

This appendix proves three statements used in the introduction and Section 4. On fixed training data, any sequence of span predictors whose training loss tends to zero converges to the minimum-norm interpolator (Appendix E.1). Along the geometric window, a constant-rate pass achieving a fixed improvement in expected test risk must have $\eta_n d_0 \to \infty$ (Appendix E.2). In contrast, on an input event of probability at least $1 - n^{-2}$, steps satisfying $0 \leq \eta_k \|\vec{x}_{i_k}\|^2 \leq 2$ keep test risk near one along the geometric window, uniformly over iterations and permitted choices of examples and rates (Proposition E.1 in Appendix E).

E.1 The only interpolator in the training span

Fix training data with $\boldsymbol{X}$ of full column rank. Put $\boldsymbol{K} = \boldsymbol{X}^\top \boldsymbol{X}$ and $\vec{u}_{\mathrm{int}} = \boldsymbol{X}\boldsymbol{K}^{-1}\vec{y}$. A span predictor $\vec{u} = \boldsymbol{X}\vec{\alpha}$ interpolates exactly when $\boldsymbol{K}\vec{\alpha} = \vec{y}$, whose unique solution is $\vec{\alpha} = \boldsymbol{K}^{-1}\vec{y}$. Any ambient interpolator has the form $\vec{u}_{\mathrm{int}} + \vec{z}$ with $\boldsymbol{X}^\top \vec{z} = 0$. Orthogonality gives $\|\vec{u}_{\mathrm{int}} + \vec{z}\|^2 = \|\vec{u}_{\mathrm{int}}\|^2 + \|\vec{z}\|^2$, so $\vec{u}_{\mathrm{int}}$ is the minimum-norm interpolator among all linear predictors.

For a span predictor $\vec{u}$, both $\vec{u}$ and $\vec{u}_{\mathrm{int}}$ lie in the training span, so $\vec{u} - \vec{u}_{\mathrm{int}} = \boldsymbol{X}\boldsymbol{K}^{-1}(\boldsymbol{X}^\top \vec{u} - \vec{y})$. Taking squared norms gives

$$\|\vec{u} - \vec{u}_{\mathrm{int}}\|^2 = (\boldsymbol{X}^\top \vec{u} - \vec{y})^\top \boldsymbol{K}^{-1} (\boldsymbol{X}^\top \vec{u} - \vec{y}) \leq \frac{n}{\lambda_{\min}(\boldsymbol{K})} L_{\mathrm{train}}(\vec{u}; \vec{\beta}). \tag{E.1}$$

Thus on these fixed training data, any sequence of span predictors whose training loss tends to zero converges to $\vec{u}_{\mathrm{int}}$. Zero-initialized GD and SGD stay in the span because each update is a

linear combination of the training inputs. This proves the introduction's statement that passes whose training loss tends to zero approach the minimum-norm interpolator.

For example, full-batch GD on half the training loss multiplies the vector of training residuals by $\boldsymbol{I}_n - \eta \boldsymbol{K}/n$ at each step with step size η, so $0 < \eta < 2n/\lambda_{\max}(\boldsymbol{K})$ makes training loss tend to zero. The classical minimum-norm limit is also discussed by Muthukumar et al. (2021, Section 2.1).

E.2 One-pass learning requires large steps on individual examples

Consider zero-initialized SGD (25) at a deterministic positive constant rate η, with a unit predetermined target $\vec{\beta}$. Fix $g \in (0,1)$. If $\eta\gamma_1 \le 1$, then $0 \le 1-\eta\gamma_1 \le 1-\eta\gamma_j$ for every j, and Jensen's inequality applied to (29) gives

$$\mathbb{E}\mathcal{E}(\vec{w}_n;\vec{\beta}) \ge \sum_j \beta_j^2 (1-\eta\gamma_j)^{2n} \ge (1-\eta\gamma_1)^{2n}. \tag{E.2}$$

If $\eta\gamma_1 > 1$, the lower bound on η in (E.3) below holds automatically, because its right side is below $1/\gamma_1$. Consequently,

$$\mathbb{E}\mathcal{E}(\vec{w}_n;\vec{\beta}) \le g \quad\Longrightarrow\quad \eta \ge \frac{1-g^{1/(2n)}}{\gamma_1}. \tag{E.3}$$

Since $1-g^{1/(2n)} = [-\frac{1}{2}\log g + o(1)]/n$, any such rates along the geometric window (8) obey

$$\eta_n d_0 \ge \left[-\tfrac{1}{2}\log g + o(1)\right]\frac{\tau}{\gamma_1} \longrightarrow \infty. \tag{E.4}$$

Equation (E.4) proves the step-size scale stated at the end of Section 4. When $q_n \to 0$, the rate (D.8) is one example, with $\eta_n d_0 = \vartheta_n \tau/\gamma_p$.

Effect on the example used by an update. For any predictor $\vec{w}$, write $e_i = y_i - \vec{w}^\top \vec{x}_i$ for its residual on training pair i. An SGD update using pair i returns $\vec{w}^+ = \vec{w} + \eta e_i \vec{x}_i$, hence

$$e_i^+ := y_i - (\vec{w}^+)^\top \vec{x}_i = (1-\eta\|\vec{x}_i\|^2) e_i. \tag{E.5}$$

This applies both when the pass first uses pair i, with $\vec{w} = \vec{w}_{i-1}$, and when a later update reuses it. If $\eta\|\vec{x}_i\|^2 > 2$ and $e_i \neq 0$, the update reverses the sign of e_i and increases $|e_i|$. On the Gaussian event (A.13), every training input satisfies $\|\vec{x}_i\|^2 \ge \|\vec{n}_i\|^2 = W_{ii} \ge \lambda_{\min}(\boldsymbol{W}) = (1-o(1))d_0$. Therefore $\min_{i\le n} \eta_n \|\vec{x}_i\|^2 \to \infty$ in probability, so with probability tending to one, every update at such a rate multiplies the residual on its pair by $1-\eta_n\|\vec{x}_i\|^2 < -1$. This concerns the pair used by the update, not the loss on the whole training set. Ranade & Sahai (2026, Section 2 and Appendix A) use this multiplier to compare fresh and reused examples in the stylized model.

E.3 Small steps from zero stay near the zero predictor

Along the geometric window, steps that never enlarge the residual on the example they use cannot learn, however often they reuse the data. For repeated passes, this is a weaker requirement than the one-pass condition $\eta_n d_0 \to \infty$, and the proposition below allows data-dependent rates.

Proposition E.1 (Small per-example steps cannot learn, even with repeated passes)**.** *Let the training inputs $\boldsymbol{X}$ be drawn from* (1)*, and let $u_W = (\sqrt{n}+z_n)/\sqrt{d_0}$ and $u_S = (\sqrt{p}+z_n)/\sqrt{n}$, with $z_n = \sqrt{2\log(4n^2)}$, be the deviation levels of Appendix A.3, assumed below one. On the event* (A.13)–(A.14)*, which has probability at least $1-n^{-2}$ and on which $\boldsymbol{X}$ has full column rank, the following holds for every unit target $\vec{\beta}$, with labels $\vec{y} = \boldsymbol{S}\vec{\beta}$. Start at $\vec{w}_0 = \vec{0}$. At iteration k, choose any training index i_k and rate $\eta_k \ge 0$, possibly using all the training data and the previous iterates, and set $\vec{w}_{k+1} = \vec{w}_k + \eta_k(y_{i_k} - \vec{w}_k^\top \vec{x}_{i_k})\vec{x}_{i_k}$. If every step satisfies $\eta_k\|\vec{x}_{i_k}\|^2 \le 2$, then every iterate satisfies*

$$\left|\sqrt{\mathcal{E}(\vec{w}_k;\vec{\beta})} - 1\right| \le 2\,\frac{1+u_S}{1-u_W}\sqrt{\frac{\gamma_1}{\tau}}. \tag{E.6}$$

Here $\tau = d_0/n$ as in (7)*. Under* (8)*, the right side of* (E.6) *tends to zero, uniformly over unit targets, iterations, and such choices of indices and rates.*

Proof. Let $\boldsymbol{K} = \boldsymbol{X}^\top \boldsymbol{X}$ and $\vec{u}_{\rm int} = \boldsymbol{X}\boldsymbol{K}^{-1}\vec{y}$ be as in Appendix E.1. A permitted step never increases the distance to $\vec{u}_{\rm int}$, so every iterate stays within $\|\vec{u}_{\rm int}\|$ of it, and $\vec{u}_{\rm int}$ has small norm. By (E.5), the step condition means that no update increases the magnitude of the residual on the pair it uses. Because $\vec{u}_{\rm int}$ interpolates, $e_i = y_i - \vec{w}^\top \vec{x}_i = -(\vec{w} - \vec{u}_{\rm int})^\top \vec{x}_i$, and

$$\|\vec{w} + \eta e_i \vec{x}_i - \vec{u}_{\rm int}\|^2 = \|\vec{w} - \vec{u}_{\rm int}\|^2 - \eta(2 - \eta\|\vec{x}_i\|^2)e_i^2. \tag{E.7}$$

The subtracted term in (E.7) is nonnegative for every step. Hence $\|\vec{w}_k - \vec{u}_{\rm int}\| \le \|\vec{u}_{\rm int}\|$ and $\|\vec{w}_k\| \le 2\|\vec{u}_{\rm int}\|$ at every iteration, however often the data are reused. Since $\boldsymbol{K} \succeq \boldsymbol{W}$, the Gaussian event gives, uniformly over unit targets,

$$\|\vec{u}_{\rm int}\|^2 = \vec{y}^\top \boldsymbol{K}^{-1}\vec{y} \le \frac{\|\boldsymbol{S}\|_{\rm op}^2}{\lambda_{\min}(\boldsymbol{W})} \le \left(\frac{1+u_S}{1-u_W}\right)^2 \frac{1}{\tau}. \tag{E.8}$$

Since $\gamma_1 > 1$, the largest eigenvalue of $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\Gamma}, \boldsymbol{I}_{d_0})$ is γ_1. The square root of (2) is the distance from $\boldsymbol{\Sigma}^{1/2}\vec{w}_k$ to the unit vector $(\vec{\beta}, \vec{0})$, so the triangle inequality gives, simultaneously for all iterates,

$$\left|\sqrt{\mathcal{E}(\vec{w}_k; \vec{\beta})} - 1\right| \le \sqrt{\gamma_1}\,\|\vec{w}_k\| \le 2\sqrt{\gamma_1}\,\|\vec{u}_{\rm int}\|. \tag{E.9}$$

Substituting (E.8) into (E.9) proves (E.6). Under (8), $u_W, u_S \to 0$, and $\gamma_1 \ll \tau$ gives $\gamma_1/\tau \to 0$. □

The proof needs neither decreasing full training loss nor convergence of the iterates.

F THE PAIRED EXAMPLE AND ITS NUMERICAL EVALUATION

Proof of Corollary 4. The scalar limit (41) is proved in Appendix F.1. We verify the other clauses below.

Both covariances in (39) have the same first and last diagonal entries. The target uses only those coordinates. Coordinates $2, \dots, M$ are independent of the target label by independence of the Gaussian signal coordinates. The pair changes their variances, not the target or the strengths of the two directions it uses. Because the label uses two different covariance eigenspaces, an orthogonal change of coordinates cannot reduce it to a single eigenspace.

The dimensions give $p/n = (M+1)/M^2 \to 0$ and $n/d_0 = M^{-5} \to 0$. The other quantities in Theorems 1 and 3 are the scale $\tau = d_0/n$ and weighted norm $B(\vec{\beta})$ of (7), and the ratios $R_{\rm wk}$ and q_n of (30):

$$\begin{gathered}\tau = M^5, \quad R_{\rm wk} = M, \quad \gamma_1/\tau = M^{-1}, \quad B(\vec{\beta})^2 = \frac{M^{-8} + M^{-6}}{2}, \\ q_n = \frac{2}{M} \text{ in A}, \qquad q_n = 1 + M^{-2} \text{ in B}.\end{gathered} \tag{F.1}$$

Since $\tau B(\vec{\beta})^2 = (M^{-1} + M^{-3})/2$ and $\tau^2 B(\vec{\beta})^2 = (M^4 + M^2)/2$, Theorem 1 proves both equivalents in (40). They compare the leading risk and training-loss terms. They do not assert that the optima for a fixed finite data set in A and B are equal.

For A, use $\eta_M = \log M/(2M^5)$. It satisfies the three sufficient conditions (33), because

$$n\eta_M\gamma_p = \tfrac{1}{2}\log M, \qquad n\eta_M^2(d_0+2) = \frac{(\log M)^2}{4M}\Big(1 + \frac{2}{M^7}\Big), \qquad \eta_M \operatorname{tr}\boldsymbol{\Gamma}_A = \frac{\log M}{M}, \tag{F.2}$$

where the first diverges and the other two tend to zero. Hence A's final expected test risk tends to zero (Appendix D.2). Family A also satisfies the joint window (37), so Markov's inequality and the converse event give divergence in probability of its raw and normalized final training loss, as in Appendix D.4.

