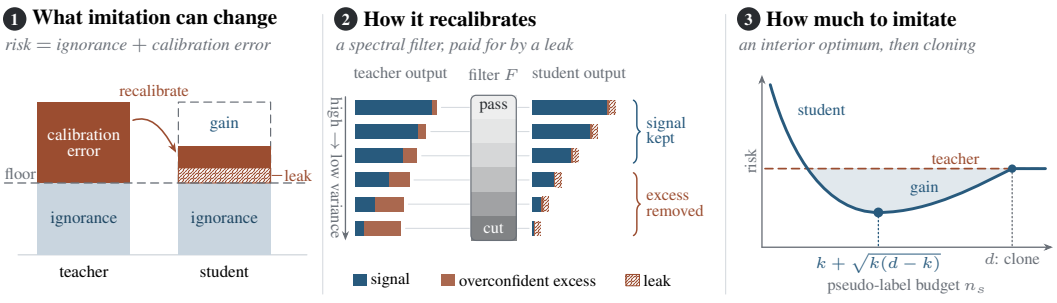
THE LIMITS OF IMITATION: A THEORY OF RECALIBRATION IN WEAK-TO-STRONG GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A student trained only on a teacher's predictions can become more accurate than the teacher, although imitation adds no information. Relative to a prior on the target, every predictor's risk splits into *ignorance*, the risk of the best readout of what it reveals, and *calibration error*, its distance from that readout. Imitation on fresh inputs cannot reduce ignorance, whatever the downstream algorithm; this one bound contains both the teacher's unobserved signal and the Bayes limit of its labels. A student with no knowledge of its own can gain only by recalibrating its teacher. In a linear Gaussian model with a sketched teacher, fitting pseudo-labels applies a spectral filter to the teacher and pays a *leak*: the energy the filter removes returns as noise. In this calculus, a minimum-norm teacher is overconfident exactly in the observed directions whose signal variance falls below one threshold, and imitating slightly less than a clone helps, to first order, when that overconfidence exceeds an explicit leak rate. With k high-variance signal directions in d dimensions, the limiting optimal budget $k + \sqrt{k(d-k)}$ of a minimum-norm student equalizes the fraction of the teacher's error it keeps and the fraction of the removed error it leaks. Along a chain of identical stages, energy and knowledge decay on two clocks, with an underconfident phase between them; at equal total budget, two stages can beat every single-stage ridge. In power-law formulas, interpolating pipelines acquire knowledge at the Bayes exponent, yet their risk plateaus at pure calibration error.

> *"When you know a thing, to hold that you know it; and when you do not know a thing, to allow that you do not know it;—this is knowledge."*
> —Confucius, *Analects* II.17, trans. James Legge (1861)



Overview (schematic). Imitation cannot lower what the teacher does not know; it recalibrates what the teacher says, cutting overconfident spectral directions at the price of a leak.

1 INTRODUCTION

Perfect imitation preserves every mistake. Yet a student trained only on a teacher's predictions can become more accurate than the teacher itself (Burns et al., 2023). If the student sees nothing but the teacher's outputs on fresh inputs, it cannot know more than its teacher. How can it be more accurate?

It can say better what the teacher knows. The squared risk of any predictor splits into two parts: its *ignorance*, the risk of the best readout of what the predictor reveals about the target, and its

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

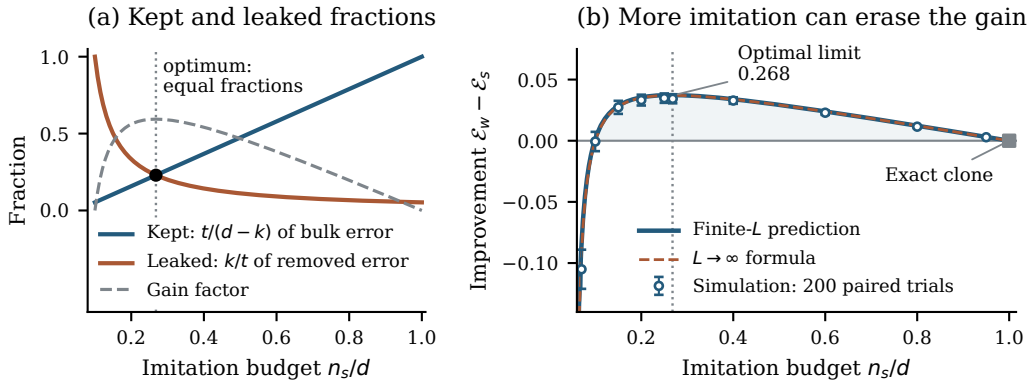


Figure 1: **Kept and leaked fractions balance.** An aligned two-level spectrum: $d = 600$ dimensions, $k = 30$ spikes of height $L = 10000$, a teacher observing $m = 180$ features and fitting $n_{\mathrm{w}} = 60$ true labels, noise variance $\sigma^2 = 0.25$, and signal power $r^2 = 1$. Left: with $t = n_{\mathrm{s}} - k$ equations in the bulk, the student keeps a fraction $t/(d-k)$ of the teacher's bulk error and returns a fraction k/t of the removed error to the spike directions (Theorem 4.2); the gain factor peaks where the two are equal. Right: teacher-minus-student risk, with finite-L and large-L predictions; dots are means over 200 paired estimator trials with bars of two standard errors. The finite-L optimizer is $0.2657d$; the dashed line marks the large-L optimizer $\theta + \sqrt{\theta(1-\theta)}$, $\theta = k/d$. The value at $n_{\mathrm{s}} = d$ is exact cloning. Error bars describe Monte Carlo uncertainty, not finite-size approximation error.

calibration error, the distance from the predictor to that readout. A teacher that reports twice its best estimate knows exactly as much as one that reports the estimate itself; only the second is calibrated. Imitation on fresh inputs cannot reduce ignorance: this is data processing, and it holds for every downstream algorithm (Theorem 2.1). A student can gain only by reducing calibration error, and never by more than the teacher's calibration error. When the student brings no knowledge of its own, weak-to-strong generalization is recalibration.

The student has no labels, so it does not know which recalibration is right. Something else must supply it. Consider a teacher that sees only the leading coordinates of a Gaussian input, and a full-dimensional student that interpolates the teacher's predictions with minimum Euclidean norm. Given the teacher parameter β_{w} and fresh inputs $\widetilde{X}$,

$$\widehat{w}_{\mathrm{s}} = \widetilde{P}\beta_{\mathrm{w}}, \qquad \widetilde{P} = \widetilde{X}^{\dagger}\widetilde{X}. \tag{1.1}$$

Imitation is a random projection. Its mean is a spectral filter that shrinks low-variance directions most; wherever the teacher is overconfident in those directions, the filter recalibrates it. The randomness has a price: the energy the filter removes returns as noise, a *leak* that we compute in closed form. A student beats its teacher when recalibration exceeds leak.

This ledger—ignorance that imitation cannot touch, calibration error that it can reduce, and a leak that it must pay—organizes the paper. A teacher that is nowhere overconfident cannot be improved by any minimum-norm or ridge student, or chain of them, and in the risk calculus a minimum-norm teacher is overconfident exactly in the observed directions whose signal variance lies below one threshold (Sections 2–3). Under isotropic inputs, minimum-norm imitation leaves the aggregate calibration slope unchanged. To first order, imitating slightly less than a clone helps when the teacher's overconfidence exceeds an explicit leak rate and hurts when it falls short; with k spikes in d dimensions, the limiting best budget $k + \sqrt{k(d-k)}$ of a minimum-norm student is where the fraction of the teacher's error it keeps equals the fraction of the removed error it leaks (Section 4, Figure 1). Along a chain of identical stages, energy and knowledge decay on two clocks, and between them the chain is underconfident; at equal total budget, a two-stage chain can beat every single-stage ridge (Section 5). True labels buy knowledge and pseudo-labels buy calibration: in the power-law formulas, interpolating pipelines acquire knowledge at the Bayes exponent while their risk stays at a floor made entirely of calibration error (Section 6). The ledger is exact and general; its spectral terms are computed in a linear Gaussian model with fixed representations and fresh Gaussian batches, at estimator level under bounded spectra and in the deterministic formulas for power laws.

Relation to existing theory. For a convex hypothesis class containing the target, weak-to-strong gain under squared loss is at least the student–teacher misfit (Charikar et al., 2024), with Bregman extensions (Mulgund & Pabbaraju, 2025); Yao et al. (2025) bound it above by the root misfit and relate student and teacher calibration errors conditioned on predicted probabilities. Our bound (2.2) comes instead from data processing under a prior; the ideal student, the readout, attains it, with misfit equal to the teacher's calibration error. Other explanations invoke pseudo-label correction (Lang et al., 2024), benign overfitting (Wu & Sahai, 2025), or representations (Dong et al., 2025; Moniri & Hassani, 2025; Oh et al., 2025). Closest to our setting, Ildiz et al. (2025) derive high-dimensional risks for surrogate-supervised regression with spectral shrinkage, and Wu et al. (2026b) analyze repeated fitting on fresh inputs with noiseless pseudo-labels, deriving risk recursions, a denoising–forgetting tradeoff, and a stopping rule; spectral denoising and recursive risk formulas are thus established, and the ledger separates their denoising, which reduces calibration error, from forgetting, which runs on two clocks. Wu et al. (2026a) already establish optimal exponent preservation in the teacher's label budget for random features (see also Medvedev et al., 2025; Wu et al., 2026c). Reusing designs or mixing original labels (Lecoiu et al., 2026; Dang et al., 2026; Garg et al., 2026) changes the available information. The calibration–refinement split of squared error is classical in forecasting (Murphy, 1973; DeGroot & Fienberg, 1983).

2 Knowledge and Calibration

Notation. We write $\langle u, v\rangle = u^\top v$, $\|u\|$ for the Euclidean norm, and $\|u\|_A^2 = u^\top A u$ for $A \succeq 0$. $A^\dagger$ is the Moore–Penrose pseudoinverse, I_p the $p \times p$ identity, $\mathrm{diag}(u)$ the diagonal matrix with entries u_i, and $\odot$ the entrywise product. Unless conditioning is explicit, $\mathbb{E}$ averages over all randomness. Subscripts w and s denote teacher and student; $\mathcal{E}_{\mathrm{w}} = \mathcal{E}(\beta_{\mathrm{w}})$ and $\mathcal{E}_{\mathrm{s}} = \mathcal{E}(\widehat{w}_{\mathrm{s}})$ are their excess risks. The dimensions are $m < d$, and n_{w} and n_{s} count true and pseudo-labels. We write $a_n \asymp b_n$ when the ratio is bounded above and below by positive constants.

Inputs are $x \sim \mathcal{N}(0, \Sigma)$ in $\mathbb{R}^d$ with $\Sigma = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, $\lambda_1 \geq \cdots \geq \lambda_d > 0$, and labels are $y = x^\top w^* + \xi$ with independent $\xi \sim \mathcal{N}(0, \sigma^2)$. The signal coordinates are independent and centered, with $\mathbb{E}(w_i^*)^2 = \omega_i^2$ and signal power $r^2 = \sum_i \lambda_i \omega_i^2$. The excess risk is $\mathcal{E}(w) = \|w - w^*\|_\Sigma^2$. The teacher observes the first m coordinates, denoted H, of an $n_{\mathrm{w}} \times d$ design X with labels y, $n_{\mathrm{w}} < m$, and interpolates with minimum norm, $\beta_{\mathrm{w}} = (X_H^\dagger y, 0_{d-m})$ (ridge where stated); with noisy labels we take $m \geq n_{\mathrm{w}} + 2$, so that its output has finite second moments. The student fits the pseudo-labels $\widetilde{X}\beta_{\mathrm{w}}$ of n_{s} fresh inputs, which gives (1.1); a ridge student uses $\widehat{w}_{\mathrm{s}}(\eta) = (\widehat{\Sigma} + \eta I)^{-1}\widehat{\Sigma}\beta_{\mathrm{w}}$ with $\widehat{\Sigma} = \widetilde{X}^\top \widetilde{X}/n_{\mathrm{s}}$. The supervised reference applies the student's rule to n_{s} noisy true labels.

The ledger. For a predictor β, its readout $\bar{\beta} = \mathbb{E}[w^* \mid \beta]$ is the best prediction of the target from β. Since $\mathbb{E}[y \mid x, \beta] = x^\top \bar{\beta}$ at a fresh input, β is calibrated exactly when $\beta = \bar{\beta}$.

Theorem 2.1 (Imitation cannot create knowledge)**.** *For a square-integrable predictor β, define the ignorance $K(\beta) = \mathbb{E}\|\bar{\beta} - w^*\|_\Sigma^2$ and the calibration error $\mathrm{CE}(\beta) = \mathbb{E}\|\beta - \bar{\beta}\|_\Sigma^2$. Then*

$$\mathbb{E}\,\mathcal{E}(\beta) = K(\beta) + \mathrm{CE}(\beta). \tag{2.1}$$

Let $\widehat{w}$ be any square-integrable procedure whose observations are the teacher output β_{w} and data U independent of $(w^*, \beta_{\mathrm{w}})$, such as fresh inputs and random seeds. Then $K(\widehat{w}) \geq K(\beta_{\mathrm{w}})$, and

$$\mathbb{E}\mathcal{E}_{\mathrm{w}} - \mathbb{E}\,\mathcal{E}(\widehat{w}) = \big[\mathrm{CE}(\beta_{\mathrm{w}}) - \mathrm{CE}(\widehat{w})\big] - \big[K(\widehat{w}) - K(\beta_{\mathrm{w}})\big] \leq \mathrm{CE}(\beta_{\mathrm{w}}). \tag{2.2}$$

The same holds coordinatewise, and with K replaced by the best linear readout risk K_{lin} when $\widehat{w} = S\beta_{\mathrm{w}}$ for a random matrix S independent of $(w^*, \beta_{\mathrm{w}})$.

Proof. The cross term in $\beta - w^* = (\beta - \bar{\beta}) + (\bar{\beta} - w^*)$ vanishes after conditioning on β. Independence gives $\mathbb{E}[w^* \mid \beta_{\mathrm{w}}, U] = \mathbb{E}[w^* \mid \beta_{\mathrm{w}}]$, and conditioning on the larger sigma-field $\sigma(\beta_{\mathrm{w}}, U) \supseteq \sigma(\widehat{w})$ cannot increase squared error. Section B.1 gives the details and the linear form. □

Three readings follow. A perfectly calibrated teacher cannot be improved by imitation. At equal risk, the teacher with the larger calibration error knows more, and has more to teach. If the student carries its own target-informative state U', such as pretrained features, the floor becomes $K(\beta_{\mathrm{w}}, U')$

and the gain acquires a second source, elicitation; our model isolates recalibration. The ledger is Bayesian: K and CE are relative to the prior, and for a fixed target (2.2) only bounds the gain by the teacher's risk. Calibration conditions on the whole model (x, β), which is stronger than calibrating the prediction value, and the fresh inputs must carry no information about the target, as here.

The theorem also locates the signal a teacher never saw. Symmetric signal coordinates outside the teacher's view have readout zero.

Proposition 2.2 (What the teacher never saw)**.** *If every signal coordinate is symmetric about zero, then any procedure as in Theorem 2.1, including any finite chain of teachers, satisfies*

$$\mathbb{E}\sum_{i>m}\lambda_i(\widehat{w}_i - w_i^*)^2 = A_m^{\mathrm{av}} + \mathbb{E}\sum_{i>m}\lambda_i\widehat{w}_i^2 \;\geq\; A_m^{\mathrm{av}}, \qquad A_m^{\mathrm{av}} = \sum_{i>m}\lambda_i\omega_i^2. \tag{2.3}$$

Proof. For $i > m$, reversing the signs of w_i^* and of column i of the teacher's design preserves the joint law and the teacher's output, so $\mathbb{E}[w_i^* \mid \beta_{\mathrm{w}}] = 0$, and $\mathbb{E}[w_i^* \mid \beta_{\mathrm{w}}, U] = 0$ by independence. The readout of $\widehat{w}$ in coordinate i is zero, and the coordinatewise ledger (2.1) gives the identity. □

The teacher predicts zero there and is already calibrated; a student can only add error. Any gain must come from the teacher's own coordinates.

Calibration slopes. Every student in this paper is linear, $\widehat{w} = S\beta_{\mathrm{w}}$ with a random fitting operator S independent of the teacher, so second moments suffice. Under our assumptions all second moments are diagonal (Lemma 3.1). Write $c_i = \mathbb{E}[w_i^*\beta_i]$, $\rho_i = \mathbb{E}\beta_i^2$ and $g_i = c_i/\rho_i$ for the *state* and *calibration slope* of β in direction i. Then $K_{\mathrm{lin}} = \sum_i \lambda_i(\omega_i^2 - c_i^2/\rho_i)$ and $\mathrm{CE}_{\mathrm{lin}} = \sum_i \lambda_i\rho_i(1-g_i)^2$ (Lemma B.2). Direction i is *overconfident* when $g_i < 1$: the teacher emits more energy there than its correlation with the target warrants.

Proposition 2.3 (What shrinkage can fix)**.** *(i) If $g_i \geq 1$ in every direction with $\rho_i > 0$, then no minimum-norm or ridge student fitted on fresh inputs, with budget and ridge fixed independently of the teacher's output, and no finite chain of them, has smaller expected risk than the teacher, at any finite dimension. (ii) If $\sum_i(\rho_i - c_i) > 0$, then a ridge student with a small positive ridge and enough fresh inputs has smaller expected risk than the teacher.*

Both proofs are short (Section B.1): the filters these students realize have means in $[0, 1]$, so they can only shrink. Condition (ii) is geometry-free because a small ridge shrinks direction i by about η/λ_i and the eigenvalue cancels in the risk. Between the two conditions, the geometry decides.

When geometry supplies no preference. Remove the geometry by setting $\Sigma = I_d$.

Proposition 2.4 (Isotropic imitation is an affine mixture)**.** *If $\Sigma = I_d$, then any teacher independent of the student's inputs satisfies*

$$\mathbb{E}\mathcal{E}_{\mathrm{s}} = \alpha\mathbb{E}\mathcal{E}_{\mathrm{w}} + (1-\alpha)r^2, \qquad \alpha = \min\{n_{\mathrm{s}}/d, 1\}, \tag{2.4}$$

for the minimum-norm student. It cannot improve on a teacher with $\mathbb{E}\mathcal{E}_{\mathrm{w}} < r^2$, and it has no strict interior optimum in sample size.

Proof. For $n_{\mathrm{s}} \leq d$, $\mathbb{E}\widetilde{P} = \alpha I$ and $\widetilde{P}^2 = \widetilde{P}$, so the aggregate correlation $\sum_i c_i$ and energy $\sum_i \rho_i$ are both multiplied by α. For $n_{\mathrm{s}} \geq d$, $\widetilde{P} = I$ almost surely. □

The projection multiplies correlation and energy by the same factor, so the aggregate calibration slope $\sum_i c_i / \sum_i \rho_i$ does not move: in aggregate, isotropic minimum-norm imitation only dilutes the teacher towards zero. Changing the rule changes this. With abundant inputs, a ridge student approaches $\kappa\beta_{\mathrm{w}}$ with $\kappa = (1+\eta)^{-1}$. Writing $\mathsf{A} = \mathbb{E}\|\beta_{\mathrm{w}}\|^2$ and $\mathsf{B} = \mathbb{E}\langle\beta_{\mathrm{w}}, w^*\rangle$,

$$R_{\mathrm{iso}}(\kappa) = \kappa^2\mathsf{A} - 2\kappa\mathsf{B} + r^2, \tag{2.5}$$

and if $0 < \mathsf{B} < \mathsf{A}$ the best shrinkage is the aggregate calibration slope,

$$\kappa^\star = \frac{\mathsf{B}}{\mathsf{A}}, \qquad R_{\mathrm{iso}}(1) - R_{\mathrm{iso}}(\kappa^\star) = \frac{(\mathsf{A}-\mathsf{B})^2}{\mathsf{A}} > 0 \tag{2.6}$$

(Lemma H.1). Uniform shrinkage fixes aggregate overconfidence. Anisotropic inputs let minimum-norm imitation shrink unequally.

3 How imitation recalibrates

The linear ledger needs only the second-order state (c, ρ), and fitting acts on it exactly and linearly.

Lemma 3.1 (Exact second-order calculus). *Let Σ be diagonal, the signal coordinates independent and centered, the teacher a minimum-norm or ridge fit with finite second moments, and every downstream stage a full-dimensional minimum-norm or ridge fit on fresh inputs. Then every predictor in the chain has diagonal second moments, and a noiseless stage with fitting operator S maps the state (c, ρ) to*

$$\big(\bar{F} \odot c,\ W\rho\big), \qquad \bar{F}_i = \mathbb{E}S_{ii}, \quad W_{ij} = \mathbb{E}S_{ij}^2, \tag{3.1}$$

where $\bar{F} \in [0,1]^d$, the column sums of W are at most one, and $W_{ii} \geq \bar{F}_i^2$.

Expected risks of chains are therefore computed by products of fixed matrices, at every finite dimension. It remains to compute $\bar{F}$ and W.

Filter and leak. The teacher's omitted coordinates contribute $A_m = \sum_{i>m} \lambda_i (w_i^*)^2$ to its test error and act as independent label noise during training. We describe every stage with one set of spectral quantities. For a stage fitted on p coordinates with eigenvalues λ_i, sample size n, ridge $\eta \geq 0$, target second moments q_i, and label-noise variance s^2, define

$$\eta = \tau \left(1 - \frac{1}{n} \sum_i \frac{\lambda_i}{\lambda_i + \tau}\right), \qquad F_i = \frac{\lambda_i}{\lambda_i + \tau}, \qquad \Xi = \frac{1}{n} \sum_i F_i^2, \tag{3.2}$$

$$\Gamma = \frac{\sum_i \lambda_i (1 - F_i)^2 q_i + s^2}{n(1 - \Xi)}, \qquad v_i = \frac{\lambda_i \Gamma}{(\lambda_i + \tau)^2}. \tag{3.3}$$

For interpolation with $\eta = 0$ and $n < p$, the fixed point is $\sum_i F_i = n$. In this calculus $\bar{F} \approx F$, and W acts as $\rho \mapsto F^2 \odot \rho + v$: a stage is a *filter* plus a fluctuation. Writing $B = \sum_i \lambda_i (1 - F_i)^2 q_i$ for the energy the filter removes and $D_2 = \sum_i F_i^2$ for its effective dimension,

$$\Gamma = \frac{B + s^2}{n - D_2}, \qquad \sum_i \lambda_i v_i = \Gamma D_2. \tag{3.4}$$

What a stage removes returns as noise, spread over its effective dimension. We call $\sum_i \lambda_i v_i$ the *leak*.

For the teacher, apply (3.2)–(3.3) on H with $q = \omega_H^2$ and $s^2 = \sigma^2 + A_m^{\mathrm{av}}$, extending $F_i^{\mathrm{w}} = v_i^{\mathrm{w}} = 0$ outside H. Its state is $c = F^{\mathrm{w}} \omega^2$ and $\rho = (F^{\mathrm{w}})^2 \omega^2 + v^{\mathrm{w}}$. The student applies the same formulas on all coordinates with $q = \rho$ and $s^2 = 0$, giving $(F^{\mathrm{s}}, v^{\mathrm{s}})$.