For B, $\beta_{M+1}^2 = 1/2$, and the total strength $H_{M+1} = \operatorname{tr}\boldsymbol{\Gamma}_B$ of directions at least as strong as direction $M+1$ (defined before Theorem 3) gives $H_{M+1}/\gamma_p = M^2 + 1$. Retaining the $j = M+1$ term in (32) gives (42). Taking logarithms of its rate-independent expression gives $2M^2\log(1 - 2/(M^2+3)) \to -4$. In particular,

$$\liminf_{M\to\infty} \inf_{\eta>0} \mathbb{E}\mathcal{E}(\vec{w}_n^{(\eta)}; \vec{\beta}) \ge \tfrac{1}{2}e^{-4}. \tag{F.3}$$

The infimum in (F.3) is over rates chosen before the training data are drawn and held fixed during the pass. □

F.1 A COMMON MULTIPLIER CANNOT MATCH BOTH TARGET AMPLITUDES

We prove both limits in (41), where the scalar may depend on the realized training data and the true distribution. Related heterogeneous-spike obstructions to scalar rescaling appear in Zhao (2026, Section 3.2 and Theorem 11), which compares scalar rescaling with negative-shifted full-batch gradient paths. We first work in probability over either family's Gaussian training data. Convergence of $\mathbb{E}Z_M$, where Z_M is the smallest test risk over scalar multiples of $\vec{u}_{\rm int}$ in (41), will follow from $0 \le Z_M \le 1$.

First compute the interpolator's signal response and nuisance variance. Write the training Gram matrix $\boldsymbol{K} = \boldsymbol{X}^\top \boldsymbol{X}$ of Appendix E.1 as $\boldsymbol{K} = \boldsymbol{W} + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S}^\top$, and take $\boldsymbol{Q} = \boldsymbol{S}^\top \boldsymbol{W}^{-1}\boldsymbol{S}$ from (A.8). Both $\boldsymbol{W}$ and $\boldsymbol{K}$ are positive definite almost surely. The matrix $\boldsymbol{\Gamma}\boldsymbol{Q}$ is similar to the positive-definite matrix $\boldsymbol{\Gamma}^{1/2}\boldsymbol{Q}\boldsymbol{\Gamma}^{1/2}$, so $\boldsymbol{I}_p + \boldsymbol{\Gamma}\boldsymbol{Q}$ is invertible. We show that the interpolator's standardized signal residual $\vec{\beta} - \vec{a}_{\rm int}$ equals

$$\vec{r}_{\rm int} = (\boldsymbol{I}_p + \boldsymbol{\Gamma}\boldsymbol{Q})^{-1}\vec{\beta}. \tag{F.4}$$

Multiplying by $\boldsymbol{K}$ verifies

$$\vec{\alpha}_{\rm int} = \boldsymbol{W}^{-1}\boldsymbol{S}\vec{r}_{\rm int}, \qquad \boldsymbol{K}\vec{\alpha}_{\rm int} = \boldsymbol{S}(\boldsymbol{I}_p + \boldsymbol{\Gamma}\boldsymbol{Q})\vec{r}_{\rm int} = \boldsymbol{S}\vec{\beta} = \vec{y}. \tag{F.5}$$

By (4) and (5), the interpolator's signal response $\vec{a}_{\rm int}$ and nuisance variance $\nu_{\rm int}$ are

$$\vec{a}_{\rm int} = \boldsymbol{\Gamma}\boldsymbol{Q}\vec{r}_{\rm int}, \qquad \nu_{\rm int} = \vec{r}_{\rm int}^{\top}\boldsymbol{Q}\vec{r}_{\rm int}. \tag{F.6}$$

The residual has its stated meaning because $\vec{\beta} - \vec{a}_{\rm int} = \vec{r}_{\rm int}$.

The two-sided bounds (A.13) and (A.14) used in the best-span analysis of Appendix A.3 give $\|\tau\boldsymbol{Q} - \boldsymbol{I}_p\|_{\rm op} \to 0$. For both families, $\tau = M^5$ and $\|\boldsymbol{\Gamma}\|_{\rm op} = M^4$. Consequently $\|\boldsymbol{\Gamma}\boldsymbol{Q}\|_{\rm op} \to 0$. By (F.4), $\vec{r}_{\rm int} = \vec{\beta} - \boldsymbol{\Gamma}\boldsymbol{Q}\vec{r}_{\rm int}$, so whenever $\|\boldsymbol{\Gamma}\boldsymbol{Q}\|_{\rm op} \le 1/2$ the triangle inequality gives $\|\vec{r}_{\rm int}\| \le 2$ and $\|\vec{r}_{\rm int} - \vec{\beta}\| \le 2\|\boldsymbol{\Gamma}\boldsymbol{Q}\|_{\rm op}$. Hence $\|\vec{r}_{\rm int} - \vec{\beta}\| \to 0$, and $\vec{r}_{\rm int}$ stays bounded. Because $\tau = M^5$, (F.6) gives $M\vec{a}_{\rm int} = (\boldsymbol{\Gamma}/M^4)\,\tau\boldsymbol{Q}\vec{r}_{\rm int}$. In both families $\|\boldsymbol{\Gamma}/M^4\|_{\rm op} = 1$, and because the target's only nonzero coordinates are the first and last, $(\boldsymbol{\Gamma}/M^4)\vec{\beta} = (\vec{e}_1 + M^{-1}\vec{e}_{M+1})/\sqrt{2}$. Adding and subtracting $(\boldsymbol{\Gamma}/M^4)\vec{\beta}$ gives

$$\left\| M\vec{a}_{\rm int} - \frac{\vec{e}_1}{\sqrt{2}} \right\| \le \left\| \frac{\boldsymbol{\Gamma}}{M^4} \right\|_{\rm op} \|\tau\boldsymbol{Q}\vec{r}_{\rm int} - \vec{\beta}\| + \frac{1}{\sqrt{2}M} \longrightarrow 0. \tag{F.7}$$

Also

$$0 \le M^2\nu_{\rm int} \le M^{-3}\|\tau\boldsymbol{Q}\|_{\rm op}\|\vec{r}_{\rm int}\|^2 \longrightarrow 0. \tag{F.8}$$

Thus, after scaling the interpolator by M, the amplitude of the stronger coordinate 1 tends to its required value $1/\sqrt{2}$, while that of the weaker coordinate $M+1$, which carries the other half of the label variance, tends to zero.

The test-risk identity (5) gives the exact scalar quadratic

$$\mathcal{E}(c\vec{u}_{\rm int}; \vec{\beta}) = 1 - 2c\,\vec{\beta}^{\top}\vec{a}_{\rm int} + c^2(\|\vec{a}_{\rm int}\|^2 + \nu_{\rm int}). \tag{F.9}$$

Its quadratic coefficient is positive almost surely. Completing the square gives

$$Z_M = 1 - \frac{(\vec{\beta}^{\top}\vec{a}_{\rm int})^2}{\|\vec{a}_{\rm int}\|^2 + \nu_{\rm int}}, \qquad c_{\rm opt} = \frac{\vec{\beta}^{\top}\vec{a}_{\rm int}}{\|\vec{a}_{\rm int}\|^2 + \nu_{\rm int}}. \tag{F.10}$$

Equations (F.7) and (F.8) imply $M\vec{\beta}^{\top}\vec{a}_{\rm int} \to 1/2$ and $M^2(\|\vec{a}_{\rm int}\|^2 + \nu_{\rm int}) \to 1/2$. Substitution proves $Z_M \to 1/2$ and $c_{\rm opt}/M \to 1$. Allowing $c = 0$ gives $0 \le Z_M \le 1$, so $\mathbb{E}Z_M \to 1/2$ because for every $\epsilon > 0$, $\mathbb{E}|Z_M - 1/2| \le \epsilon + \mathbb{P}\{|Z_M - 1/2| > \epsilon\}$. This proves (41) and completes the proof of Corollary 4.

The unmodified interpolator has test risk tending to one by (F.9). Its best scalar multiple has test risk tending to $1/2$, so inflation helps but cannot make test risk vanish. On each realization of the training data, every scalar choice independent of the fresh test pair has risk at least Z_M. Thus its expected risk has lower limit at least $1/2$, and its risk is at least $1/2 - \epsilon$ with probability tending to one for each $\epsilon > 0$. In particular, $c_{\rm opt}$ exceeds one with probability tending to one.

F.2 NEGATIVE SPAN RIDGE CAN LEARN BOTH MEMBERS OF THE MATCHED PAIR

The guarded estimator (F.20) below uses only the training data and the supplied n and d_0, and its expected test risk is $O(M^{-1})$ in both families. The ordinary negative-ridge calculation gives the risk bound with high probability, and the spectral check in (F.20) controls the exceptional data sets that the expectation also counts. A negative ridge penalty can subtract most of the large nuisance term from the Gram matrix (Tsigler & Bartlett, 2023, Section 8). We subtract slightly less than $d_0\boldsymbol{I}_n$ from the training Gram matrix, leaving a margin for the fluctuations of $\boldsymbol{W}$ around $d_0\boldsymbol{I}_n$, so that the inverse in (F.11) stays bounded.

A deterministic bound under two input conditions. Consider the span-restricted objective

$$\frac{1}{n}\|\boldsymbol{X}^\top\vec{u}-\vec{y}\|^2+\lambda\|\vec{u}\|^2, \qquad \vec{u}\in\mathrm{col}(\boldsymbol{X}).$$

For $\lambda<0$, an off-span component would reduce the penalty without changing the training predictions, so the unrestricted objective would be unbounded below. Within the span, the objective has a unique minimizer whenever $\boldsymbol{K}+n\lambda\boldsymbol{I}_n\succ 0$, where $\boldsymbol{K}=\boldsymbol{X}^\top\boldsymbol{X}$. The minimizer is

$$\vec{u}_\lambda=\boldsymbol{X}(\boldsymbol{K}+n\lambda\boldsymbol{I}_n)^{-1}\vec{y}. \tag{F.11}$$

A single ridge parameter is not a common multiplier of the interpolator $\vec{u}_{\mathrm{int}}=\boldsymbol{X}\boldsymbol{K}^{-1}\vec{y}$. If $\boldsymbol{K}\vec{v}=k\vec{v}$ with $\|\vec{v}\|=1$, the component of $\vec{u}_\lambda$ along $\boldsymbol{X}\vec{v}$ is $k/(k+n\lambda)$ times that of $\vec{u}_{\mathrm{int}}$, so eigenvectors of the observed Gram matrix with different eigenvalues receive different multipliers. These eigenvectors come from the observed training inputs, not from the hidden signal directions.

Let $\Delta>0$ and choose $n\lambda=2\Delta-d_0$. The diagonal shift $n\lambda=2\Delta-d_0$ is negative when $\Delta<d_0/2$, as happens eventually for the choice in (F.18). Suppose the training inputs satisfy

$$\|\boldsymbol{W}-d_0\boldsymbol{I}_n\|_{\mathrm{op}}\le\Delta, \qquad \sigma_{\min}(\boldsymbol{S})\ge b>0. \tag{F.12}$$

Put $\boldsymbol{C}_\lambda=\boldsymbol{W}+n\lambda\boldsymbol{I}_n$, the shifted nuisance Gram matrix. The first condition in (F.12) gives the first bound below, so $\boldsymbol{C}_\lambda$ is invertible. With $\boldsymbol{Q}_\lambda=\boldsymbol{S}^\top\boldsymbol{C}_\lambda^{-1}\boldsymbol{S}$, the shifted counterpart of $\boldsymbol{Q}$ in (A.8), the second condition then gives the second bound:

$$\Delta\boldsymbol{I}_n\preceq\boldsymbol{C}_\lambda\preceq 3\Delta\boldsymbol{I}_n, \qquad \boldsymbol{Q}_\lambda\succeq\frac{b^2}{3\Delta}\boldsymbol{I}_p. \tag{F.13}$$

In particular the inverse in (F.11) is positive definite. Write $\vec{u}_\lambda=\boldsymbol{X}\vec{\alpha}_\lambda$. The test-risk identity (5) needs its signal residual $\vec{r}_\lambda=\vec{\beta}-\boldsymbol{\Gamma}\boldsymbol{S}^\top\vec{\alpha}_\lambda$ and its nuisance variance $\vec{\alpha}_\lambda^\top\boldsymbol{W}\vec{\alpha}_\lambda$. The residual and the coefficient vector are

$$\vec{r}_\lambda=(\boldsymbol{\Gamma}^{-1}+\boldsymbol{Q}_\lambda)^{-1}\boldsymbol{\Gamma}^{-1}\vec{\beta}, \qquad \vec{\alpha}_\lambda=\boldsymbol{C}_\lambda^{-1}\boldsymbol{S}\vec{r}_\lambda, \tag{F.14}$$

where $\|\boldsymbol{\Gamma}^{-1}\vec{\beta}\|=B(\vec{\beta})$. Multiplication verifies that this $\vec{\alpha}_\lambda$ solves $(\boldsymbol{K}+n\lambda\boldsymbol{I}_n)\vec{\alpha}_\lambda=\boldsymbol{S}\vec{\beta}$, as (F.11) requires, and that $\vec{\beta}-\boldsymbol{\Gamma}\boldsymbol{S}^\top\vec{\alpha}_\lambda=\vec{r}_\lambda$. Because $\boldsymbol{\Gamma}^{-1}+\boldsymbol{Q}_\lambda\succeq b^2/(3\Delta)\boldsymbol{I}_p$ by (F.13), its inverse has operator norm at most $3\Delta/b^2$.

Since $\boldsymbol{\Gamma}^{-1}\vec{\beta}=(\boldsymbol{\Gamma}^{-1}+\boldsymbol{Q}_\lambda)\vec{r}_\lambda$ by (F.14), $\vec{r}_\lambda^\top\boldsymbol{\Gamma}^{-1}\vec{\beta}=\vec{r}_\lambda^\top\boldsymbol{\Gamma}^{-1}\vec{r}_\lambda+\vec{r}_\lambda^\top\boldsymbol{Q}_\lambda\vec{r}_\lambda\ge\vec{r}_\lambda^\top\boldsymbol{Q}_\lambda\vec{r}_\lambda$. With the operator-norm bound just derived, $\|\boldsymbol{\Gamma}^{-1}\vec{\beta}\|=B(\vec{\beta})$, and Cauchy–Schwarz,

$$\|\vec{r}_\lambda\|\le\frac{3\Delta}{b^2}B(\vec{\beta}), \qquad \vec{r}_\lambda^\top\boldsymbol{Q}_\lambda\vec{r}_\lambda\le\vec{r}_\lambda^\top\boldsymbol{\Gamma}^{-1}\vec{\beta}\le\frac{3\Delta}{b^2}B(\vec{\beta})^2. \tag{F.15}$$

Also $\vec{\alpha}_\lambda^\top\boldsymbol{C}_\lambda\vec{\alpha}_\lambda=\vec{r}_\lambda^\top\boldsymbol{Q}_\lambda\vec{r}_\lambda$. Since $\boldsymbol{C}_\lambda\succeq\Delta\boldsymbol{I}_n$ and $\boldsymbol{W}\preceq(d_0+\Delta)\boldsymbol{I}_n$, the nuisance variance obeys

$$\vec{\alpha}_\lambda^\top\boldsymbol{W}\vec{\alpha}_\lambda\le\frac{d_0+\Delta}{\Delta}\,\vec{\alpha}_\lambda^\top\boldsymbol{C}_\lambda\vec{\alpha}_\lambda=\frac{d_0+\Delta}{\Delta}\,\vec{r}_\lambda^\top\boldsymbol{Q}_\lambda\vec{r}_\lambda\le\frac{3(d_0+\Delta)}{b^2}B(\vec{\beta})^2. \tag{F.16}$$

Putting the squared residual bound from (F.15) and the nuisance bound (F.16) into the test-risk identity (5) gives

$$\mathcal{E}(\vec{u}_\lambda;\vec{\beta})\le\left[\frac{9\Delta^2}{b^4}+\frac{3(d_0+\Delta)}{b^2}\right]B(\vec{\beta})^2. \tag{F.17}$$

The bound (F.17) depends on the nuisance Gram fluctuation Δ and the lower bound b on the smallest singular value of the standardized signal matrix $\boldsymbol{S}$. It has no signal-trace term. Raising only the strengths of directions irrelevant to the label leaves $B(\vec{\beta})$ unchanged, and neither Δ nor b depends on the strengths. For the target of (39), the weakest direction carries half the label variance, so vanishing expected risk for constant-rate SGD requires $\eta<2/(\mathrm{tr}\,\boldsymbol{\Gamma}+2\gamma_p)$ eventually, as in the proof of (32). The signal trace $\mathrm{tr}\,\boldsymbol{\Gamma}$ enters that SGD requirement but not (F.17).

Gaussian inputs meet the conditions with high probability. For the two families of Corollary 4, take

$$\theta_n = \sqrt{2\log(8n^6)}, \qquad \Delta_n = 2\sqrt{d_0}(\sqrt{n}+\theta_n) + (\sqrt{n}+\theta_n)^2, \qquad n\lambda_n = 2\Delta_n - d_0. \quad \text{(F.18)}$$

The Gaussian singular-value bounds used in Appendix A give (F.12), with $\Delta = \Delta_n$ and $b_n = \sqrt{n} - \sqrt{p} - \theta_n > 0$, on an input event G_n of probability at least $1 - n^{-6}$ for all sufficiently large M. Here $\Delta_n = (2 + o(1))\sqrt{nd_0} = o(d_0)$ and $b_n^2/n \to 1$. The penalty is negative eventually. Substituting these into (F.17) makes its two coefficients $(36+o(1))\tau$ and $(3+o(1))\tau$. Equation (F.17) therefore gives, on G_n simultaneously for all unit targets,

$$\mathcal{E}(\vec{u}_{\lambda_n}; \vec{\beta}) \le (39 + o(1))\tau B(\vec{\beta})^2. \quad \text{(F.19)}$$

For the target $\vec{\beta}$ of (39), $\tau B(\vec{\beta})^2 = (M^{-1} + M^{-3})/2$. Extending the estimator by zero when its shifted Gram matrix is not positive definite proves $\mathcal{E}(\vec{u}_{\lambda_n}; \vec{\beta}) = O_{\mathbb{P}}(M^{-1})$ in both families. This in-probability bound does not control the expected risk, because on exceptional data sets the shifted Gram matrix $\boldsymbol{K} + n\lambda_n \boldsymbol{I}_n$ can be positive definite yet arbitrarily close to singular.

A spectral safety check permits an expected-risk comparison. To exclude arbitrarily near-singular accepted solves, define

$$\widehat{\vec{u}}_{\lambda_n} = \begin{cases} \boldsymbol{X}(\boldsymbol{K} + n\lambda_n \boldsymbol{I}_n)^{-1}\vec{y}, & \lambda_{\min}(\boldsymbol{K} + n\lambda_n \boldsymbol{I}_n) \ge \Delta_n,\ \Delta_n < d_0/2, \\ \vec{0}, & \text{otherwise.} \end{cases} \quad \text{(F.20)}$$

The check uses the observed Gram matrix and the supplied n, d_0. It uses no signal directions, target coordinates, or validation observations. On G_n the guard accepts for all sufficiently large M, because $\boldsymbol{K} \succeq \boldsymbol{W}$ and (F.13) give $\boldsymbol{K} + n\lambda_n \boldsymbol{I}_n \succeq \boldsymbol{C}_{\lambda_n} \succeq \Delta_n \boldsymbol{I}_n$, and $\Delta_n = o(d_0)$.

Whenever it accepts, diagonalizing $\boldsymbol{K}$ and using $n\lambda_n < 0$ gives

$$\|\widehat{\vec{u}}_{\lambda_n}\|^2 = \vec{y}^\top (\boldsymbol{K} + n\lambda_n \boldsymbol{I}_n)^{-1} \boldsymbol{K} (\boldsymbol{K} + n\lambda_n \boldsymbol{I}_n)^{-1} \vec{y} \le \frac{d_0}{\Delta_n^2} \|\vec{y}\|^2. \quad \text{(F.21)}$$

Indeed $k/(k + n\lambda_n)^2$ is decreasing for $k \ge \Delta_n - n\lambda_n = d_0 - \Delta_n$. The bound also holds for the zero output.

Since $(y - \widehat{\vec{u}}^\top \vec{x})^2 \le 2y^2 + 2(\widehat{\vec{u}}^\top \vec{x})^2$ for $\widehat{\vec{u}} = \widehat{\vec{u}}_{\lambda_n}$, and $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\Gamma}, \boldsymbol{I}_{d_0})$ has $\|\boldsymbol{\Sigma}\|_{\mathrm{op}} = M^4$, the test risk satisfies

$$\mathcal{E}(\widehat{\vec{u}}_{\lambda_n}; \vec{\beta}) \le 2 + 2\widehat{\vec{u}}^\top \boldsymbol{\Sigma} \widehat{\vec{u}} \le 2 + 2M^4 \|\widehat{\vec{u}}_{\lambda_n}\|^2.$$

With (F.21) and $\Delta_n^2 \ge 4d_0 n$, the right side is at most $2 + M^4\|\vec{y}\|^2/(2n)$. Cauchy–Schwarz gives $\mathbb{E}[\|\vec{y}\|^2 \mathbf{1}_{G_n^c}] \le \sqrt{\mathbb{E}\|\vec{y}\|^4\, \mathbb{P}(G_n^c)}$, which needs no independence between $\vec{y}$ and G_n. With $\mathbb{E}\|\vec{y}\|^4 = n(n+2)$ for every predetermined unit target and $\mathbb{P}(G_n^c) \le n^{-6}$, this gives

$$\mathbb{E}[\mathcal{E}(\widehat{\vec{u}}_{\lambda_n}; \vec{\beta}) \mathbf{1}_{G_n^c}] \le 2n^{-6} + \frac{M^4}{2}\sqrt{1 + 2/n}\, n^{-3} = O(M^{-2}). \quad \text{(F.22)}$$

Both bounds hold for every predetermined unit target, and $\sup_{\|\vec{\beta}\|=1} \tau B(\vec{\beta})^2 = \tau/\gamma_p^2 = 1/M$ because both families have $\tau = M^5$ and $\gamma_p = M^3$. Combining (F.19) and (F.22) therefore yields

$$\mathbb{E}\mathcal{E}(\widehat{\vec{u}}_{\lambda_n}; \vec{\beta}) \le (39 + o(1))\tau B(\vec{\beta})^2 + O(M^{-2}), \qquad \sup_{\|\vec{\beta}\|=1} \mathbb{E}\mathcal{E}(\widehat{\vec{u}}_{\lambda_n}; \vec{\beta}) = O(M^{-1}). \quad \text{(F.23)}$$

Thus the guarded ridge estimator has vanishing expected risk even in B. The guarded predictor remains a span predictor, so by Theorem 1 and Markov's inequality its raw and normalized training losses still diverge in probability for every predetermined unit target.

F.3 A SHARPER TRAINING-LOSS BOUND FOR FAMILY A

Corollary 4 needs only divergence. The converse (10) uses the largest strength γ_1, which for this pair gives the training-loss scale $(\tau/\gamma_1)^2 = M^2$. Figure 3 uses the larger target-dependent scale $\tau^2 B(\vec{\beta})^2 = (M^4 + M^2)/2$ of the best span predictor in (9). To obtain a lower bound at this scale for A's SGD pass, we keep the target dependence through the second response bound in (A.5). Let $\vec{w}_n$ be the final iterate of (25) for family A at the rate $\eta_M = \log M/(2M^5)$ of Corollary 4, with the target $\vec{\beta}$ of (39), and abbreviate $B = B(\vec{\beta})$. For every fixed $\epsilon \in (0, 1)$,

$$\mathbb{P}\left\{ \min\left(L_{\mathrm{train}}(\vec{w}_n; \vec{\beta}), \frac{L_{\mathrm{train}}(\vec{w}_n; \vec{\beta})}{L_0(\vec{\beta})} \right) \geq (1-\epsilon)\frac{M^4 + M^2}{2} \right\} \to 1. \tag{F.24}$$

To prove (F.24), no rate of convergence for the risk is needed. Corollary 4 gives $\mathbb{E}\mathcal{E}(\vec{w}_n; \vec{\beta}) \to 0$, so the realized risk $g_M := \mathcal{E}(\vec{w}_n; \vec{\beta})$ tends to zero in probability. The final iterate lies in the training span, and the finite bounds (A.3)–(A.4) hold for every $g \geq 0$ at once, so they apply with $g = g_M$. For this target, $\gamma_p B = \sqrt{(1 + M^{-2})/2} \to 1/\sqrt{2}$ and $\tau B = \sqrt{(M^4 + M^2)/2} \to \infty$, so $\sqrt{g_M}/(\gamma_p B)$ and $\sqrt{g_M}/(\tau B)$ tend to zero in probability. On the input event (A.2) with $\delta = n^{-2}$, put the target-dependent response bound of (A.5) into (A.3) and (A.4), and divide both by $\tau^2 B^2$. Since $\Lambda_\delta/d_0 \to 1$ and $s_\delta/\sqrt{n} \to 1$, both resulting lower coefficients tend to one in probability. The input event fails with probability at most n^{-2}, so (F.24) follows. The proof uses SGD only through two facts, that its final iterate lies in the training span and that its test risk tends to zero. This is not an upper bound on the SGD training loss.