Theorem 3.2 (Spectral composition). *Let $d \to \infty$ with $m/d \to \psi \in (0,1)$, $m/n_{\mathrm{w}} \to \bar{\kappa_{\mathrm{w}}} > 1$, and $d/n_{\mathrm{s}} \to \bar{\kappa_{\mathrm{s}}} > 1$. Assume $M^{-1} \leq \lambda_i \leq M$, $\sigma^2 \geq 0$, fixed nonnegative ridges, $\sup_{d,i} d\omega_i^2 < \infty$, a uniformly bounded fourth-moment ratio for the signal coordinates, and convergence of the empirical law of $(i/d, \lambda_i, d\omega_i^2)$. Then $\mathcal{E}_{\mathrm{s}} - \mathcal{R}_s^{(d)} \to 0$ in probability and in expectation, where*

$$\mathcal{R}_s^{(d)} = \underbrace{\sum_i \lambda_i (1 - F_i^{\mathrm{s}} F_i^{\mathrm{w}})^2 \omega_i^2}_{\textit{signal distortion}} + \underbrace{\sum_i \lambda_i (F_i^{\mathrm{s}})^2 v_i^{\mathrm{w}}}_{\textit{teacher fluctuations retained}} + \underbrace{\sum_i \lambda_i v_i^{\mathrm{s}}}_{\textit{leak}}. \tag{3.5}$$

The same holds for the teacher, with $\mathcal{R}_w^{(d)} = \sum_i \lambda_i [(1 - F_i^{\mathrm{w}})^2 \omega_i^2 + v_i^{\mathrm{w}}]$, and for the supervised reference.

The proof conditions on the teacher and applies the Gaussian surrogate of Han & Xu (2023) to the student's linear and quadratic functionals; since that theorem needs positive label noise, we add a small noise variance, which moves estimator and surrogate by the order of its square root, and let $d \to \infty$ before it vanishes (Section G). Given the teacher's law, Lemma 3.1 and the single-stage law reprove the student's expected risk one stage at a time (Section B.4).

The ledger in the calculus. With $g_i = c_i/\rho_i$ and $\pi_i = \lambda_i \rho_i$, completing the square coordinatewise gives

$$\mathcal{R}_s^{(d)} = K_{\mathrm{w}} + \|F^{\mathrm{s}} - g\|_\pi^2 + \mathfrak{T}, \qquad \mathfrak{T} = \sum_i \lambda_i v_i^{\mathrm{s}}, \tag{3.6}$$

$$K_{\mathrm{w}} = \sum_i \lambda_i \Big(\omega_i^2 - \frac{c_i^2}{\rho_i}\Big), \qquad \|h\|_\pi^2 = \sum_i \pi_i h_i^2,$$

with $0/0 = 0$. K_{w} is the teacher's linear ignorance, and the teacher itself is the identity filter, so

$$\mathcal{R}_w^{(d)} - \mathcal{R}_s^{(d)} = \underbrace{\|\mathbf{1} - g\|_\pi^2 - \|F^{\mathrm{s}} - g\|_\pi^2}_{\text{recalibration}} - \underbrace{\mathfrak{T}}_{\text{leak}}. \tag{3.7}$$

The student gains by moving its filter from the identity towards the teacher's calibration slope, and pays the leak. Computing g requires the unknown signal and error profiles; the student never computes it, but its filter can land near it. Where the teacher is overconfident is therefore the crux.

Proposition 3.3 (Where a minimum-norm teacher is overconfident)**.** *Let the teacher interpolate, and let τ_{w} and Γ_{w} be its threshold and noise level in (3.2)–(3.3). Its calculus state satisfies, for $i \le m$,*

$$\rho_i - c_i = \frac{\lambda_i(\Gamma_{\mathrm{w}} - \tau_{\mathrm{w}}\omega_i^2)}{(\lambda_i + \tau_{\mathrm{w}})^2}, \qquad \sum_{i \le m} (\rho_i - c_i) = \frac{\sigma^2 + A_m^{\mathrm{av}}}{\tau_{\mathrm{w}}}. \tag{3.8}$$

Among observed directions $i \le m$ it is overconfident exactly where $\omega_i^2 < \Gamma_{\mathrm{w}}/\tau_{\mathrm{w}}$, and in aggregate whenever $\sigma^2 + A_m^{\mathrm{av}} > 0$.

Whether a direction is overconfident depends only on its signal strength, not on its eigenvalue. When the signal is aligned with the spectrum, the teacher is overconfident in its low-variance directions—exactly where a student's filter shrinks hardest. This coincidence is the mechanism of the gain.

4 How much imitation is useful?

With $n_{\mathrm{s}} \ge d$ fresh Gaussian inputs, the minimum-norm student copies its teacher exactly, at every finite d. Any gain must disappear by that point, so the path to cloning is a path along which recalibration is withdrawn. The first question is whether imitating slightly less than a clone helps.

Proposition 4.1 (When imitating less helps)**.** *Let the teacher have state (c, ρ), overconfidence $\mathfrak{O} = \sum_i (\rho_i - c_i)$ and clone leak rate $\ell = \frac{d}{2} \sum_i (\rho_i/\lambda_i) / \sum_i \lambda_i^{-1}$. As $n_{\mathrm{s}} \uparrow d$, the calculus gives*

$$\mathcal{R}_s^{(d)} = \mathcal{R}_w^{(d)} - \frac{2(\mathfrak{O} - \ell)(d - n_{\mathrm{s}})}{\mathrm{tr}(\Sigma^{-1})} + O\big((1 - n_{\mathrm{s}}/d)^2\big). \tag{4.1}$$

Under isotropy $\mathfrak{O} - \ell = (\mathcal{R}_w^{(d)} - r^2)/2$. For spectra in $[M^{-1}, M]$ the remainder is uniform in d.

Near the clone, the student shrinks direction i by about τ/λ_i. The eigenvalue cancels in the recalibration, $2\tau(\rho_i - c_i)$ per direction, but not in the leak: soft directions give up more energy, and that energy returns across all d directions. Under isotropy it reduces to an incompetent teacher (Proposition 2.4). In the limit, a competent teacher with $\mathfrak{O} > \ell$ therefore has an interior optimal budget strictly below both endpoints, and under bounded spectra the actual student beats the actual teacher there with probability tending to one (Corollary B.3). In the spiked family below, $\mathfrak{O} > \ell$ reduces to $d > 2k$.

Kept and leaked fractions. Put k eigenvalues at height L and the rest at one, with all signal on the spikes, $\omega_i^2 = r^2/(kL)$, inside the teacher's view. Take the proportional limit of the formulas first, holding $\theta = k/d$, m/d, n_{w}/d, n_{s}/d and L fixed, and then let $L \to \infty$; dimension symbols below denote ratios. At fixed L the spectrum is bounded, so with a uniformly bounded fourth-moment ratio Theorem 3.2 makes these the limits of the actual risks.

Theorem 4.2 (An optimal budget for a fixed interpolation rule)**.** *In this limit, with $k < n_{\mathrm{w}} < m$, $2k < d$, $k < n_{\mathrm{s}} < d$ and $\sigma^2 > 0$, put $t = n_{\mathrm{s}} - k$, $D = d - k$ and $V = \sigma^2(n_{\mathrm{w}} - k)/(m - n_{\mathrm{w}})$, the teacher's error in the bulk. The student keeps a fraction t/D of V and returns a fraction k/t of what*

it removes to the spike directions, so

$$G(n_{\mathrm{s}}) = V\Big(1 - \frac{t}{D}\Big)\Big(1 - \frac{k}{t}\Big) = \sigma^2 \frac{(d - n_{\mathrm{s}})(n_{\mathrm{w}} - k)(n_{\mathrm{s}} - 2k)}{(d - k)(n_{\mathrm{s}} - k)(m - n_{\mathrm{w}})}. \tag{4.2}$$

The gain is positive precisely for $2k < n_{\mathrm{s}} < d$ and is maximized where the kept and leaked fractions are equal, $t/D = k/t$:

$$n_{\mathrm{s}}^\star = k + \sqrt{k(d-k)}, \qquad \frac{n_{\mathrm{s}}^\star}{d} = \theta + \sqrt{\theta(1-\theta)}, \qquad G(n_{\mathrm{s}}^\star) = V\Big(1 - \sqrt{\tfrac{k}{d-k}}\Big)^2. \tag{4.3}$$

Every increase beyond this budget reduces the gain until cloning removes it. Strict improvement persists at sufficiently large finite L.

The student spends k equations on the spikes and t on the isotropic bulk, where it spans a fraction t/D and keeps that share of the teacher's error. By the leak law (3.4), the error it removes returns as noise, and the k spike directions receive a fraction k/t of it. More budget keeps more and leaks less; at the optimum the two fractions coincide. This theorem and Proposition 4.1 are one first-order balance between recalibration and leak, read at two budgets (fixed points in Sections B.2 and I.4). In the subsequent, ordered limit $\theta \downarrow 0$, $n_{\mathrm{s}}^\star/d \sim \sqrt{\theta}$, with no uniform claim for fixed k.

Samples pay for filters. Every point above is a fully solved interpolation problem, in which sample size sets the filter through the rank of the design. Ridge decouples the two. For $\tau > 0$, write $F_i(\tau) = \lambda_i/(\lambda_i + \tau)$, $D_j(\tau) = \sum_i F_i(\tau)^j$, and $B(\tau) = \sum_i \lambda_i \rho_i (1 - F_i(\tau))^2$, and set $R(F) = \|F - g\|_\pi^2 + K_{\mathrm{w}}$. At any $n \geq D_1(\tau)$, the ridge $\eta = \tau[1 - D_1(\tau)/n]$ realizes that filter, and by (3.4)

$$\mathcal{R}_n(\tau) = R(F(\tau)) + \frac{B(\tau) D_2(\tau)}{n - D_2(\tau)}. \tag{4.4}$$

More samples buy the same recalibration with a smaller leak, so the optimally tuned formula risk is nonincreasing in budget (Proposition I.5). The interior optimum belongs to minimum-norm imitation; it does not prescribe discarding data when ridge can be retuned.

5 What does another distillation stage buy?

One student chooses how far to recalibrate. A further student changes the filter shapes available, and Theorem 2.1 fixes the terms of trade: no stage can recover knowledge an earlier stage leaked. Let the teacher be stage zero and fit stage $t = 1, \ldots, T$ on an independent batch of n_t pseudo-labels, with $\sum_t n_t = N$. Initialize $(\alpha^{(0)}, \nu^{(0)}) = (F^{\mathrm{w}}, v^{\mathrm{w}})$ and apply (3.2)–(3.3) with $q_i = (\alpha_i^{(t-1)})^2 \omega_i^2 + \nu_i^{(t-1)}$ and $s^2 = 0$:

$$\alpha_i^{(t)} = F_i^{(t)} \alpha_i^{(t-1)}, \qquad \nu_i^{(t)} = (F_i^{(t)})^2 \nu_i^{(t-1)} + v_i^{(t)}. \tag{5.1}$$

For bounded spectra, the risk-relevant aggregates of the chain converge to this recursion (Theorem D.9). With $F^{\mathrm{chain}} = \prod_t F^{(t)}$, the terminal formula is $R(F^{\mathrm{chain}}) + \mathfrak{T}_{\mathrm{chain}}$, where

$$\mathfrak{T}_{\mathrm{chain}} = \sum_i \lambda_i \sum_{j=1}^{T} v_i^{(j)} \prod_{u=j+1}^{T} (F_i^{(u)})^2. \tag{5.2}$$

Each stage filters the leaks before it and adds its own.