F.4 EXACT EXPECTED RISKS USE A THREE-DIMENSIONAL RECURRENCE

The proofs above use no numerical optimization or sampled runs. This and the following subsections describe the exact expected-risk calculations and the sampled illustrations. For computation only, write $\boldsymbol{\Gamma}_{M,k} = \mathrm{diag}(M^4 \boldsymbol{I}_k, M^3 \boldsymbol{I}_{M+1-k})$. A has $k = 1$ and B has $k = M$. Sum the risk contributions $T_{t,j}$ of (26) over the k coordinates of strength M^4 and over the $M + 1 - k$ coordinates of strength M^3, and keep N_t. With $a = \eta M^4$, $b = a/M$, $r = M + 1 - k$, and $\delta_a = a^2/M = \eta^2 d_0$, the transition on the strong, weak, and nuisance totals is the following matrix, where $\eta = a/M^4$:

$$\boldsymbol{A}_{M,k}(a) = \begin{pmatrix} 1 - 2a + (k+2)a^2 & ka^2 & ka^2 \\ rb^2 & 1 - 2b + (r+2)b^2 & rb^2 \\ \delta_a & \delta_a & 1 - 2\eta + \delta_a + 2\eta^2 \end{pmatrix}. \tag{F.25}$$

The initial totals are $(1/2, 1/2, 0)^\top$. Multiplying by $\boldsymbol{A}_{M,k}(a)^{M^2}$ and summing gives the exact expected test risk from (27) and (28). This is not a Monte Carlo estimate.

Figure 1(b) uses $\{2^j : 2 \leq j \leq 12\} \cup \{3 \cdot 2^j : 1 \leq j \leq 10\}$, which has 21 values of M. A uses its specified rate. For B, a diagnostic search takes 4,097 geometrically spaced a values in $[2^{-24}, 2/3]$, and also includes risk one for the zero predictor. Every grid point is retained. Powers of the transition matrix are rescaled during the computation, and logarithms record their scale. The smallest grid value is not a certificate of the continuous minimum. The separate curve (42) is the proved bound for every positive constant rate. At $M = 4, 64, 4096$, doubling the grid lowers its minimum by less than 9×10^{-6} in relative terms, and 80-decimal evaluations at A's rate and at B's grid minimizer agree with double precision to relative error below 4×10^{-10}.

F.5 EXACT GAUSSIAN GRAM SAMPLING AVOIDS AN ENORMOUS NUISANCE MATRIX

The span coefficients and losses in these experiments depend on the nuisance inputs only through $\boldsymbol{W} = \boldsymbol{N}^\top \boldsymbol{N}$, so we sample $\boldsymbol{S}$ and $\boldsymbol{W}$ from their exact joint law instead of forming the $d_0 \times n$ matrix $\boldsymbol{N}$. Figure 3 uses $M \in \{4, 6, 8, 12, 16\}$ with 64 paired training data sets per value, and panel (a) of Figure 1 uses the 64 data sets at $M = 16$ from the same experiment. The root seed is 20260919. A NumPy `SeedSequence` (NumPy 2.3.5) spawns 320 child sequences, ordered first by M and then by replicate, each seeding a `default_rng` generator (PCG64). Each child generator draws an $n \times p$ standard Gaussian matrix $\boldsymbol{S}$. It next draws the entries below the diagonal of a lower-triangular $n \times n$ matrix $\boldsymbol{L}$, in row-major order, as independent standard normals divided by $\sqrt{d_0}$.

Finally, it draws diagonal entries $L_{ii} = \sqrt{\chi^2_{d_0-i+1}/d_0}$ independently. The Bartlett representation gives $\overline{\boldsymbol{W}} = \boldsymbol{L}\boldsymbol{L}^\top$ with the same law as $\boldsymbol{W}/d_0$ (Muirhead, 1982). The same $\boldsymbol{S}$ and $\overline{\boldsymbol{W}}$ are used for A and B. The unnormalized formulas are applied with $\boldsymbol{W} = d_0\overline{\boldsymbol{W}}$, and the stored coefficient vector is $\widetilde{\vec{\alpha}} = d_0\vec{\alpha}$.

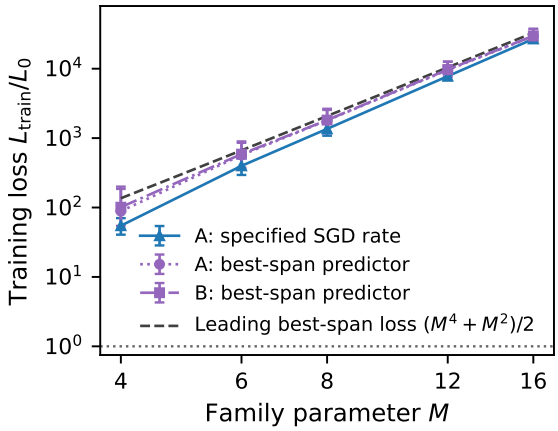


Figure 3: The best-span training loss is asymptotic to $(M^4 + M^2)/2$ in both families, and (F.24) gives the same asymptotic lower bound for A's SGD pass; all are far above the zero predictor's. We plot final training loss divided by L_0 for these three predictors. Markers are medians and bars are pointwise 10th–90th percentiles over 64 paired Gaussian data sets at each M. The best span predictors use population information. The dashed curve is the best-span asymptotic equivalent, not a finite-data bound. The dotted line is one, the zero predictor's normalized training loss.

The best-span coefficients are computed from (A.9) and (A.11), using positive-definite solves. For SGD, let $\boldsymbol{K} = \boldsymbol{X}^\top\boldsymbol{X}$ and let $\boldsymbol{K}_<$ be its strictly lower-triangular part in the order the examples are processed. The coefficient assigned to example i is $\alpha_i = \eta(y_i - \sum_{j<i} K_{ij}\alpha_j)$. Therefore

$$\vec{\alpha}_{\text{SGD}} = (\eta^{-1}\boldsymbol{I}_n + \boldsymbol{K}_<)^{-1}\vec{y}. \tag{F.26}$$

The triangular solve (F.26) evaluates the same coefficients as the sequential SGD updates. It is a computational shortcut for the experiment, not information supplied to the learner. For Figure 3, evaluate test risk by (5) and training loss by (6), dividing the latter by the same data set's L_0.

The percentiles in Figure 3 use linear interpolation over the 64 data sets at each M. No result is excluded.

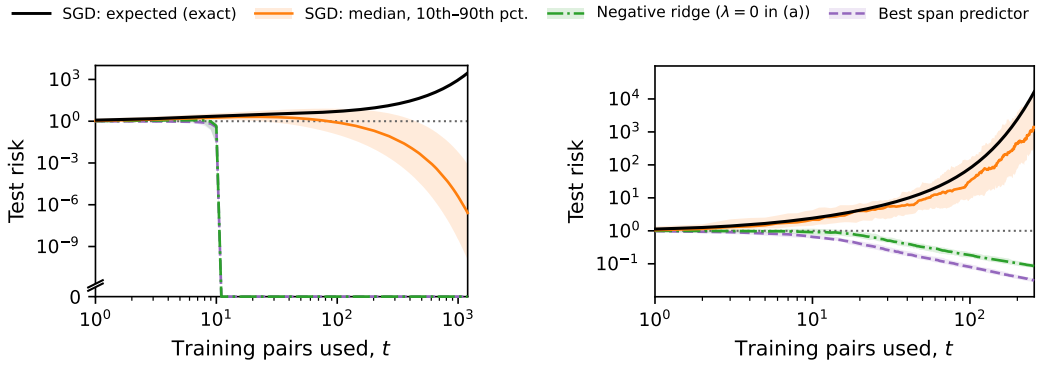
At $M = 4$, A's SGD pass has median test risk 0.237 and median normalized training loss 54.7. At $M = 6$, the finite bounds of Appendix A.1 with $\delta = 0.05$ and the target-dependent bound of (A.5) give probability at least 0.95 that every span predictor for the target of (39) with test risk at most $1/4$ has training loss above 2.5 times the zero predictor's. That certificate is not vacuous, because 58 of the 64 runs of A's SGD pass have test risk at most $1/4$, each with normalized training loss above 256. The archives contain every risk, loss, seed, and tested rate.

Panel (a) of Figure 1 uses the 64 data sets at $M = 16$, with child sequences 256 through 319 and the rate η_M of Corollary 4. At time t, set $\vec{\alpha}^{(t)}$ to the first t coefficients of (F.26), followed by zeros. The coefficient vector $\vec{\alpha}^{(t)}$ gives the SGD iterate $\vec{w}_t = \boldsymbol{X}\vec{\alpha}^{(t)}$. Evaluate its test risk by (5) and its training loss by (6). At every time, the latter uses all n training pairs, including pairs not yet used in an update. This evaluation does not change the updates or select the rate. No trajectory is smoothed or excluded, and the final values agree with the archived A-SGD results. The curves and bands summarize the data sets separately at each time.

Panel (a) plots training loss without dividing by L_0, which is near one at $n = 256$. After the first update, the median training loss is about 260. The residual multiplier $1 - \eta_M\|\vec{x}_1\|^2$ of (E.5) explains the immediate overshoot on that first training pair.

F.6 TYPICAL RUNS CAN ESCAPE A GROWING MEAN; SIMULATED B RUNS DO NOT

Theorem 3 controls expected test risk, which averages over training runs. Figure 4 compares that average with the median and the 10th and 90th percentiles of the test risk of SGD across runs. Expected risk need not describe most runs. In panel (a), an eleven-dimensional example, the exact expected risk grows while the median falls. In panel (b), family B at the rate $\eta = 2/(\operatorname{tr}\mathbf{\Gamma} + 2\gamma_p)$, both grow. Both panels also show negative ridge and the best span predictor, fit to the same first t training pairs.



(a) The 11-dimensional example

(b) Family B at $M = 16$

Figure 4: The expected test risk of SGD can grow while most runs improve (a), or grow along with most runs (b). Black curves are exact expected test risks from Gaussian recurrences. The other curves are medians, with bands from the 10th to the 90th percentile across runs at each t. Negative ridge and the best span predictor are fit to the first t training pairs of each run. (a) $\mathbf{\Sigma} = \operatorname{diag}(3.5\boldsymbol{I}_{10}, 1)$, label $s[11]$, $\eta = 19/400$, and 200,000 runs. With $d_0 = 0$, the green curve is the minimum-norm interpolator. It and the best span predictor have test risk exactly zero for $t \geq 11$, drawn on the horizontal axis below the axis break. (b) Family B at $M = 16$ with $\eta = 2/(\operatorname{tr}\mathbf{\Gamma} + 2\gamma_p)$, on the 64 B data sets of Appendix F.5. Both axes are logarithmic apart from that zero. The dotted line is the zero predictor's risk, one.

Panel (a) uses a small example without a nuisance block, so $d_0 = 0$ and it lies outside the misfitting window. Take $\mathbf{\Sigma} = \operatorname{diag}(\gamma\boldsymbol{I}_{10}, 1)$ with strength $\gamma = 7/2$, inputs $\vec{x}_t = \mathbf{\Sigma}^{1/2}\vec{s}_t$ with $\vec{s}_t \sim \mathcal{N}(\vec{0}, \boldsymbol{I}_{11})$, labels $y_t = s_t[11]$, and the update (25) at the rate $\eta = 19/400$. The standardized residual $\vec{r}_t = \vec{e}_{11} - \mathbf{\Sigma}^{1/2}\vec{w}_t$ gives test risk $\|\vec{r}_t\|^2$. Let $T_t^{\text{weak}} = \mathbb{E}r_{t,11}^2$ and $T_t^{\text{strong}} = \mathbb{E}\sum_{j=1}^{10} r_{t,j}^2$. Applying (C.1) with this $\mathbf{\Sigma}$ and summing the diagonal of $\mathbf{\Sigma}\mathbf{\Omega}_t$ over the two blocks gives, from $(T_0^{\text{weak}}, T_0^{\text{strong}}) = (1, 0)$,

$$\begin{pmatrix} T_{t+1}^{\text{weak}} \\ T_{t+1}^{\text{strong}} \end{pmatrix} = \begin{pmatrix} 1 - 2\eta + 3\eta^2 & \eta^2 \\ 10\eta^2\gamma^2 & 1 - 2\eta\gamma + 12\eta^2\gamma^2 \end{pmatrix} \begin{pmatrix} T_t^{\text{weak}} \\ T_t^{\text{strong}} \end{pmatrix}. \tag{F.27}$$

Both columns of the nonnegative transition matrix in (F.27) sum to at least 1.001425, so the expected test risk $T_t^{\text{weak}} + T_t^{\text{strong}}$ is at least 1.001425^t and diverges. Yet in 200,000 simulated runs at this rate, the median test risk falls below 10^{-3} only at $t = 594$, is below 10^{-6} at $t = 1200$, and the 90th percentile is below 10^{-3} at $t = 1200$. A small fraction of runs with very large risk determines the expectation, as Nascimento & Sayed (2000, Section III-C) observed for LMS adaptive filters.

The growing mean is a property of the rate $\eta = 19/400$, about twice the mean-square-safe rate $1/(\operatorname{tr}\mathbf{\Sigma} + 2\|\mathbf{\Sigma}\|_{\text{op}}) = 1/43$ treated below. At that smaller rate the recursion (F.27) gives expected test risk about 9.6×10^{-24} at $t = 1200$. In panel (a), once $t \geq 11$, the inputs span $\mathbb{R}^{11}$ almost surely, so the minimum-norm interpolator and the best span predictor both recover the true coefficients and have test risk exactly zero.

Panel (b) uses family B at $M = 16$ on the 64 data sets of Appendix F.5, at the rate $\eta = 2/(\operatorname{tr}\mathbf{\Gamma} + 2\gamma_p)$. At and above this rate, (35) shows that the expected squared error V_t in the original signal coefficients cannot decrease. The expected risk of SGD is exact from (F.25), and the iterates come from (F.26). At $t = n$, the median test risks are 0.031 for the best span predictor, 0.086 for negative

ridge, and 0.205 for the guarded estimator of Appendix F.2. The median test risk of SGD instead rises from one to about $1{,}400$.

The mean-square-safe comparison rate. The mean-square-safe rate guarantees mean-square convergence on an unlimited stream of fresh pairs. For the following calculation only, take any fixed covariance $\boldsymbol{\Sigma} \succ 0$ in the noiseless Gaussian model and set $\eta = 1/(\operatorname{tr}\boldsymbol{\Sigma} + 2\|\boldsymbol{\Sigma}\|_{\mathrm{op}})$. Write $\vec{h}_t = \vec{w}_\star - \vec{w}_t$ for the coefficient error relative to the zero-risk predictor $\vec{w}_\star$. The Gaussian identity $\mathbb{E}[\|\vec{x}\|^2 \vec{x}\vec{x}^\top] = (\operatorname{tr}\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^2$ gives, given the first t pairs,

$$\begin{aligned}\mathbb{E}[\|\vec{h}_{t+1}\|^2 \mid \mathcal{F}_t] &= \vec{h}_t^\top \big[\boldsymbol{I} - 2\eta\boldsymbol{\Sigma} + \eta^2\big((\operatorname{tr}\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^2\big)\big]\vec{h}_t \\ &\le \vec{h}_t^\top(\boldsymbol{I} - \eta\boldsymbol{\Sigma})\vec{h}_t \le \big(1 - \eta\lambda_{\min}(\boldsymbol{\Sigma})\big)\|\vec{h}_t\|^2.\end{aligned} \tag{F.28}$$

The first inequality uses $\boldsymbol{\Sigma}^2 \preceq \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\boldsymbol{\Sigma}$ and $\eta(\operatorname{tr}\boldsymbol{\Sigma} + 2\|\boldsymbol{\Sigma}\|_{\mathrm{op}}) = 1$, and the second uses the smallest eigenvalue of $\boldsymbol{\Sigma}$. Since the test risk is $\vec{h}_t^\top\boldsymbol{\Sigma}\vec{h}_t \le \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\|\vec{h}_t\|^2$ and $\vec{h}_0 = \vec{w}_\star$, iterating (F.28) gives $\mathbb{E}\mathcal{E}(\vec{w}_t;\vec{\beta}) \le \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\big(1 - \eta\lambda_{\min}(\boldsymbol{\Sigma})\big)^t\|\vec{w}_\star\|^2$, which tends to zero as $t \to \infty$ for this fixed $\boldsymbol{\Sigma}$.

Computing the comparison predictors. Negative ridge uses (F.11) with n replaced by t and penalty $t\lambda = \frac{5}{2}\sqrt{td_0} - d_0$. This penalty is chosen for illustration. It differs from the guarded estimator (F.20) used in the proof. In panel (b), the shifted Gram matrix of negative ridge stayed positive definite, with smallest eigenvalue above $\frac{1}{2}\sqrt{td_0}$ at every t on every data set, and the guarded estimator keeps a larger margin. When $d_0 = 0$ the penalty vanishes, and the green curve is the minimum-norm interpolator $\boldsymbol{X}_t(\boldsymbol{X}_t^\top\boldsymbol{X}_t)^\dagger\vec{y}_{1:t}$, where $\boldsymbol{X}_t$ holds the first t inputs, $\vec{y}_{1:t}$ their labels, and † denotes the Moore–Penrose pseudoinverse. The best span predictor has the smallest test risk in the span of those pairs, which requires the true distribution. In panel (b), with $\boldsymbol{W}_t$ and $\boldsymbol{S}_t$ the nuisance Gram matrix and standardized signal matrix of the first t pairs, its coefficients solve $(\boldsymbol{W}_t + \boldsymbol{S}_t\boldsymbol{\Gamma}^2\boldsymbol{S}_t^\top)\vec{\alpha} = \boldsymbol{S}_t\boldsymbol{\Gamma}\vec{\beta}$, the normal equations of the test-risk quadratic (5) on those pairs. This needs only $\boldsymbol{W}_t \succ 0$, not full column rank of $\boldsymbol{S}_t$, so it applies to every prefix, unlike the formulas of Appendix A. In panel (a), which has no nuisance block, the test risk of the best span predictor is one minus the squared norm of the projection of $\vec{e}_{11}$ onto the span of $\boldsymbol{\Sigma}\vec{s}_1, \dots, \boldsymbol{\Sigma}\vec{s}_t$. This shows when the pairs used so far make an accurate predictor available.

Related distinctions between second-moment and typical behavior under multiplicative noise arise in the stability of SGD (Chemnitz & Engel, 2025, Appendix A) and in control through actuation channels with random gains, where different moments of stability have different control capacities (Ranade & Sahai, 2019).

F.7 A STABLE RATE CAN BE TOO SLOW FOR ONE PASS

Figure 5 varies the constant rate while keeping the $M = 16$ training data and the horizon n fixed. We use 89 geometrically spaced values of $n\eta\gamma_p$ from 10^{-3} to 2.5, plus B's mean-square-safe rate (Appendix F.6), B's trace threshold $2/(\operatorname{tr}\boldsymbol{\Gamma}_B + 2\gamma_p)$, and A's rate η_M of Corollary 4.

Smaller rates help typical runs in B but do not close the gap to negative ridge and the best span predictor. Among these 92 rates, the median final test risk of SGD in B is never below 0.208, against 0.086 for the illustrative negative-ridge predictor of Appendix F.6 (0.205 for the guarded estimator of Appendix F.2) and 0.031 for the best span predictor on the same data sets. At every tested rate with $n\eta\gamma_p \le 1.5$, B's median final test risk is within 6% of its expected final test risk, and at the larger tested rates both grow. On the paired A data sets, SGD reaches a median of 0.061 at the best tested rate and 0.064 at A's specified rate.

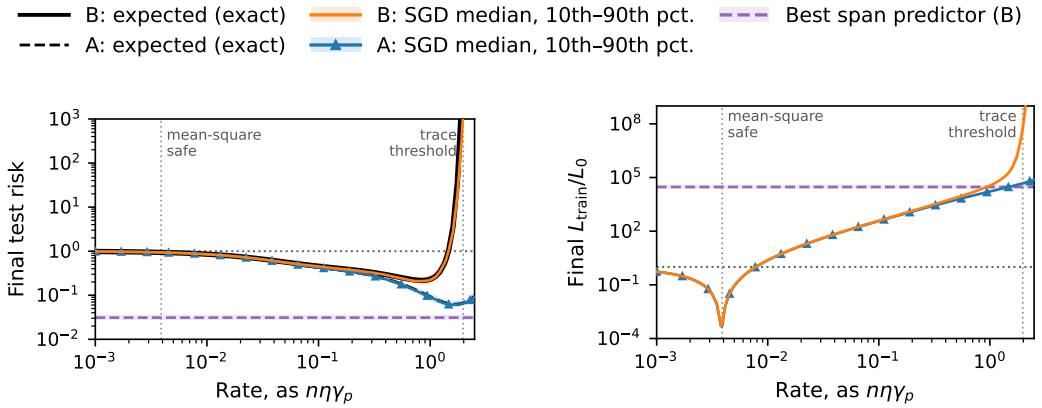
For family B, take the mean-square-safe rate $\eta = 1/(\operatorname{tr}\boldsymbol{\Sigma}_B + 2\|\boldsymbol{\Sigma}_B\|_{\mathrm{op}})$ with $\boldsymbol{\Sigma}_B = \operatorname{diag}(\boldsymbol{\Gamma}_B, \boldsymbol{I}_{d_0})$. By (F.28), this rate guarantees mean-square convergence on an unlimited stream of fresh pairs. At $t = n$, however, the expected test risk is 0.938 and the median is 0.936, while the median normalized training loss is 7.1×10^{-4}. Here $\eta d_0 \approx 1$, so the multiplier $1 - \eta\|\vec{x}_i\|^2$ in (E.5) is near zero and each update nearly fits its own pair, and the observed pass nearly interpolates its training data yet learns little of the label. The contraction factor in (F.28) is $1 - \eta$ because $\lambda_{\min}(\boldsymbol{\Sigma}_B) = 1$, and $n\eta \approx n/d_0 = M^{-5}$, so the one-pass budget is far too short for that guarantee to help. In both families, every tested rate whose median final test risk is below one half ends with median normalized training loss L_{train}/L_0 above 100, like the best span predictor.

At $M = 64$ and $M = 256$, B's median final test risk is within 2% of its expected final test risk at every tested rate with $n\eta\gamma_p \le 1.5$, and at the larger tested rates both grow. With 2,000 runs per rate at 45 geometrically spaced values of $n\eta\gamma_p$ from 0.05 to 2.5, the smallest median final risk is 0.148 at $M = 64$ and 0.117 at $M = 256$, and the exact expected risks at those same rates, 0.149 and 0.118, are also the smallest on the grid. These are finite-grid observations. Theorem 3 bounds only expected risk.

Rotational symmetry within the strong block and within the nuisance block lets an exact simulation of the pass track only three numbers: the weak residual $r_{\text{weak}} = r_{t,M+1}$, the length r_{strong} of the strong-block residual, and the length v of the nuisance part. At each update, draw independent $z, u, \zeta \sim \mathcal{N}(0,1)$, $c_1 \sim \chi^2_{M-1}$, and $c_2 \sim \chi^2_{d_0-1}$. They represent the new pair's weak coordinate, its strong and nuisance components along the current residual directions, and the squared lengths of its remaining strong and nuisance components. With prediction error $e = r_{\text{weak}} z + r_{\text{strong}} u - v\zeta$ on the new pair,

$$\begin{aligned} r^+_{\text{weak}} &= r_{\text{weak}} - \eta\gamma_p e z, \\ r^+_{\text{strong}} &= \sqrt{(r_{\text{strong}} - \eta\gamma_1 e u)^2 + (\eta\gamma_1 e)^2 c_1}, \\ v^+ &= \sqrt{(v + \eta e\zeta)^2 + (\eta e)^2 c_2}, \end{aligned} \tag{F.29}$$

starting from $r_{\text{weak}} = r_{\text{strong}} = 1/\sqrt{2}$ and $v = 0$, with test risk $r^2_{\text{weak}} + r^2_{\text{strong}} + v^2$. The runs use a NumPy `SeedSequence` with entropy $[20260922, M]$ and share the scalar draws across rates. Because the residual-aligned coordinates depend on the rate, this coupling does not run different rates on the same training data, so we report only each rate's marginal risk distribution.



(a) Final test risk

(b) Final training loss on the training data

Figure 5: At B's mean-square-safe rate the pass nearly interpolates yet learns little. Every tested rate with median test risk below one half has median training loss over 100 times the zero predictor's; in B, none has median test risk below 0.2. One SGD pass at $M = 16$ against its constant rate, shown as $n\eta\gamma_p$, for family B (orange) and paired family A (blue) on the 64 paired data sets of Appendix F.5. Panel (b) divides training loss by L_0. Blue and orange curves are medians with 10th–90th percentile bands, and black curves are exact expected test risks (solid for B, dashed for A). The purple line and band give B's best span predictor. The first dotted vertical line marks B's mean-square-safe rate $1/(\operatorname{tr}\boldsymbol{\Sigma}_B + 2\|\boldsymbol{\Sigma}_B\|_{\text{op}})$. By (F.28) it guarantees mean-square convergence on an unlimited stream of fresh pairs. The second marks B's trace threshold $2/(\operatorname{tr}\boldsymbol{\Gamma}_B + 2\gamma_p)$, at and above which the expected squared error V_t in the original signal coefficients cannot decrease by (35). At the largest rates, B's curves continue above the top of each panel. The dotted horizontal line marks the zero predictor's value, one.