Purification, then forgetting. If n_t/d stays in a fixed compact subset of $(0, 1)$, with bounded ridges and a deterministic schedule, the risk approximation holds uniformly in depth, including depths growing with d (Theorem J.2); for expected risks it also follows, given the teacher's law, from Lemma 3.1 and the single-stage law by telescoping (Proposition B.6). Long chains obey an exact bound:

$$\mathbb{E}\|\beta^{(T)}\|^2 \leq \mathbb{E}\|\beta_{\mathrm{w}}\|^2 \prod_{t=1}^{T} \left[1 - \frac{1 - n_t/d}{\kappa_\Sigma}\right], \qquad \kappa_\Sigma = \frac{\lambda_1}{\lambda_d} \tag{5.3}$$

(Section J.2). The ledger refines this. For identical stages the exact calculus iterates one positive matrix W: energy decays at its Perron root $\mu < 1$, and the knowledge held in direction i at rate

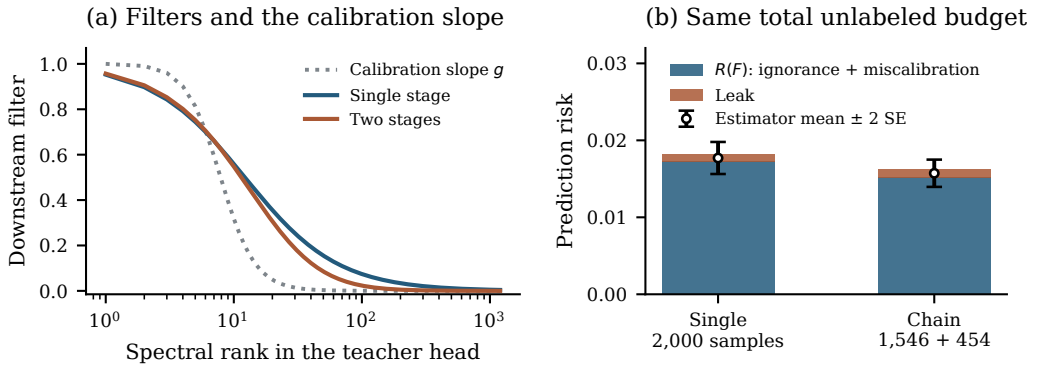


Figure 2: **Depth buys shape with knowledge.** The power-law instance of (5.6). Left: the downstream filters of the best single student and the certified two-stage witness, with the teacher's calibration slope g. Right: predicted risk split into $R(F)$, the teacher's linear ignorance plus the calibration error left by the filter, and the leak. Markers show means and two standard errors from 40 paired estimator trials; ridges were chosen from the formulas, with the prior known, before the data were drawn. Both use the same total number of unlabeled inputs.

$\bar{F}_i^2/\mu < 1$ (Proposition B.4). Every such chain is eventually uninformed, but between the two clocks it is underconfident, still knowing much of what it no longer says. In the calculus, a power-law chain with $n_t = d/2$ halves its energy every 470 stages but keeps its top direction's knowledge for about 1.8×10^8; its linear calibration error falls from 0.172 to 0.0025 at depth 47, where the risk 0.066 nearly reaches the teacher's linear ignorance 0.063, and returns to 0.758 by depth 5000 while the ignorance is still 0.210 (Table 1). Under isotropy the clocks coincide.

What richer filters can achieve. At fixed d, with downstream samples growing and population ridges t_j fixed, the leaks vanish and the chain reaches the filter families

$$\mathcal{R}_T = \left\{ \prod_{j=1}^{T} \varphi_{t_j} : t_j \in [0, \infty] \right\}, \qquad \varphi_t(\lambda_i) = \frac{\lambda_i}{\lambda_i + t}. \tag{5.4}$$

These are nested, since $\varphi_0 = 1$. One positive-ridge resolvent decays linearly in a small eigenvalue and two decay quadratically, so depth can suppress a noisy tail more sharply (Figure 2, left).

Proposition 5.1 (A sufficient test for a useful extra stage)**.** *Let $F^\star = \varphi_{t^\star}$ be a global minimizer of R over $\mathcal{R}_1$ with $0 < t^\star < \infty$. If*

$$\left\langle F^\star - g, \frac{F^\star}{\lambda} \right\rangle_\pi > 0, \tag{5.5}$$

then $\min_{F \in \mathcal{R}_2} R(F) < \min_{F \in \mathcal{R}_1} R(F)$.

Proof. The feasible path $F(u) = F^\star \varphi_u$ has $F'(0) = -F^\star/\lambda$, so $\frac{d}{du} R(F(u))|_{0^+} < 0$ under (5.5). □

For $\lambda_i = i^{-a}$ and $\omega_i^2 \propto i^{-b}$ with a noisy minimum-norm teacher, a sign analysis gives a population tie at $b = 0$ and a strict two-stage improvement for $b \geq a$ (Theorem D.11). The test has no local converse: at a certified instance it points away from improvement, yet a nonlocal two-stage choice lowers the risk by more than 2.2×10^{-10} (Theorem D.12), too small a margin to matter in practice.

Whether finite data can pay for depth. At a fixed budget, the filters and leaks both depend on the allocation (n_t); let $R^\star_{T,N}$ be the smallest formula risk of a T-stage chain with total budget N, over allocations and ridges. For $(d, m, n_w, N) = (4000, 1200, 400, 2000)$, $(a, b) = (1.2, 2)$, $r^2 = 1$ and $\sigma^2 = 0.25$, a directed-rounding certificate proves

$$R^\star_{2,N} \leq 0.016302 < 0.017217 \leq R^\star_{1,N}. \tag{5.6}$$

Replicating each of the 4000 spectral levels q times, dividing its signal variance by q, and multiplying all budgets by q leaves every formula unchanged and the spectrum bounded. As $q \to \infty$, with probability tending to one, the two-stage witness beats *every* single-stage ridge by at least 9.15×10^{-4} under the same total budget, including a ridge tuned after seeing the data (Theorem J.30). The levels stay fixed, so this is not a growing power-law theorem. In our two-level spectra, one stage won in all but one case (Section B.3): a chain spends knowledge twice.

6 Scaling: labels buy knowledge, pseudo-labels buy calibration

Pseudo-labels change how a teacher's knowledge is expressed and can only lose knowledge; only true labels can increase it. Since the teacher is a function of the labeled data (X, y), Theorem 2.1 gives every downstream pipeline without further signal-dependent information at least the ignorance of an oracle that sees (X, y). For a Gaussian prior with $\Omega = \mathrm{diag}(\omega_i^2)$ and $\sigma^2 > 0$, the oracle's posterior covariance is $(\Omega^{-1} + \sigma^{-2} X^\top X)^{-1}$, and operator convexity with $\mathbb{E} X^\top X = n_{\mathrm{w}} \Sigma$ gives, for every such pipeline,

$$\mathbb{E}\mathcal{E}(\widehat{w}) \geq \sum_{i=1}^{d} \frac{\lambda_i \omega_i^2}{1 + (n_{\mathrm{w}}/\sigma^2)\lambda_i \omega_i^2}. \tag{6.1}$$

For $\lambda_i = i^{-a}$ and $\omega_i^2 = G i^{-b}$, with $a > 1$, $b \geq 0$ and G fixing r^2, this is $\Omega(n_{\mathrm{w}}^{-(a+b-1)/(a+b)})$ once $d \geq c n_{\mathrm{w}}^{1/(a+b)}$. These spectra violate the lower spectral bound of Theorem 3.2; the statements below are asymptotics of the deterministic formulas (Theorem D.14 and Section K).

Proposition 6.1 (The interpolation floor is calibration error)**.** *Along this power law with $\sigma^2 > 0$ and fixed ratios $(\psi, \bar{\kappa_{\mathrm{w}}}, \bar{\kappa_{\mathrm{s}}})$, the minimum-norm student's formula risk tends to $\sigma^2 C_{\mathrm{floor}} > 0$, while the linear ignorance of both teacher and student is $O(n_{\mathrm{w}}^{-(a+b-1)/(a+b)})$: the limiting risk is all calibration error.*

In the formulas, the teacher at fixed shape acquires knowledge at the Bayes exponent, yet interpolation keeps a constant noise floor, and the student passes it on. Regularization removes it. With both ridges tuned, $m = \Theta(n_{\mathrm{w}})$, $n_{\mathrm{s}}/n_{\mathrm{w}} \to \infty$ and $d/n_{\mathrm{s}} \to \bar{\kappa_{\mathrm{s}}} > 1$, the pipeline attains the supervised exponent

$$\mathcal{R}_s^\star(n_{\mathrm{w}}) \asymp n_{\mathrm{w}}^{-(a+b-1)/(a+b)}, \qquad b < a + 1, \tag{6.2}$$

matching the lower bound (6.1). A noiseless exact teacher knows everything and is calibrated; imitating it with $\liminf d/n_{\mathrm{s}} > 1$ costs only filter and leak, at rate $n_{\mathrm{s}}^{-\min\{a+b-1, 2a\}}$ (up to a logarithm at $b = a + 1$).

7 Discussion and Conclusion

The ledger separates two ways to beat a teacher: saying better what it knows, or knowing something it does not. Our model isolates the first. The bound $K(\beta_{\mathrm{w}}, U')$ shows what a theory of the second must supply: a state U' of the student, from pretraining or a richer representation, informative about the target given the teacher. A prior linking the teacher's observed and missing signs breaks the symmetry of Proposition 2.2 without adding knowledge, letting recalibration reach unobserved coordinates; elicitation needs a target-correlated representation.

The ledger also suggests measurements: with teacher, representation and rule fixed, vary the unlabeled budget and track its three terms, estimating calibration slopes from a held-out labeled set pooled across directions or tasks, and separating filtering from leak with repeated fresh batches (Appendix N). Budget and ridge choose a filter and pay for it with a leak, and more fresh inputs at a fixed filter lower only the leak (4.4). Choosing a filter from teacher outputs alone, an estimator theorem at a growing power-law spectrum with errors small relative to the vanishing risk, and a simple spectral family in which depth pays at equal budget remain open.

We asked how a student can be more accurate than its teacher without knowing more; it recalibrates. Perfect imitation preserves every mistake; selective imitation can correct some of them, but never by adding knowledge. Every gain a student makes over its teacher is recalibration, paid for by a leak, and the spectrum decides which recalibrations are available and at what price. A student can become stronger than its teacher by saying better what the teacher knows; to know more, it needs knowledge of its own.

Reproducibility Statement

The supplement contains complete statements and proof details for the results used in the main text. The accompanying package includes the deterministic risk evaluator, the two directed-rounding certificates, and the scripts, seeds, and per-trial results for the numerical checks reported here; `audit/check_ledger.py` recomputes every number quoted for the knowledge–calibration ledger. Numerical evaluations of a deterministic formula are distinguished from simulations of trained estimators. Appendix N.3 describes the exact symbolic checks and the two interval certificates, and what each of them covers.

AI Use Statement

The original research ideas and initial manuscript draft were developed by the authors. AI tools were used to assist with checking mathematical details and the correctness of arguments, filling in routine proof steps, coding assistance, and language polishing. The authors take full responsibility for the content and correctness of the final manuscript.

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.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. doi: 10.1007/s10208-006-0196-8.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hien Dang, Pratik Patil, and Alessandro Rinaldo. Optimal unconstrained self-distillation in ridge regression: Strict improvements, precise asymptotics, and one-shot tuning. *arXiv preprint arXiv:2602.17565v1*, 2026. URL https://arxiv.org/abs/2602.17565v1.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1–2):12–22, 1983. doi: 10.2307/2987588.