F.8 FINITE EXAMPLES AT ORDINARY SCALE

The powers of M in (39) make the asymptotic comparisons simple. This subsection examines three finite problems with explicitly stored inputs, each with target $\vec{\beta} = (1, 1)/\sqrt{2}$ and 32 data sets drawn with seeds 1–32 of NumPy's `default_rng`. The first shows accurate misfitting predictors and one SGD pass reaching them. In the second, the finite converse also certifies that sufficiently accurate span predictors must misfit. The third has a larger strength ratio and shows the common-multiplier obstruction. The finite orderings $\gamma_1 < d_0/n < \gamma_p^2$ quoted below place each problem relative to the window (8), whose separations are asymptotic.

Table 3 reports the first problem, the model (1) at $n = 400$, $d_0 = 40{,}000$, and $\mathbf{\Gamma} = \mathrm{diag}(30, 25)$. Then $\tau = 100$ lies between $\gamma_1 = 30$ and $\gamma_p^2 = 625$, $R_{\mathrm{wk}} = 6.25$, $q_n = 0.0055$, $\tau B(\vec{\beta})^2 = 0.136$, and $\tau^2 B(\vec{\beta})^2 = 13.6$. Each data set stores the $n \times d$ array of input rows, $\boldsymbol{X}^\top$, explicitly, and SGD runs its updates in $\mathbb{R}^d$. Test risk is evaluated by (2) and training loss by (3) on all n pairs, divided by the same data set's L_0.

The rate (D.8), which uses only the strengths and dimensions, is $\eta \approx 0.90/(n\gamma_p)$; it is the one quoted in Section 5. The exact expected risk there is 0.200, and the simulated mean is 0.201. The rate $\eta^\star \approx 1.28/(n\gamma_p)$ minimizes the exact expected risk of (27)–(28) over 25 geometrically spaced values of $n\eta\gamma_p$ in $[0.1, 3]$. This search uses the strengths and the target, though no data, so $\eta^\star$ is a population-tuned comparison rather than a rate Theorem 3 supplies. At this rate $\eta^\star \|\vec{x}_i\|^2 \approx 5$, so each update multiplies the residual on its own pair by about -4, and the exact expected risk 0.162 matches the simulated mean 0.162.

The repeated-pass rows use $\eta_k = 1/\max_i \|\vec{x}_i\|^2$ for 20 reshuffled epochs, so every step satisfies the hypothesis of Proposition E.1. At this size the bound of that proposition is not informative, since $\gamma_1/\tau = 0.3$. After 20 epochs the training loss is below 10^{-30} and the test risk agrees with the interpolator's to the displayed precision, consistent with Appendix E.1, which identifies the limit whenever training loss tends to zero.

Negative ridge uses the illustrative penalty $n\lambda = \frac{5}{2}\sqrt{nd_0} - d_0$ of Appendix F.6. The guarded estimator (F.20), with the thresholds (F.18), subtracts less of d_0 at this size. It roughly halves the zero predictor's risk with training loss 0.2 times the zero predictor's. This does not contradict the converse (10), which is asymptotic, and whose finite form below does not exclude it at this size.

The two strengths differ by only 20%, and the best scalar multiple of the interpolator, whose multiplier (F.10) uses the distribution, is nearly as accurate as the best span predictor, with median test risk 0.12 at training loss 9.3 times the zero predictor's. The instance shows accurate misfitting predictors at ordinary scale, but not the common-multiplier obstruction of Corollary 4, which the last instance below, with strength ratio 4, makes visible.

Table 3: Medians and 10th–90th percentiles over 32 data sets at $n = 400$, $d = 40{,}002$, strengths 30 and 25.

Predictor	Test risk	L_{train}/L_0
Zero predictor	1.00	1.0
Minimum-norm interpolator	0.62 [0.61, 0.64]	0.0
Interpolator, best scalar multiple (uses the distribution)	0.12 [0.11, 0.13]	9.3 [8.6, 10.9]
SGD at rate $1/\max_i \|\vec{x}_i\|^2$, 20 epochs	0.62 [0.61, 0.64]	$< 10^{-30}$
Best span predictor (uses the distribution)	0.12 [0.11, 0.13]	9.4 [8.8, 11.1]
One SGD pass at population-tuned rate $\eta^\star$	0.16 [0.15, 0.18]	8.1 [7.9, 8.5]
One SGD pass at rate (D.8)	0.20 [0.19, 0.22]	4.2 [4.1, 4.3]
Negative ridge, penalty of Appendix F.6	0.23 [0.22, 0.26]	3.1 [2.9, 3.5]
Guarded negative ridge (F.20)	0.48 [0.47, 0.51]	0.2 [0.19, 0.21]

A finite necessity guarantee. The finite converse of Appendix A.1 is loose at the first problem. With $\delta = 0.05$, $g = 1/4$, and the target-dependent bound of (A.5), (A.4) gives only 0.17. The second problem has $n = 200$, $d_0 = 100{,}000$, and $\mathbf{\Gamma} = \mathrm{diag}(50, 40)$, where $\tau = 500$ (Table 4, which also reports the third problem). There the same bound gives 4.9, so with probability at least 0.95 every span predictor for this target with test risk at most $1/4$ has training loss above 4.9 times the zero predictor's. The best span predictor has median test risk 0.199, below that threshold, at 74

times the zero predictor's training loss. One SGD pass at the population-tuned rate $\eta^\star$ has median 0.252 at 64 times, just above $1/4$. To compare with the population-tuned SGD pass, take $g = 0.30$. The target-free first bound in (A.5) then gives 3.5 for every unit target at once, and the SGD median 0.252 and 90th percentile 0.273 both lie below 0.30.

Table 4: Medians and 10th–90th percentiles over 32 data sets for the second problem ($n = 200$, $d = 100{,}002$, strengths 50 and 40) and the third ($n = 200$, $d = 140{,}002$, strengths 200 and 50). For each problem, $\eta^\star$ minimizes that problem's exact expected risk over the grid used for the first problem, giving $n\eta^\star\gamma_p \approx 0.97$ for both problems here.

Predictor	Test risk	L_{train}/L_0
Second problem		
Minimum-norm interpolator	0.842 [0.825, 0.854]	0
Interpolator, best scalar multiple (uses the distribution)	0.201 [0.187, 0.221]	72 [61, 86]
Best span predictor (uses the distribution)	0.199 [0.182, 0.214]	74 [61, 86]
One SGD pass at population-tuned rate $\eta^\star$	0.252 [0.236, 0.273]	64 [59, 68]
One SGD pass at rate (D.8)	0.333 [0.303, 0.367]	24 [22, 25]
Third problem		
Minimum-norm interpolator	0.737 [0.714, 0.757]	0
Interpolator, best scalar multiple (uses the distribution)	0.249 [0.230, 0.284]	17 [14, 23]
Best span predictor (uses the distribution)	0.116 [0.107, 0.128]	61 [52, 72]
One SGD pass at population-tuned rate $\eta^\star$	0.158 [0.148, 0.183]	56 [53, 63]
One SGD pass at rate (D.8)	0.194 [0.176, 0.223]	25 [24, 27]

A larger strength ratio. The third problem has $n = 200$, $d_0 = 140{,}000$, and $\mathbf{\Gamma} = \operatorname{diag}(200, 50)$, so the strength ratio is 4 and $\tau = 700$ lies between $\gamma_1 = 200$ and $\gamma_p^2 = 2500$. There the best scalar multiple of the interpolator is less accurate than the best span predictor, with median test risk 0.25 against 0.12 (Table 4). One SGD pass at the rate (D.8) has median test risk 0.19 at 25 times the zero predictor's training loss, and its 90th percentile, 0.22, lies below the 10th percentile of the rescaled interpolator, 0.23. With $\delta = 0.05$, $g = 1/4$, and the target-dependent bound of (A.5), (A.4) gives 1.56, so with probability at least 0.95 every span predictor for this target with test risk at most $1/4$ has training loss above 1.56 times the zero predictor's.

Norm and training loss together. In all 32 data sets of each of the three problems, the interpolator has smaller norm than the SGD pass at the rate (D.8), zero training loss, and larger test risk. For the problem of Table 3, the two median norms are 0.089 and 0.27. So no bound of the form (G.1) can certify a risk for that pass below the interpolator's test risk, whose median in Table 3 is 0.62. The script `ordinary_scale_check.py` in the supplement produces these numbers, including the finite bounds; it is independent of the Gram-sampling code of Appendix F.5.

G Relations to SGD and interpolation results

This appendix gives context and additional consequences, not steps in the proofs of Theorems 1–3 or Corollary 4. The SGD comparisons here concern noiseless least squares and the final iterate after one use of each independent example. We use the exact Gaussian recursion of Bordelon & Pehlevan (2022, Theorem 1 and eq. (6)), which applies to both families in (39); Appendix C quotes it and translates it into our notation. Weighted-error identities and transient mean-square analysis are classical LMS tools (Sayed, 2008, Chapters 22–23).

The trace lower bound (32) follows by applying the Gaussian recurrence to the coefficient error V_t in (34). The geometric converse instead bounds training loss for every accurate span predictor on one realization of the training data. Bordelon–Pehlevan also analyze training and test losses when observations are reused. Here the distinct conclusion is that the converse (10) holds for every accurate predictor in the span, not only for a particular learning trajectory.

Training fit and generalization bounds. Many explanations of generalization bound test risk by training loss plus a nonnegative gap term,

$$\mathcal{E}(\vec{u};\vec{\beta}) \le L_{\text{train}}(\vec{u};\vec{\beta}) + \Delta_D(\vec{u}), \qquad \Delta_D(\vec{u}) \ge 0.$$

Koehler et al. (2021) use uniform convergence with norm bounds to explain benign overfitting of interpolators, and Nagarajan & Kolter (2019) show that uniform convergence can fail to explain generalization in deep networks. Our converse concerns the training-loss term itself. In the window (8), every span predictor with training loss at most the zero predictor's has test risk above any fixed $g < 1$ on the event of Theorem 1, so for an accurate span predictor the bound's first term is already large. This rules out certifying small risk through that template.

Bounds nondecreasing in norm and training loss. Adding a norm does not help. Let $C_D(r, \ell)$ be nondecreasing in both arguments and valid on the training data D:

$$\mathcal{E}(\vec{v}; \vec{\beta}) \le C_D\big(\|\vec{v}\|, L_{\mathrm{train}}(\vec{v}; \vec{\beta})\big) \qquad \text{for every span predictor } \vec{v}. \tag{G.1}$$

In the window (8), accurate span predictors exceed the minimum-norm interpolator $\vec{u}_{\mathrm{int}} = \boldsymbol{X}\boldsymbol{K}^{-1}\vec{y}$, where $\boldsymbol{K} = \boldsymbol{X}^\top\boldsymbol{X}$, in both training loss and norm, so such a bound cannot certify them below the interpolator's test risk.

For a span predictor $\vec{u} = \boldsymbol{X}\vec{\alpha}$ with training predictions $\vec{z} = \boldsymbol{K}\vec{\alpha}$, we have $\|\vec{u}\|^2 = \vec{z}^\top\boldsymbol{K}^{-1}\vec{z}$, $\|\vec{u}_{\mathrm{int}}\|^2 = \vec{y}^\top\boldsymbol{K}^{-1}\vec{y}$, $nL_{\mathrm{train}}(\vec{u}; \vec{\beta}) = \|\vec{z} - \vec{y}\|^2$, and $\|\vec{y}\|^2 = nL_0(\vec{\beta})$. Bounding the two quadratic forms by the extreme eigenvalues of $\boldsymbol{K}$ and using $\|\vec{z}\| \ge \|\vec{z} - \vec{y}\| - \|\vec{y}\|$ gives

$$\frac{\|\vec{u}\|^2}{\|\vec{u}_{\mathrm{int}}\|^2} \ge \frac{\lambda_{\min}(\boldsymbol{K})}{\lambda_{\max}(\boldsymbol{K})}\left[\sqrt{\frac{L_{\mathrm{train}}(\vec{u}; \vec{\beta})}{L_0(\vec{\beta})}} - 1\right]_+^2. \tag{G.2}$$

On the event (A.13)–(A.14), $\boldsymbol{K}/d_0 \to \boldsymbol{I}_n$ in operator norm, because $\boldsymbol{W}/d_0 \to \boldsymbol{I}_n$ and $\|\boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S}^\top\|_{\mathrm{op}} \le \gamma_1\|\boldsymbol{S}\|_{\mathrm{op}}^2 = O(\gamma_1 n) = o(d_0)$. So the eigenvalue ratio in (G.2) tends to one, every span predictor whose training loss is large compared with $L_0(\vec{\beta})$ has $\|\vec{u}\|^2 = (1 + o(1))L_{\mathrm{train}}(\vec{u}; \vec{\beta})/\tau$, and all span predictors with a common large training loss have the same leading-order norm.

With $\delta = n^{-2}$, on the event (A.2) and the rank event, every span predictor with test risk at most a fixed $g < 1$ has $L_{\mathrm{train}}(\vec{u}; \vec{\beta})/L_0(\vec{\beta}) \ge \widetilde{\ell}_n(g, n^{-2}) = (1 + o(1))(1 - \sqrt{g})^2(\tau/\gamma_1)^2$ by (A.7). By (G.2), its squared norm is then at least $(1 + o(1))(1 - \sqrt{g})^2(\tau/\gamma_1)^2$ times the interpolator's. This norm ratio diverges because $\gamma_1 \ll \tau$. On the same event the interpolator has test risk above g, because its zero training loss is below that bound. Once the norm ratio exceeds one, every span predictor $\vec{u}$ with test risk at most g therefore satisfies

$$C_D\big(\|\vec{u}\|, L_{\mathrm{train}}(\vec{u}; \vec{\beta})\big) \ge C_D(\|\vec{u}_{\mathrm{int}}\|, 0) \ge \mathcal{E}(\vec{u}_{\mathrm{int}}; \vec{\beta}) > g, \tag{G.3}$$

by monotonicity and then by (G.1) at the interpolator. If (G.1) holds only with high probability, the comparison holds on the intersection of its event with those above. Appendix F.8 shows the same ordering at ordinary scale.

Noiseless and last-iterate guarantees. The rate η_M with which family A learns eventually exceeds the rates allowed by the upper bounds discussed next, as (G.4) shows. Varre et al. (2021, Assumptions 1–2 and Theorem 1) study constant-step last-iterate convergence under realizability from zero initialization. Their specified rate is $(4R_4 \log n)^{-1}$ for $n \ge 2$, where $\mathbb{E}[\|\vec{x}\|^2\vec{x}\vec{x}^\top] \preceq R_4\boldsymbol{\Sigma}$. For Gaussian features, $\mathbb{E}[\|\vec{x}\|^2\vec{x}\vec{x}^\top] = (\operatorname{tr}\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^2$, so $R_4 = \operatorname{tr}\boldsymbol{\Sigma} + 2\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$ is admissible. Their interpolation assumption is that a common noiseless solution exists. Our model satisfies it even when a finite trained iterate does not interpolate its training data.

Averages and prescribed decaying schedules. The upper bound in Zou et al. (2023, Theorem 2.1) concerns averaged constant-step SGD and uses $\eta < 1/(\alpha_4 \operatorname{tr}\boldsymbol{\Sigma})$, where α_4 is their fourth-moment constant. Their separate Theorem 2.2 permits $\eta < 1/\lambda_1$ for an averaged-iterate lower bound, where $\lambda_1 = \|\boldsymbol{\Sigma}\|_{\mathrm{op}} = \gamma_1$ is the largest eigenvalue of the full feature covariance, with its other assumptions and $n \ge 500$. Their lower-bound theorem therefore allows a larger rate range than their upper-bound theorem.

Wu et al. (2022a, Theorems 4.1–4.2) study the last iterate under a prescribed constant-prefix and geometrically decaying schedule. Their upper-bound theorem requires initial rate $\eta_0 < [3\alpha_4 \operatorname{tr}\boldsymbol{\Sigma} \log(s + K)]^{-1}$ in their schedule notation. Their lower-bound theorem permits $\eta_0 < 1/\lambda_1$

and $K \geq 10$. It does not quantify over every predictable scalar schedule. The more recent comparison of Wu et al. (2026, Proposition 2.2 of arXiv:2509.17251v2) uses a specified decaying last iterate with initial rate at most $1/(4 \operatorname{tr} \boldsymbol{\Sigma})$, $n \geq 100$, and its stated moment and noise assumptions.

For family A, the rate in Corollary 4 satisfies

$$\eta_M \operatorname{tr} \boldsymbol{\Sigma}_A = \frac{1}{2} M^2 \log M + \frac{\log M}{M} \to \infty. \tag{G.4}$$

Thus A's rate eventually exceeds the rates allowed by the cited upper bounds. The exact Gaussian recursion of Bordelon & Pehlevan (2022) remains applicable. Exceeding the inverse-trace rate $1/\operatorname{tr} \boldsymbol{\Sigma}$ is not specific to multiple spikes. For every constant-rate pass achieving a fixed expected improvement in the geometric window, including $p = 1$, (E.4) gives $\eta_n \operatorname{tr} \boldsymbol{\Sigma} \geq \eta_n d_0 \to \infty$. What multiple spikes add is the signal-trace obstruction (32).

The bi-level ensemble. The bi-level model of Muthukumar et al. (2021, Definition 5) has one common variance in its preferred directions and a smaller common variance in the remaining directions. For a fixed ambient dimension, the total variance is fixed. In their notation the number of preferred directions is n^r, with $r < 1$.

After one common rescaling makes the smaller variance one, the real-label model has our form with $\boldsymbol{\Gamma} = \gamma \boldsymbol{I}_p$, $p/n \to 0$, and $n/d_0 \to 0$. Here p counts the preferred directions. In these units, with $\tau = d_0/n$ from (7), the three strict regimes of their Theorem 13 are $\tau \ll \gamma$, $\gamma \ll \tau \ll \gamma^2$, and $\gamma^2 \ll \tau$. In the first, minimum-norm interpolation succeeds for both real-label regression and binary-label classification. In the second, it fails for regression but succeeds for classification. This is our misfitting window (8). Theorem 1 shows that accurate span prediction is possible there, but it must sacrifice training fit.

All span predictors in the third regime. Our span formulas also explain the third regime, where their interpolating classifier fails. The lower bound $\boldsymbol{M}_{\mathrm{resp}} \succeq \tau \boldsymbol{\Gamma}^{-2}/(1 + e_n)$ in (A.15), on the response matrix of (A.8), was obtained using only $p/n \to 0$ and $n/d_0 \to 0$. With $\boldsymbol{\Gamma} = \gamma \boldsymbol{I}_p$ and $\gamma^2 \ll \tau$, this lower bound makes every eigenvalue of $\boldsymbol{M}_{\mathrm{resp}}$ diverge. Substituting into (A.12) shows that the best-span regression risk tends to one.

For classification, rotate within the equal-strength signal subspace so that the target is $\operatorname{sign}(s[1])$, where $s[1]$ is the first coordinate of the fresh test input's standardized signal vector $\vec{s}$. For a nonzero span predictor $\boldsymbol{X}\vec{\alpha}$, classify by the sign of its output $(\boldsymbol{X}\vec{\alpha})^\top \vec{x}$. With signal response $\vec{a} = \gamma \boldsymbol{S}^\top \vec{\alpha}$, its survival and contamination are $\mathrm{SU} = a_1$ and $\mathrm{CN}^2 = \sum_{j=2}^{p} a_j^2 + \vec{\alpha}^\top \boldsymbol{W} \vec{\alpha}$. Survival is the coefficient of the label-bearing coordinate $s[1]$ in the prediction, and contamination is the standard deviation of the remaining prediction terms, which are independent of $s[1]$. Because $\boldsymbol{W} \succ 0$, the contamination is positive. Using $\boldsymbol{Q}$ from (A.8), Cauchy–Schwarz gives

$$\frac{|\mathrm{SU}|}{\mathrm{CN}} \leq \frac{\gamma |\vec{e}_1^\top \boldsymbol{S}^\top \vec{\alpha}|}{\sqrt{\vec{\alpha}^\top \boldsymbol{W} \vec{\alpha}}} \leq \gamma \sqrt{\vec{e}_1^\top \boldsymbol{Q} \vec{e}_1} \leq (1 + o(1)) \frac{\gamma}{\sqrt{\tau}} \longrightarrow 0. \tag{G.5}$$

The first inequality in (G.5) drops $\sum_{j \geq 2} a_j^2$ from CN^2. The second applies Cauchy–Schwarz to $(\boldsymbol{W}^{-1/2} \boldsymbol{S} \vec{e}_1)^\top (\boldsymbol{W}^{1/2} \vec{\alpha})$. The third uses $\|\tau \boldsymbol{Q} - \boldsymbol{I}_p\|_{\mathrm{op}} \to 0$ from Appendix A.3. The final bound holds on one event of probability tending to one and is independent of the predictor. By Muthukumar et al. (2021, Proposition 17), its classification error is $1/2 - \pi^{-1} \arctan(\mathrm{SU}/\mathrm{CN})$, which therefore tends to $1/2$ uniformly over nonzero span predictors. Assigning either fixed class to the zero score gives error $1/2$. These all-span conclusions follow from the geometry of the training inputs regardless of whether the fitting labels are real or binary. They are not claims attributed to the interpolation theorem.

The trace ratio behaves differently. With equal signal strengths it is $q_n = p/n \to 0$ in all three regimes, so the obstruction in family B does not arise. Increasing the number of preferred directions in their model also changes the preferred variance. Our pair instead holds both label-bearing strengths and every dimension fixed between A and B.

Where the assumptions are used. The converse of Theorem 1 compares the nuisance prediction $\boldsymbol{W}\vec{\alpha}$ on training inputs in (6) with the nuisance variance $\vec{\alpha}^\top \boldsymbol{W} \vec{\alpha}$ on fresh test inputs in (5). Within

the window (8), it needs only that $\boldsymbol{W}/d_0$ and $\boldsymbol{S}^\top\boldsymbol{S}/n$ concentrate. The projection result of Theorem 2 compares a frame $\boldsymbol{U}$ with its mirror $\boldsymbol{J}_X\boldsymbol{U}$, and the isotropic nuisance block makes the mirror another frame of the same model, with conjugate covariance and equal likelihood at the observed sample. Gaussian sampling supplies the concentration events, the likelihood formula, and the exact recursion (27)–(28). The noiseless linear label means that, at each fixed frame, one linear predictor gives the label exactly on every input, which makes both loss formulas exact. The unit target and the ordering of strengths give simple constants.

What changes with several spikes. In the geometric window, the leading best-span risk is $\tau\|\boldsymbol{\Gamma}^{-1}\vec{\beta}\|^2$. The SGD trace ratio also counts directions irrelevant to the label. The pair isolates this difference by changing only the strengths of those directions. Its leading best-span risk and training loss remain unchanged, but constant-rate SGD achieves vanishing expected risk in A and not in B.

G.1 Training on one family and testing on the other

Could B's error floor arise only because its test covariance penalizes coefficient errors more heavily? To answer this, we evaluate each family's trained pass under both input laws, without changing the trained predictor. This is the only comparison in the paper in which the training and test distributions differ.

The two families in (39) share their labeling rule. Because $\boldsymbol{\Gamma}_A$ and $\boldsymbol{\Gamma}_B$ agree in coordinates 1 and $M+1$, where $\vec{\beta}$ is supported, the true predictor $\vec{u}_{\vec{\beta}} = (\boldsymbol{\Gamma}^{-1/2}\vec{\beta}, \vec{0})$ of Section 2 is the same for both, and only the input covariance differs. Write $\mathcal{E}_A$ and $\mathcal{E}_B$ for the test risk (2) of this target under the two input laws. Coordinates $2, \dots, M$ have $\beta_j = 0$, so for every fixed predictor $\vec{u}$,

$$\mathcal{E}_B(\vec{u}) = \mathcal{E}_A(\vec{u}) + (M^4 - M^3)\sum_{j=2}^{M}(u_j^{\mathrm{sig}})^2. \tag{G.6}$$

The comparison is thus a covariate shift with a fixed labeling rule, a setting studied for SGD by Wu et al. (2022b) and for kernel regression by Canatar et al. (2021). By (G.6), $\mathcal{E}_B \geq \mathcal{E}_A$, but this ordering alone gives neither conclusion below. Under B the weights on coordinates $2, \dots, M$ count M times as much, and a floor under B need not persist under A. In $\mathbb{E}\mathcal{E}_B(\vec{w}_n^A)$, the superscript A names the training law over which the outer expectation is taken, and the subscript B names the law of the fresh test pair.