Yijun Dong, Yicheng Li, Yunai Li, Jason D. Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through the lens of intrinsic dimension. In *International Conference on Machine Learning (ICML)*, 2025.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, 1996. doi: 10.1007/978-94-009-1740-8.

Anvit Garg, Sohom Bhattacharya, and Pragya Sur. Preventing model collapse under overparametrization: Optimal mixing ratios for interpolation learning and ridge regression. *arXiv preprint arXiv:2509.22341v2*, 2026. URL https://arxiv.org/abs/2509.22341v2.

Qiyang Han and Xiaocong Xu. The distribution of ridgeless least squares interpolators. *arXiv preprint arXiv:2307.02044*, 2023.

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

M. Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Radu Lecoiu, Debarghya Mukherjee, and Pragya Sur. Self-distillation is optimal among spectral shrinkage estimators in spiked covariance models. *arXiv preprint arXiv:2605.17778*, 2026.

James Legge. *The Chinese Classics, Volume I: Confucian Analects, The Great Learning, and The Doctrine of the Mean*. Trübner, London, 1861.

Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-to-strong generalization even in random feature networks, provably. In *International Conference on Machine Learning (ICML)*, 2025.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. Self-distillation amplifies regularization in hilbert space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Behrad Moniri and Hamed Hassani. On the mechanisms of weak-to-strong generalization: A theoretical perspective. *arXiv preprint arXiv:2505.18346*, 2025.

Abhijeet Mulgund and Chirag Pabbaraju. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. *arXiv preprint arXiv:2501.19105*, 2025.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973. doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Junsoo Oh, Jerry Song, and Chulhee Yun. From linear to nonlinear: Provable weak-to-strong generalization through feature learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok (eds.), *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012.

David X. Wu and Anant Sahai. Provable weak-to-strong generalization via benign overfitting. In *International Conference on Learning Representations (ICLR)*, 2025.

Diyuan Wu, Lehan Chen, Theodor Misiakiewicz, and Marco Mondelli. Improved scaling laws via weak-to-strong generalization in random feature ridge regression. *arXiv preprint arXiv:2603.05691v3*, 2026a. URL https://arxiv.org/abs/2603.05691v3.

Mingqi Wu, Archer Y. Yang, and Qiang Sun. Why self-training helps and hurts: Denoising vs. signal forgetting. *arXiv preprint arXiv:2602.14029*, 2026b. URL https://arxiv.org/abs/2602.14029v1.

Wendao Wu, Fangqing Zhang, Haihan Zhang, and Cong Fang. What makes a strong model? a unified spectral analysis of knowledge transfer over high-dimensional linear regression. *arXiv preprint arXiv:2606.01292v1*, 2026c. URL https://arxiv.org/abs/2606.01292v1.

Wei Yao, Wenkai Yang, Gengze Xu, Ziqiao Wang, Yankai Lin, and Yong Liu. The capabilities and limitations of weak-to-strong generalization: Generalization and calibration. *arXiv preprint arXiv:2502.01458*, 2025.

CONTENTS OF THE APPENDIX

A The dependency of results, and notation **15**

A.1 Dependency of results . . . 15

A.2 Notation . . . 16

B The ledger: knowledge, calibration, and leak **17**

B.1 Exact statements . . . 17

B.2 The ledger in the deterministic calculus . . . 19

B.3 Equal-budget depth on two-level spectra . . . 22

B.4 A second proof of the chain approximation in expectation . . . 22

C Detailed setup and spectral calculus **23**

C.1 The interpolation channel . . . 24

C.2 The Wiener form of the risk . . . 25

D Detailed statements and extensions **27**

D.1 The composition theorem . . . 27

D.2 Structural corollaries of the single hop . . . 29

D.3 The imitation sweet spot . . . 30

D.4 Distillation chains: when ladders beat leaps . . . 32

D.5 Weak-to-strong scaling laws . . . 36

E Notation and proof routes **37**

E.1 Notation . . . 37

E.2 Proof routes . . . 38

E.3 Scope of the supporting results . . . 38

F The interpolation channel: statement and provenance **38**

F.1 The Wiener risk decomposition . . . 39

G Proof of the composition theorem **40**

G.1 A single-stage law including zero noise . . . 41

G.2 Composition by conditioning . . . 43

G.3 Random signals and expected risks . . . 43

H Structural corollaries **44**

H.1 Isotropic minimum-norm imitation . . . 44

H.2 Super-ceiling threshold . . . 46

H.3 Teacher capacity formulas and the transferred double descent . . . 47

I Proof of the sweet-spot theorem **48**
I.1 Master fixed-point identity . . . 48
I.2 Part (i): the clone plateau (exact, finite d) . . . 49
I.3 Part (ii): continuity and the two boundary limits . . . 49
I.4 Part (iii): a provable non-monotone instance in the spiked family . . . 50
I.5 Corollary: sweet spot $\iff$ gain, and the gain-free overshoot witness . . . 52
I.6 The cost of holding the filter fixed . . . 53

J Proofs for chains **54**
J.1 Aggregate recursion and its stability in depth . . . 54
J.2 An exact finite-dimensional forgetting bound . . . 56
J.3 The population filter objective and the strictness criterion . . . 57
J.4 Global necessity: the trichotomy and the certified counterexample . . . 60
J.5 The power-law threshold . . . 68
J.6 Finite budgets: the phase transition . . . 69
J.7 Mechanism: composition raises the filter order . . . 79
J.8 From a finite formula certificate to estimator separation . . . 79

K Proofs of the scaling theorems **81**
K.1 Spectral toolkit . . . 82
K.2 (i) The imitation rate . . . 86
K.3 (ii) The noise floor . . . 88
K.4 (iii) No scaling penalty for weak supervision . . . 88

L The Gaussian sketch (isotropic features) **94**

M Finite-sample information bounds **95**
M.1 The observation model in Proposition 2.2 . . . 95
M.2 Gaussian Bayes lower bound . . . 95

N Numerical checks and reproducible artifacts **96**
N.1 Finite-dimensional checks of the spiked example . . . 96
N.2 Paired chain comparison . . . 96
N.3 Symbolic checks and interval certificates . . . 96

ROADMAP TO THE APPENDIX

The appendix proves every statement of the main text at the level at which it is stated, and records the protocol behind every number. Start with Section A: Figure 3 shows which results build on which and sorts them into three levels (exact at every finite dimension, the deterministic risk calculus, and estimators under bounded spectra), and Section A.2 collects the notation of the main text; symbols local to the derivations are listed in Section E.1. Section D restates the main results in full generality, and the remaining appendices follow the main text section by section.

Knowledge and calibration (Section 2). Section B.1 proves the ledger (Theorem 2.1) by conditioning, and its linear form through the diagonal ledger (Lemma B.2). It also proves Proposition 2.2, the exact second-order calculus (Lemma 3.1), and the limits of shrinkage (Proposition 2.3). The observation model behind Proposition 2.2 and the Gaussian Bayes bound (6.1) are treated in Section M; the isotropic identity (Proposition 2.4) and the ridge comparison (2.6) in Section H.

How imitation recalibrates (Section 3). Section C fixes the setup and defines the stage calculus (Section C.1) and its Wiener form (Section C.2). Section F states the one imported result, the Gaussian surrogate of Han & Xu (2023) (Lemma F.1), and Section G turns it into the composition theorem (Theorem 3.2, in full as Theorem D.1) through a small-noise coupling (Lemma G.2), aggregate laws (Lemma G.3) and conditioning on the teacher. Section B.2 derives the leak law and proves Proposition 3.3; Section B.4 reproves the expected chain risks from the exact calculus. Further corollaries of a single stage are stated in Section D.2 and proved in Sections H.2 and H.3.

The imitation budget (Section 4). Section B.2 also proves the clone criterion (Proposition 4.1), the interior budget it implies (Corollary B.3), and the kept-and-leaked reading of Theorem 4.2. The spiked theorem in full (Theorem D.6) is proved in Section I, which ends with the cost of holding the filter fixed, (4.4) and Proposition I.5.

Depth (Section 5). Section J proves the chain recursion (Theorem D.9) and its extension uniform in depth (Theorem J.2), the forgetting bound (5.3) (Proposition J.3), the extra-stage test (Proposition 5.1) with its trichotomy and certified counterexample (Theorems D.11 and D.12), and the transfer of the certificate (5.6) to actual estimators (Theorem J.30). The two clocks (Proposition B.4) are proved in Section B.2, and the equal-budget comparisons on two-level spectra are in Section B.3.

Scaling (Section 6). Section K proves the power-law asymptotics of the formulas (Theorem D.14): a spectral toolkit, the imitation rate, the noise floor, and the absence of a scaling penalty for weak supervision. Proposition 6.1 is proved in Section B.2, and Section L extends the model to Gaussian sketches under isotropy.

Numerics and checks. Section N records the protocols, seeds and paired trials behind Figures 1 and 2, and Section N.3 describes the exact symbolic checks and the two interval certificates, with what each of them covers. The supplementary archive contains every script, seed and raw result: `code/REPRODUCE.md` lists each script with its protocol, seed and expected output, and both certificates end with an explicit verdict.

Levels and vocabulary. The exact statements of Sections B.1 and M need no deterministic-equivalent approximation. The composition argument of Section G and its extension uniform in chain depth in Section J assume bounded spectra. The power-law analyses of Sections J and K concern the deterministic risk calculus; an estimator theorem at a growing power-law spectrum remains open. The detailed appendices keep some earlier vocabulary: the Wiener filter g of Definition C.3 is the teacher's calibration slope, Irr its linear ignorance, $\|\mathbf{1} - g\|_\pi^2$ its calibration error, and the toll $\mathfrak{T}$ the leak. Section E.2 summarizes the proof routes.

Reading paths. For the framework alone, Section A, Section B.1 and Section M suffice; none of them uses random matrix theory. For the spectral calculus and its probabilistic justification, add Sections C, F and G; the budget, depth and scaling results are then developed in Sections I to K, and Section E.3 records the scope of the supporting results.

A THE DEPENDENCY OF RESULTS, AND NOTATION

This section shows how the results of the paper build on one another and collects the notation of the main text; symbols used only in the detailed derivations are listed in Section E.1. Every result holds at one of three levels: exactly at every finite dimension, in the deterministic risk calculus, or for the actual estimators under bounded spectra.

A.1 DEPENDENCY OF RESULTS

Figure 3 sorts the results by level. Apart from self-contained results, the exact ones descend from two roots, the ledger (Theorem 2.1 and Definition B.1) and the exact second-order calculus (Lemma 3.1). The deterministic calculus starts from the stage formulas (3.2)–(3.3) and meets the ledger in (3.6)–(3.7), through the diagonal ledger (Lemma B.2). The estimator-level results prove that actual risks converge to the calculus formulas and transfer comparisons made there. Besides standard facts of matrix analysis and random matrix theory, and interval-arithmetic certificates, they import one theorem: the Gaussian surrogate of Han & Xu (2023), which a small-noise coupling turns into the single-stage law.