Proposition G.1 (Cross-evaluation of the pair)**.** *Let $\vec{w}_n^A$ and $\vec{w}_n^B$ be the final iterates of one pass* (25) *on training data from A and from B. At the rate η_M of Corollary 4, $\mathbb{E}\mathcal{E}_B(\vec{w}_n^A) = O(M^{-1/2}\log^2 M) \to 0$. For every deterministic positive constant rate,*

$$\mathbb{E}\mathcal{E}_A(\vec{w}_n^B) \geq \frac{1}{2}\Big(\frac{M^2+1}{M^2+3}\Big)^{2M^2} \longrightarrow \frac{e^{-4}}{2}. \tag{G.7}$$

Proof. For the first claim, the label-irrelevant coordinates need separate control, because B multiplies their contribution to test risk by M. All recurrence quantities in this paragraph refer to the pass on A. Write $b = \eta_M M^3 = \log M/(2M^2)$ and $\rho = 1-2b+2b^2 \leq e^{-b}$, the signal multiplier of coordinates $2, \dots, M+1$ in A. The quantities ϕ and ψ_n of Lemma D.1 are $\phi = \eta_M \operatorname{tr}\boldsymbol{\Gamma}_A = (\log M)/M$ and $\psi_n = (1+o(1))(\log^2 M)/(4M)$. For all sufficiently large M the lemma's hypotheses hold, and its proof gives $T_{\max} \leq 2$ and $N_{\max} \leq 2\psi_n$. Summing (D.6) over j at time t and adding $N_t \leq 2\psi_n$ gives, for $0 \leq t \leq n$,

$$R_t = \mathbb{E}\mathcal{E}_A(\vec{w}_t^A) \leq e^{-bt} + \varepsilon_M, \qquad \varepsilon_M = 2\phi(1+\psi_n) + 2\psi_n = O(M^{-1}\log^2 M). \tag{G.8}$$

For $2 \leq j \leq M$, the recurrence (27) starts from $T_{0,j} = \beta_j^2 = 0$, so

$$T_{n,j} = b^2\sum_{t=0}^{n-1}\rho^{n-1-t}R_t \leq b^2 n e^{-b(n-1)} + b\,\varepsilon_M, \tag{G.9}$$

using $\rho^{n-1-t}e^{-bt} \le e^{-b(n-1)}$ and $\sum_{k\ge 0}\rho^k = 1/(2b-2b^2) \le 1/b$. Since $T_{n,j} = M^3\mathbb{E}(u_j^{\mathrm{sig}})^2$ for $\vec{u} = \vec{w}_n^A$ at these coordinates, (G.6) gives

$$\mathbb{E}\mathcal{E}_B(\vec{w}_n^A) = R_n + (M-1)\sum_{j=2}^{M} T_{n,j} \le e^{-bn} + \varepsilon_M + (M-1)^2\big[b^2 n e^{-b(n-1)} + b\,\varepsilon_M\big]. \quad \text{(G.10)}$$

The first factor $M-1$ is the change M^4/M^3-1 in the test weight, and the second counts coordinates $2,\dots,M$. With $n = M^2$ and $bn = \frac{1}{2}\log M$, the right side of (G.10) is $O(M^{-1/2}\log^2 M + M^{-1}\log^3 M) = O(M^{-1/2}\log^2 M)$.

For the second claim, let $\eta > 0$ be the rate of the pass on B. Suppose first that $\eta < 2/(\operatorname{tr}\mathbf{\Gamma}_B + 2\gamma_p) = 2/(M^5+3M^3)$, so that $0 < \eta M^3 < 2/(M^2+3)$. The weak coordinate $M+1$ has strength M^3 in both families, and by (29) its mean residual after the pass on B is $(1-\eta M^3)^n/\sqrt{2}$. Its squared residual is part of $\mathcal{E}_A(\vec{w}_n^B)$, so Jensen's inequality gives

$$\mathbb{E}\mathcal{E}_A(\vec{w}_n^B) \ge \frac{1}{2}(1-\eta M^3)^{2n} \ge \frac{1}{2}\Big(\frac{M^2+1}{M^2+3}\Big)^{2M^2}.$$

Otherwise, (35) makes B's coefficient error $V_t = \mathbb{E}\|\mathbf{\Gamma}_B^{-1/2}\vec{\beta} - \vec{w}_t^{\mathrm{sig}}\|^2$ nondecreasing. Every strength in $\mathbf{\Gamma}_A$ is at least M^3, and $\mathbf{\Gamma}_A^{-1/2}\vec{\beta} = \mathbf{\Gamma}_B^{-1/2}\vec{\beta}$, so

$$\mathbb{E}\mathcal{E}_A(\vec{w}_n^B) \ge M^3 V_n \ge M^3 V_0 = \frac{1}{2} + \frac{1}{2M},$$

which exceeds the right side of (G.7). □

The supplement's script `cross_eval_check.py` evaluates the expected risks of both passes under both input laws from the exact recurrences (27)–(28).

The identity (G.6) separates two effects of the label-irrelevant strengths. At test time they penalize the weights u_j^{sig} with $2 \le j \le M$, and training on A keeps those weights small enough that the penalty vanishes even under B. During training they enter the trace obstruction of Theorem 3, and (G.7) shows that testing on the more favorable law A does not remove it. In particular, for test inputs from B, the pass trained on A eventually has smaller expected risk than every constant-rate pass trained on B, by (42). In this pair, a constant-rate pass can learn when its training inputs have weaker label-irrelevant directions, whichever of the two laws it is tested on.

H An additional bound for scalar steps chosen from the past

For comparison with constant rates, this appendix gives a lower bound for rates chosen using earlier examples. The bound is not used in the constant-rate characterization and does not extend the A/B separation to predictable schedules. Let $\mathcal{F}_t$ contain the first t training pairs and the independent learner randomness used before update $t+1$. A *predictable* rate η_t is a finite real number chosen using only $\mathcal{F}_t$. The next pair is independent of $\mathcal{F}_t$. The learner starts at $\vec{w}_0 = \vec{0}$, applies (25) with rate η_t, and returns the final iterate without averaging.

Before the training data are drawn, fix a nonempty set J of signal coordinates, called a signal block. Because $\mathbf{\Gamma}$ is diagonal, the covariance maps the span of these coordinates into itself. Let $\vec{\beta}_J$ be the target restricted to J, let $\mathbf{\Lambda} = \mathbf{\Gamma}_J$ be its covariance block, and put $s_J = \operatorname{tr}(\mathbf{\Lambda}^2)/\|\mathbf{\Lambda}\|_{\mathrm{op}}^2 = \sum_{j\in J}(\gamma_j/\max_{i\in J}\gamma_i)^2$, the effective number of directions in J, which satisfies $1 \le s_J \le |J|$. Then every such schedule obeys

$$\mathbb{E}\mathcal{E}(\vec{w}_n;\vec{\beta}) \ge \|\vec{\beta}_J\|^2\left(1 - \frac{1}{s_J+2}\right)^n. \quad \text{(H.1)}$$

If $n \ll s_J$, the bound shows that the expected risk is at least $(1-o(1))\|\vec{\beta}_J\|^2$, nearly the block's whole initial contribution. When all strengths in J are equal, $s_J = |J|$. The resulting factor $1 - 1/(|J|+2)$ already appears in the Gaussian LMS learning curves of Werfel et al. (2004) and

Bordelon & Pehlevan (2022, eq. (7)). The extension here is the conditional argument for predictable steps and its application to a signal block used by the target.

The expectation in (H.1) may be infinite. Conditioning on $\mathcal{F}_t$ fixes the finite rate η_t and the iterate $\vec{w}_t$, and the next iterate is a polynomial in a fresh Gaussian vector with a finite conditional second moment. So the proof bounds how much even the best real rate could reduce the block error in one update, and then iterates. To prove (H.1), use the full covariance $\mathbf{\Sigma} = \mathrm{diag}(\mathbf{\Gamma}, \boldsymbol{I}_{d_0})$ and zero-risk predictor $\vec{w}_\star$ from Appendix C. Write the standardized coefficient error as $\vec{\epsilon}_t = \mathbf{\Sigma}^{1/2}(\vec{w}_\star - \vec{w}_t)$. In the notation of Section 4, $\vec{\epsilon}_t = (\vec{r}_t, -\vec{v}_t)$. The iterate has test risk $\mathcal{E}(\vec{w}_t; \vec{\beta}) = \|\vec{\epsilon}_t\|^2$, and the error evolves by $\vec{\epsilon}_{t+1} = (\boldsymbol{I} - \eta_t \mathbf{\Sigma}\vec{\zeta}\vec{\zeta}^\top)\vec{\epsilon}_t$ where $\vec{\zeta} \sim \mathcal{N}(\vec{0}, \boldsymbol{I}_d)$ represents the fresh input $\vec{x} = \mathbf{\Sigma}^{1/2}\vec{\zeta}$. Let $\vec{b}_t$ be the restriction of $\vec{\epsilon}_t$ to J. Condition on $\mathcal{F}_t$, which fixes $\vec{b} = \vec{b}_t$ and the error $\vec{c}$ in all other coordinates, signal and nuisance. Write $\vec{\zeta}_J$ and $\vec{\zeta}_{J^c}$ for the coordinates of $\vec{\zeta}$ in J and outside J. Because $\mathbf{\Sigma}$ is diagonal, the error update restricted to J is

$$\vec{b}_{t+1} = \vec{b} - \eta_t \mathbf{\Lambda}\vec{\zeta}_J\big(\vec{\zeta}_J^\top \vec{b} + \vec{\zeta}_{J^c}^\top \vec{c}\big).$$

The fresh Gaussian coordinates in J and outside J are independent and centered, so the cross terms between the parts driven by $\vec{b}$ and by $\vec{c}$ have zero conditional expectation. In the expansion of $\|\vec{b}_{t+1}\|^2$, the term linear in η_t uses $\mathbb{E}[\vec{\zeta}_J\vec{\zeta}_J^\top] = \boldsymbol{I}$, and the terms quadratic in η_t use the Gaussian fourth moments

$$\mathbb{E}\big[\|\mathbf{\Lambda}\vec{\zeta}_J\|^2(\vec{\zeta}_J^\top\vec{b})^2 \mid \mathcal{F}_t\big] = \mathrm{tr}(\mathbf{\Lambda}^2)\|\vec{b}\|^2 + 2\vec{b}^\top\mathbf{\Lambda}^2\vec{b},$$
$$\mathbb{E}\big[\|\mathbf{\Lambda}\vec{\zeta}_J\|^2(\vec{\zeta}_{J^c}^\top\vec{c})^2 \mid \mathcal{F}_t\big] = \mathrm{tr}(\mathbf{\Lambda}^2)\|\vec{c}\|^2.$$

Together these give

$$\begin{aligned}\mathbb{E}[\|\vec{b}_{t+1}\|^2 \mid \mathcal{F}_t] = \|\vec{b}\|^2 &- 2\eta_t\vec{b}^\top\mathbf{\Lambda}\vec{b} \\ &+ \eta_t^2[\mathrm{tr}(\mathbf{\Lambda}^2)\|\vec{b}\|^2 + 2\vec{b}^\top\mathbf{\Lambda}^2\vec{b} + \mathrm{tr}(\mathbf{\Lambda}^2)\|\vec{c}\|^2].\end{aligned} \tag{H.2}$$

For $\vec{b} \neq 0$, discard the last nonnegative term and minimize the remaining quadratic in (H.2) over all real steps. The decrease is at most

$$\frac{(\vec{b}^\top\mathbf{\Lambda}\vec{b})^2}{\mathrm{tr}(\mathbf{\Lambda}^2)\|\vec{b}\|^2 + 2\vec{b}^\top\mathbf{\Lambda}^2\vec{b}} \le \frac{\|\vec{b}\|^2}{s_J + 2}. \tag{H.3}$$

For (H.3), apply Cauchy–Schwarz to the numerator. Write $x = \|\mathbf{\Lambda}\vec{b}\|^2/\|\vec{b}\|^2 \le \|\mathbf{\Lambda}\|_{\mathrm{op}}^2$. The resulting bound is $\|\vec{b}\|^2 x/[\mathrm{tr}(\mathbf{\Lambda}^2) + 2x]$. This expression increases with $x \ge 0$. Evaluating it at $x = \|\mathbf{\Lambda}\|_{\mathrm{op}}^2$ gives the stated bound. The zero-block case follows by nonnegativity. In every case, (H.2) and (H.3) give

$$\mathbb{E}[\|\vec{b}_{t+1}\|^2 \mid \mathcal{F}_t] \ge \Big(1 - \frac{1}{s_J + 2}\Big)\|\vec{b}_t\|^2. \tag{H.4}$$

Iterating (H.4) with the tower property for nonnegative random variables, from $\vec{b}_0 = \vec{\beta}_J$, gives $\mathbb{E}\|\vec{b}_n\|^2 \ge \|\vec{\beta}_J\|^2\big(1 - \frac{1}{s_J+2}\big)^n$. The test risk $\|\vec{\epsilon}_n\|^2$ is at least its block part $\|\vec{b}_n\|^2$, which proves (H.1).

The rate is chosen before seeing the current pair, so the result does not cover division by that pair's observed squared norm. Such steps are covered separately by Proposition E.1, which allows every choice with $0 \le \eta_k\|\vec{x}_{i_k}\|^2 \le 2$ and rules out learning in the geometric window even with repeated passes.

For the families in (39), every block has $s_J \le |J| \le M+1$ while $n = M^2$, so the right side of (H.1) is at most $(1 - \frac{1}{M+3})^{M^2} \to 0$. The bound therefore does not extend the constant-rate separation of Corollary 4 to predictable schedules.

The proof conditions one update at a time. Substituting an entire adaptive rate sequence into a fixed matrix-power formula would not be valid, because conditioning on that sequence can change the independence of the incoming pairs.