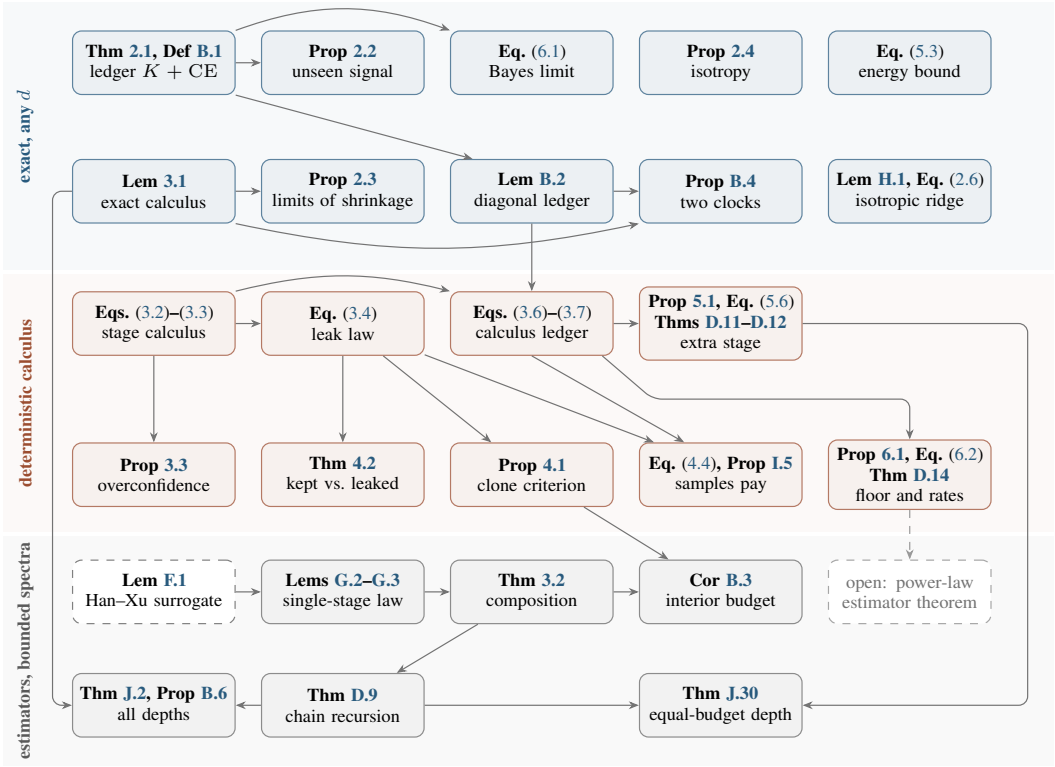


Figure 3: **Dependency of results.** An arrow points from a result to one that builds on it; arrows implied by transitivity are omitted, and unconnected results are proved directly. Blue nodes are exact at every finite dimension; umber nodes are statements about the deterministic calculus; gray nodes concern the actual estimators under bounded spectra. The dashed node is the one imported theorem (standard matrix and random-matrix facts are not drawn), and the dashed open node marks the missing estimator theorem at a growing power-law spectrum.

A.2 NOTATION

Symbol	Meaning	Defined in
Conventions and model		
$\langle\cdot,\cdot\rangle$, $\|\cdot\|_A$, $A^\dagger$, $\odot$, $\asymp$	inner product; norm $(u^\top Au)^{1/2}$; pseudoinverse; entrywise product; same order	§2
$\mathbf{1}$, λ, c	all-ones vector; vector of eigenvalues; a generic constant, not c_i	§3, §5, §6
d, m, H	ambient dimension; the teacher observes $H = \{1, \dots, m\}$	§2
x, y, ξ, X, X_H	input, label, label noise; labeled design and its observed columns	§2
n_{w}, n_{s}	true-label budget; pseudo-label budget on fresh inputs	§2
$\Sigma = \mathrm{diag}(\lambda_i)$	input covariance, $\lambda_1 \geq \cdots \geq \lambda_d > 0$	§2
w^*, ω_i^2	target, with independent centered coordinates; their variances	§2
σ^2, r^2, η	label-noise variance; signal power $\sum_i \lambda_i \omega_i^2$; ridge	§2
β_{w}, $\widehat{w}_{\mathrm{s}}$, $\widetilde{X}$, $\widetilde{P}$	teacher and student; fresh design and projection onto its row space	(1.1)
$\widehat{\Sigma}$	sample covariance $\widetilde{X}^\top \widetilde{X}/n_{\mathrm{s}}$ of the fresh design	§2
$\mathcal{E}(w)$, $\mathcal{E}_{\mathrm{w}}$, $\mathcal{E}_{\mathrm{s}}$	excess risk $\|w - w^*\|_\Sigma^2$; its values for teacher and student	§2
A_m^{av}, A_m	omitted signal energy $\sum_{i>m} \lambda_i \omega_i^2$; its realized value $\sum_{i>m} \lambda_i (w_i^*)^2$	(2.3), §3
Ledger and isotropy		
$\bar{\beta}$	readout $\mathbb{E}[w^* \mid \beta]$ of a predictor β	§2
K, CE	ignorance and calibration error; $\mathbb{E}\,\mathcal{E} = K + \mathrm{CE}$	(2.1)
$\widehat{w}$, U	procedure fed by β_{w} and data U independent of $(w^*, \beta_{\mathrm{w}})$	Thm 2.1
S	random fitting matrix independent of $(w^*, \beta_{\mathrm{w}})$; a linear student is $S\beta_{\mathrm{w}}$	Thm 2.1
K_{lin}, $\mathrm{CE}_{\mathrm{lin}}$	ignorance and calibration error with linear readouts	Thm 2.1, §2
U'	a target-informative state of the student	§2
c_i, ρ_i	second-order state $\mathbb{E}[w_i^* \beta_i]$ and $\mathbb{E}\beta_i^2$	§2
g_i	calibration slope c_i/ρ_i; direction i is overconfident when $g_i < 1$	§2
α	factor $\min\{n_{\mathrm{s}}/d, 1\}$ of isotropic minimum-norm imitation	(2.4)
A, B	teacher energy $\mathbb{E}\|\beta_{\mathrm{w}}\|^2$ and correlation $\mathbb{E}\langle \beta_{\mathrm{w}}, w^* \rangle$	(2.5)
κ, $\kappa^\star$, R_{iso}	uniform shrinkage factor; its best value B/A; its risk	(2.5), (2.6)
Exact calculus and deterministic calculus		
$\bar{F}_i$, W_{ij}	exact maps of a fresh stage, $\mathbb{E}S_{ii}$ and $\mathbb{E}S_{ij}^2$	(3.1)
p, n, q_i, s^2	stage size and samples; target second moments; label noise	§3
τ, F_i, Ξ	stage threshold; filter $\lambda_i/(\lambda_i + \tau)$; $\Xi = \sum_i F_i^2/n$	(3.2)
Γ, v_i	noise level; fluctuation profile	(3.3)
B, D_1, D_2	removed energy; $D_j = \sum_i F_i^j$, D_2 the effective dimension	(3.4), (4.4)
F^{w}, v^{w}, F^{s}, v^{s}	teacher and student profiles in the calculus	§3
M, ψ, $\bar{\kappa}_{\mathrm{w}}$, $\bar{\kappa}_{\mathrm{s}}$	spectral bound $M^{-1} \leq \lambda_i \leq M$; limiting ratios m/d, m/n_{w}, d/n_{s}	Thm 3.2
$\mathcal{R}_w^{(d)}$, $\mathcal{R}_s^{(d)}$	deterministic risk formulas of teacher and student	(3.5)
K_{w}, π_i, $\|\cdot\|_\pi$, $\mathfrak{T}$	teacher's linear ignorance; weight $\lambda_i \rho_i$, its norm; leak $\sum_i \lambda_i v_i^{\mathrm{s}}$	(3.6)
τ_{w}, Γ_{w}	teacher threshold and noise level	Prop 3.3
Budget		
$\mathfrak{O}$, ℓ	overconfidence $\sum_i (\rho_i - c_i)$; clone leak rate	Prop 4.1
k, L, θ	number and height of spikes; spike fraction k/d	§4
t, D, V	excess budget $n_{\mathrm{s}} - k$; bulk size $d - k$; the teacher's bulk error	Thm 4.2
$G(n_{\mathrm{s}})$, $n_{\mathrm{s}}^\star$	gain of the spiked student; optimal budget $k + \sqrt{k(d-k)}$	(4.2), (4.3)
$R(F)$, $\mathcal{R}_n(\tau)$	filter objective $\|F - g\|_\pi^2 + K_{\mathrm{w}}$; risk of its ridge with n samples	(4.4)
Depth		
t, T, n_t, N	stage index; chain depth; stage budgets; total budget	§5
$\alpha^{(t)}$, $\nu^{(t)}$	chain signal coefficient and fluctuation profile	(5.1)
F^{chain}, $\mathfrak{T}_{\mathrm{chain}}$	product filter; accumulated leak	(5.2)
$\beta^{(t)}$, κ_Σ, μ	stage-t predictor; condition number λ_1/λ_d; Perron root of W	(5.3), §5
φ_t, t_j, $\mathcal{R}_T$	resolvent at population ridge t; ridges of a chain; T-stage filters	(5.4)
$F^\star$, $t^\star$	best single-stage filter and its ridge	Prop 5.1
$R_{T,N}^\star$, q	best formula risk of T stages at total budget N; replication factor	§5
Scaling		
a, b, G	exponents in $\lambda_i = i^{-a}$, $\omega_i^2 = G i^{-b}$; normalization G (not $G(n_{\mathrm{s}})$)	§5, §6
Ω	prior covariance $\mathrm{diag}(\omega_i^2)$; $\Omega(\cdot)$ and $\Theta(\cdot)$ are orders of growth	(6.1)
C_{floor}	interpolation floor constant	Prop 6.1
$\mathcal{R}_s^\star$	formula risk with both ridges tuned	(6.2)

B THE LEDGER: KNOWLEDGE, CALIBRATION, AND LEAK

This appendix proves the statements that organize the main text. The first subsection is exact: it holds at every finite dimension and, where stated, for every prior and every downstream algorithm. The second subsection works in the deterministic calculus of Section C.1. The last subsection gives a second proof, in expectation, of the chain approximations, from the exact calculus, the single-stage law, the teacher's law and uniform integrability.

B.1 EXACT STATEMENTS

Throughout, $\Sigma \succ 0$ is fixed and all random vectors are square integrable. For a random vector Z, write $\bar{w}(Z) = \mathbb{E}[w^* \mid Z]$.

Definition B.1 (Readout, ignorance, calibration error)**.** For a predictor β, its *readout* is $\bar{\beta} = \mathbb{E}[w^* \mid \beta]$. Its *ignorance* is $K(\beta) = \mathbb{E}\|\bar{\beta} - w^*\|_\Sigma^2$ and its *calibration error* is $\mathrm{CE}(\beta) = \mathbb{E}\|\beta - \bar{\beta}\|_\Sigma^2$. Its *linear ignorance* is $K_{\mathrm{lin}}(\beta) = \min_{A \in \mathbb{R}^{d\times d}} \mathbb{E}\|A\beta - w^*\|_\Sigma^2$, and $\mathrm{CE}_{\mathrm{lin}}(\beta) = \mathcal{E}_{\mathrm{av}}(\beta) - K_{\mathrm{lin}}(\beta)$, where $\mathcal{E}_{\mathrm{av}}(\beta) = \mathbb{E}\|\beta - w^*\|_\Sigma^2$.

Every prediction of the model at a fresh input x has conditional mean $\mathbb{E}[y \mid x, \beta] = x^\top \bar{\beta}$. Hence β is calibrated, in the sense that its predictions equal the conditional mean of the label given the input and the model, exactly when $\beta = \bar{\beta}$; and $\mathrm{CE}(\beta) = \mathbb{E}(x^\top \beta - x^\top \bar{\beta})^2$ is the squared calibration error of its predictions.

Proof of Theorem 2.1. **Step 1: the decomposition.** Write $\beta - w^* = (\beta - \bar{\beta}) + (\bar{\beta} - w^*)$. The first vector is $\sigma(\beta)$-measurable and $\mathbb{E}[\bar{\beta} - w^* \mid \beta] = 0$. The cross term therefore vanishes after conditioning on β, and $\mathcal{E}_{\mathrm{av}}(\beta) = K(\beta) + \mathrm{CE}(\beta)$.

Step 2: fresh data carry no information about the target. Let $\widehat{w}$ be measurable with respect to $\mathcal{G} = \sigma(\beta_{\mathrm{w}}, U)$, where U is independent of $(w^*, \beta_{\mathrm{w}})$. For bounded measurable h and k, $\mathbb{E}[w^* h(\beta_{\mathrm{w}}) k(U)] = \mathbb{E}[\mathbb{E}[w^* \mid \beta_{\mathrm{w}}] h(\beta_{\mathrm{w}})]\,\mathbb{E}k(U)$, so $\mathbb{E}[w^* \mid \mathcal{G}] = \mathbb{E}[w^* \mid \beta_{\mathrm{w}}]$ by a monotone-class argument. Conditional expectation onto a larger sigma-field has smaller squared error. Since $\sigma(\widehat{w}) \subseteq \mathcal{G}$,

$$K(\widehat{w}) = \mathbb{E}\|\mathbb{E}[w^* \mid \widehat{w}] - w^*\|_\Sigma^2 \geq \mathbb{E}\|\mathbb{E}[w^* \mid \mathcal{G}] - w^*\|_\Sigma^2 = K(\beta_{\mathrm{w}}).$$

Step 1 then gives $\mathcal{E}_{\mathrm{av}}(\widehat{w}) \geq K(\beta_{\mathrm{w}})$, with equality only if $\widehat{w} = \mathbb{E}[w^* \mid \beta_{\mathrm{w}}]$ almost surely, and the gain identity (2.2) follows by subtracting the two decompositions.

Step 3: coordinates and linear readouts. Replacing Σ by $e_i e_i^\top$ in Steps 1 and 2 gives the coordinatewise bound. For the linear statement, let $\widehat{w} = S\beta_{\mathrm{w}}$ with S a random matrix independent of $(w^*, \beta_{\mathrm{w}})$. For each fixed value s of S and each matrix B, Bs is a deterministic matrix, so $\mathbb{E}[\|BS\beta_{\mathrm{w}} - w^*\|_\Sigma^2 \mid S = s] \geq K_{\mathrm{lin}}(\beta_{\mathrm{w}})$. Averaging over S proves $K_{\mathrm{lin}}(S\beta_{\mathrm{w}}) \geq K_{\mathrm{lin}}(\beta_{\mathrm{w}})$; taking $B = I$ proves $\mathcal{E}_{\mathrm{av}}(S\beta_{\mathrm{w}}) \geq K_{\mathrm{lin}}(\beta_{\mathrm{w}})$. □

Proof of Proposition 2.2. Fix $i > m$. Section M shows that $\mathbb{E}[w_i^* \mid \mathcal{O}] = 0$, where $\mathcal{O}$ is the sigma-field of the teacher output, fresh inputs, and independent seeds; in particular $\mathbb{E}[w_i^* \mid \beta_{\mathrm{w}}] = 0$. For any data U independent of $(w^*, \beta_{\mathrm{w}})$, Step 2 of the proof of Theorem 2.1 gives $\mathbb{E}[w_i^* \mid \beta_{\mathrm{w}}, U] = \mathbb{E}[w_i^* \mid \beta_{\mathrm{w}}] = 0$. Since $\widehat{w}$ is $\sigma(\beta_{\mathrm{w}}, U)$-measurable, its readout in coordinate i is $\mathbb{E}[w_i^* \mid \widehat{w}] = 0$, and the coordinatewise form of (2.1) reads $\mathbb{E}(\widehat{w}_i - w_i^*)^2 = \omega_i^2 + \mathbb{E}\widehat{w}_i^2$: the ignorance there is the whole prior variance, and a downstream procedure can only add calibration error. Weighting by λ_i and summing over $i > m$ gives (2.3). □

Second-order states. The remaining exact statements use only second moments. Call a pair of random vectors (w^*, β) *diagonal* if $\mathbb{E}[\beta_i \beta_j] = \mathbb{E}[w_i^* \beta_j] = \mathbb{E}[w_i^* w_j^*] = 0$ for $i \neq j$, and write $c_i = \mathbb{E}[w_i^* \beta_i]$ and $\rho_i = \mathbb{E}\beta_i^2$ for its *state*. By Cauchy–Schwarz $c_i^2 \leq \omega_i^2 \rho_i$, so $c_i = 0$ whenever $\rho_i = 0$; on such coordinates we set $c_i^2/\rho_i = 0$ and $g_i = 0$.

Lemma B.2 (The diagonal ledger)**.** *For a diagonal pair with state* (c, ρ)*, the optimal linear readout is* $\mathrm{diag}(g)$ *with* $g_i = c_i/\rho_i$*, and*

$$K_{\mathrm{lin}}(\beta) = \sum_i \lambda_i \Big(\omega_i^2 - \frac{c_i^2}{\rho_i}\Big), \qquad \mathrm{CE}_{\mathrm{lin}}(\beta) = \sum_i \lambda_i \rho_i (1 - g_i)^2.$$

Proof. For a matrix A, $\mathbb{E}\|A\beta - w^*\|_\Sigma^2 = \sum_{i,j} \lambda_i A_{ij}^2 \rho_j - 2\sum_i \lambda_i A_{ii} c_i + \sum_i \lambda_i \omega_i^2$, because all off-diagonal second moments vanish. Off-diagonal entries of A only add nonnegative terms. The diagonal entries minimize $\lambda_i(A_{ii}^2\rho_i - 2A_{ii}c_i + \omega_i^2)$ at $A_{ii} = g_i$, with value $\lambda_i(\omega_i^2 - c_i^2/\rho_i)$. The calibration term is $\lambda_i(\omega_i^2 - 2c_i + \rho_i) - \lambda_i(\omega_i^2 - c_i^2/\rho_i) = \lambda_i\rho_i(1-g_i)^2$. □

A linear fitting rule Ψ maps a design X and labels y to $\Psi(X)y$. It is *sign-equivariant* if $\Psi(XD) = D\Psi(X)$ for every diagonal sign matrix D. Minimum-norm interpolation and ridge regression are sign-equivariant, with and without the truncation sketch. Throughout this subsection second moments are finite: for a minimum-norm teacher with noisy labels this requires $m \geq n_{\mathrm{w}} + 2$, since $\mathbb{E}\,\mathrm{tr}\{(X_H X_H^\top)^{-1}\}$ is finite exactly then; ridge teachers need no condition.

Proof of Lemma 3.1. **Step 1: the teacher is diagonal.** Let the teacher be $\beta_{\mathrm{w}} = \Psi(X)(Xw^* + \xi)$, with Ψ a sign-equivariant rule that uses the head columns. Fix i and let D_i flip coordinate i. Since $Xw^* = (XD_i)(D_i w^*)$ and XD_i has the law of X, the conditional law of β_{w} given $w^* = w$ equals the law of $D_i\beta_{\mathrm{w}}$ given $w^* = D_i w$. The fit is linear in (w^*, ξ), so $f(w) = \mathbb{E}[(\beta_{\mathrm{w}})_i(\beta_{\mathrm{w}})_j \mid w^* = w]$ is a quadratic polynomial in w, with finite coefficients because second moments are finite, and $f(w) = -f(D_i w)$ for $j \neq i$. Hence every monomial of f contains w_i exactly once, and $\mathbb{E}f(w^*) = 0$ because the coordinates of w^* are independent and centered. The same argument, applied to $w_j\mathbb{E}[(\beta_{\mathrm{w}})_i \mid w^* = w]$, shows $\mathbb{E}[w_j^*(\beta_{\mathrm{w}})_i] = 0$ for $j \neq i$.

Step 2: one fresh stage. Let S be the fitting operator of a fresh full-dimensional design, so a noiseless stage returns $S\beta$. The rule is sign-equivariant and the design law is invariant under column sign changes, so DSD has the law of S. Therefore $\mathbb{E}S_{ij} = 0$ and $\mathbb{E}[S_{ik}S_{jk}] = 0$ for $i \neq j$. Independence of S from the diagonal pair (w^*, β) gives

$$\mathbb{E}[w_i^*(S\beta)_j] = \mathbb{E}S_{ji}\, c_i, \qquad \mathbb{E}[(S\beta)_i(S\beta)_j] = \sum_k \mathbb{E}[S_{ik}S_{jk}]\rho_k.$$

Both vanish for $i \neq j$, and for $i = j$ they are $\bar{F}_i c_i$ and $(W\rho)_i$. By induction every stage of a chain is diagonal, and its state is obtained by composing these maps.

Step 3: the properties of $(\bar{F}, W)$**.** A full-dimensional minimum-norm or ridge operator S is symmetric with $0 \preceq S \preceq I$, so $S_{ii} \in [0,1]$ and $\sum_i S_{ij}^2 = (S^2)_{jj} \leq S_{jj} \leq 1$. Taking expectations gives $\bar{F}_i \in [0,1]$, column sums of W at most $\bar{F}_j$, and $W_{ii} = \mathbb{E}S_{ii}^2 \geq \bar{F}_i^2$ by Jensen's inequality. □

For a minimum-norm stage, S is an orthogonal projection, so $S^2 = S$ and the column sums of W equal $\bar{F}_j$ exactly. This is the exact counterpart of the row-sum identity (J.3). A downstream fit restricted to a subset of coordinates is not symmetric, and its column sums can exceed one; the chains in this paper use full-dimensional fits.

Proof of Proposition 2.3. **(i) A teacher that is nowhere overconfident.** Consider a chain of k fresh minimum-norm or ridge stages with maps $(\bar{F}^{(t)}, W_t)$. By Lemma 3.1 its state is $c^{(k)} = \mu \odot c$ with $\mu = \prod_t \bar{F}^{(t)} \in [0,1]^d$, and, since every W_t is entrywise nonnegative with $(W_t\rho)_i \geq (\bar{F}_i^{(t)})^2\rho_i$, $\rho_i^{(k)} \geq \mu_i^2\rho_i$. Therefore

$$\mathcal{E}_{\mathrm{av}}(\beta^{(k)}) - \mathcal{E}_{\mathrm{av}}(\beta_{\mathrm{w}}) \geq \sum_i \lambda_i\big[2(1-\mu_i)c_i - (1-\mu_i^2)\rho_i\big] = \sum_i \lambda_i(1-\mu_i)\big[2c_i - (1+\mu_i)\rho_i\big].$$

If $c_i \geq \rho_i$, the bracket is at least $(1-\mu_i)\rho_i \geq 0$. Coordinates with $\rho_i = 0$ contribute nothing negative.

(ii) A teacher that is overconfident in aggregate. Fix $\eta > 0$. As the fresh sample size $n_{\mathrm{s}} \to \infty$ at fixed d, $\widehat{\Sigma} \to \Sigma$ almost surely and $S = (\widehat{\Sigma} + \eta I)^{-1}\widehat{\Sigma} \to (\Sigma + \eta I)^{-1}\Sigma$. Since $\|S\| \leq 1$, dominated

convergence gives $\bar{F} \to F(\eta)$ and $W \to \mathrm{diag}(F(\eta)^2)$, with $F_i(\eta) = \lambda_i/(\lambda_i + \eta)$. The limiting risk is $R_\infty(\eta) = \sum_i \lambda_i(\omega_i^2 - 2F_i c_i + F_i^2 \rho_i)$, and

$$R'_\infty(0) = \sum_i \lambda_i(2\rho_i - 2c_i)\Big(-\frac{1}{\lambda_i}\Big) = -2\sum_i(\rho_i - c_i).$$

If this is negative, some small $\eta > 0$ has $R_\infty(\eta) < R_\infty(0) = \mathcal{E}_{\mathrm{av}}(\beta_{\mathrm{w}})$, and so has the ridge student at every sufficiently large n_{s}. □

The two conditions do not meet. Between them, whether a given student helps depends on the geometry through $(\bar{F}, W)$; the rest of the paper computes it.

B.2 The ledger in the deterministic calculus

The leak law. Write the stage of Definition C.2 with eigenvalues λ_i (its c_i) and target second moments q_i (its u_i^2). Put $D_1 = \sum_i F_i$ and $D_2 = \sum_i F_i^2$, so that $\Xi = D_2/n$, and let $B = \sum_i \lambda_i(1 - F_i)^2 q_i$ be the energy that the mean filter removes from the stage target. Then

$$\Gamma = \frac{B + s^2}{n - D_2}, \qquad \sum_i \lambda_i v_i = \Gamma D_2, \qquad \frac{B + s^2\Xi}{1 - \Xi} = B + (B + s^2)\frac{D_2}{n - D_2}. \tag{B.1}$$

The first identity is (C.6) rewritten; the second is $\lambda_i v_i = \Gamma F_i^2$; the third is the aggregate stage risk (C.7). A stage charges the energy it removes once as bias, and then again, together with its label noise, as fluctuation scaled by $D_2/(n - D_2)$. In the main text the second charge is the *leak*. For a downstream stage it equals the toll $\mathfrak{T}$ of Section C.2.

Dictionary. In the calculus, a teacher with profiles $(F^{\mathrm{w}}, v^{\mathrm{w}})$ has state $c = F^{\mathrm{w}}\omega^2$ and $\rho = (F^{\mathrm{w}})^2\omega^2 + v^{\mathrm{w}}$. The Wiener filter g of Definition C.3 is its calibration slope, Irr in Proposition C.4 is its linear ignorance, and $\|\mathbf{1} - g\|_\pi^2$ is its calibration error. Proposition C.6 (i) is the calculus form of Proposition 2.3.

Proof of Proposition 3.3. With $v_i = \lambda_i\Gamma_{\mathrm{w}}/(\lambda_i + \tau_{\mathrm{w}})^2$ and $F_i = \lambda_i/(\lambda_i + \tau_{\mathrm{w}})$,

$$\rho_i - c_i = F_i^2\omega_i^2 + v_i - F_i\omega_i^2 = \frac{\lambda_i^2\omega_i^2 + \lambda_i\Gamma_{\mathrm{w}} - \lambda_i(\lambda_i + \tau_{\mathrm{w}})\omega_i^2}{(\lambda_i + \tau_{\mathrm{w}})^2} = \frac{\lambda_i(\Gamma_{\mathrm{w}} - \tau_{\mathrm{w}}\omega_i^2)}{(\lambda_i + \tau_{\mathrm{w}})^2}.$$

This proves the pointwise statement. Put $S_1 = \sum_{i \le m} \lambda_i\omega_i^2/(\lambda_i + \tau_{\mathrm{w}})^2$ and $S_2 = \sum_{i \le m} \lambda_i/(\lambda_i + \tau_{\mathrm{w}})^2$ (the A and B of (F.2)). Summing and using (F.3) gives $\sum_{i \le m}(\rho_i - c_i) = \Gamma_{\mathrm{w}} S_2 - \tau_{\mathrm{w}} S_1 = s_{\mathrm{w}}^2/\tau_{\mathrm{w}}$ with $s_{\mathrm{w}}^2 = \sigma^2 + A_m^{\mathrm{av}}$, as in the proof of Remark C.7 (b). It is positive exactly when $s_{\mathrm{w}}^2 > 0$; with $\sigma^2 = 0$ and all signal observed it vanishes. □

Proof of Proposition 4.1. **Step 1: parametrize by the threshold.** For a minimum-norm student, $n = \sum_i F_i$ with $F_i = \lambda_i/(\lambda_i + \tau)$, a smooth decreasing bijection from $\tau \in (0, \infty)$ onto $(0, d)$. Its expansion $n = d - \tau\sum_i \lambda_i^{-1} + O(\tau^2)$ gives $\tau = (d - n)/\operatorname{tr}(\Sigma^{-1}) + O((d - n)^2)$.

Step 2: expand the leak. For $\eta = 0$, $n - D_2 = \sum_i F_i(1 - F_i) = \tau\sum_i \lambda_i/(\lambda_i + \tau)^2$ and $B = \tau^2\sum_i \lambda_i\rho_i/(\lambda_i + \tau)^2$, so by (B.1)

$$\Gamma = \tau\,\frac{\sum_i \lambda_i\rho_i/(\lambda_i + \tau)^2}{\sum_i \lambda_i/(\lambda_i + \tau)^2} = \tau\,\frac{\sum_i \rho_i/\lambda_i}{\sum_i 1/\lambda_i} + O(\tau^2).$$

Since $D_2 = d + O(\tau)$, the leak is $\Gamma D_2 = 2\ell\tau + O(\tau^2)$.

Step 3: expand the filter. The filter part $\sum_i \lambda_i(\omega_i^2 - 2F_i c_i + F_i^2\rho_i)$ equals $\mathcal{R}_w$ at $\tau = 0$. Its derivative there is $\sum_i \lambda_i(2\rho_i - 2c_i)(-1/\lambda_i) = -2\mathfrak{O}$. Adding Steps 2 and 3 and substituting Step 1 proves (4.1).

Step 4: isotropy and uniformity. At $\Sigma = I_d$, $\ell = \frac{1}{2}\sum_i \rho_i$ and $\mathcal{R}_w - r^2 = \sum_i(\rho_i - 2c_i) = 2(\mathfrak{O} - \ell)$. At fixed d the remainder is $O(\tau^2)$, and τ and $1 - n_{\mathrm{s}}/d$ are proportional to first order. For spectra in

$[M^{-1}, M]$ three facts make the constant uniform in d. First, $\sum_i \lambda_i/(\lambda_i+\tau)^2 \geq d/(M(M+\tau)^2)$, so $\Gamma = \tau h(\tau)$ with h smooth and bounded, and $1 - n_{\mathrm{s}}/d = \tau\, d^{-1} \sum_i (\lambda_i + \tau)^{-1}$ inverts with uniformly bounded derivatives. Second, $\sum_i \rho_i$ and $\sum_i |c_i|$ are bounded, because the teacher's energy is bounded when $1 - \Xi_{\mathrm{w}}$ is bounded below (Lemma G.1). Third, the filter part has second τ-derivative at most $C(M) \sum_i (|c_i| + \rho_i)$. The remainder is therefore $O((1 - n_{\mathrm{s}}/d)^2)$ uniformly in d, and the expansion passes to the limiting formula as $d/n_{\mathrm{s}} \downarrow 1$. □

Corollary B.3 (An interior budget from the clone criterion)**.** *Assume the bounded-spectrum hypotheses of Theorem 3.2. Then the teacher's overconfidence $\mathfrak{O}_d$ and clone leak rate ℓ_d converge to limits $\bar{\mathfrak{O}}$ and $\bar{\ell}$, integrals against the limiting law of $(i/d, \lambda_i, d\omega_i^2)$. Suppose $\bar{\mathfrak{O}} > \bar{\ell}$ and $\bar{\mathcal{E}}_w < r^2$. Then the limiting risk curve attains its minimum over $\bar{\kappa}_{\mathrm{s}} \in (1, \infty)$ at an interior point, strictly below both $\bar{\mathcal{E}}_w$ and r^2, and $\Pr(\mathcal{E}_{\mathrm{s}} < \mathcal{E}_{\mathrm{w}}) \to 1$ at every ratio $\bar{\kappa}_{\mathrm{s}}$ with $\bar{\mathcal{E}}_s(\bar{\kappa}_{\mathrm{s}}) < \bar{\mathcal{E}}_w$, in particular at every limiting minimizer.*

Proof. The uniform expansion gives a limiting budget with $\bar{\mathcal{E}}_s < \bar{\mathcal{E}}_w$, and Corollary D.8 (a) supplies the interior minimum. At any fixed ratio with $\bar{\mathcal{E}}_s(\bar{\kappa}_{\mathrm{s}}) < \bar{\mathcal{E}}_w$, Theorem 3.2 applies to both risks, so $\mathcal{E}_{\mathrm{s}} - \mathcal{E}_{\mathrm{w}} \to \bar{\mathcal{E}}_s - \bar{\mathcal{E}}_w < 0$ in probability. □

Proof of the reading in Theorem 4.2. Use the large-L limits displayed in Section I.4, with $t = n_{\mathrm{s}} - k$, $D = d - k$, $\mu_{\mathrm{s}} = D/t$ and $\mu_{\mathrm{w}} = (m-k)/(n_{\mathrm{w}} - k)$. The teacher's bulk error is $(n_{\mathrm{w}} - k)\Gamma_{\mathrm{w}}/\mu_{\mathrm{w}} = \sigma^2/(\mu_{\mathrm{w}} - 1) = \sigma^2(n_{\mathrm{w}} - k)/(m - n_{\mathrm{w}}) = V$. The displayed student leak level is $\Gamma_{\mathrm{s}} = (\mu_{\mathrm{s}} - 1)V/(\mu_{\mathrm{s}} t) = (1 - t/D)V/t$. The student's bulk error is $V/\mu_{\mathrm{s}}^2 + \Gamma_{\mathrm{s}} t/\mu_{\mathrm{s}} = (t/D)V$, and the error it adds to the k spike directions is $k\Gamma_{\mathrm{s}} = (k/t)(1 - t/D)V$. The teacher's spike error $k\Gamma_{\mathrm{w}}$ passes unchanged. Subtracting,

$$\bar{\mathcal{E}}_w - \bar{\mathcal{E}}_s = V\Big(1 - \frac{t}{D}\Big)\Big(1 - \frac{k}{t}\Big),$$

which is (4.2). On $k < t < D$ its logarithmic derivative is $-(D-t)^{-1} + (t-k)^{-1} - t^{-1}$, which has the sign of $kD - t^2$ (Remark D.7). The gain therefore increases up to $t^2 = kD$ and decreases after it; there $t/D = k/t = \sqrt{k/D}$ and the gain is $V(1 - \sqrt{k/D})^2$. The kept amount $(t/D)V$ and the leaked amount $(k/t)(1 - t/D)V$ are not equal; the fractions are. □

The clone criterion in the spiked family. In the same limit the clone criterion of Proposition 4.1 reads $d > 2k$. The teacher's observed bulk has $m - k$ coordinates of energy $V/(m-k)$ and zero signal, the spikes contribute $O(1/L)$ to $\mathfrak{O}$ and to $\sum_i \rho_i/\lambda_i$, and $\sum_i \lambda_i^{-1} = D + O(1/L)$. Hence $\mathfrak{O} \to V$ and $\ell \to dV/(2D)$, and $\mathfrak{O} > \ell$ exactly when $2D > d$, that is, when $d > 2k$.

Proof of Proposition 6.1. Write $s = a + b > 1$. Constants below depend on $(a, b, r^2, \sigma^2, \psi, \bar{\kappa_{\mathrm{w}}}, \bar{\kappa_{\mathrm{s}}})$ only.

Step 1: both thresholds and noise levels. Lemma K.4, applied on the head with $n_{\mathrm{w}} = m/\bar{\kappa_{\mathrm{w}}}$ and on all coordinates with $n_{\mathrm{s}} = d/\bar{\kappa_{\mathrm{s}}}$, gives $\tau_{\mathrm{w}} \asymp n_{\mathrm{w}}^{-a}$ and $1 - \Xi_{\mathrm{w}}, 1 - \Xi_{\mathrm{s}} \geq c_0 > 0$. The teacher numerator in (C.6) is at most $r^2 + \sigma^2 + r^2$, so $\Gamma_{\mathrm{w}} \leq C/n_{\mathrm{w}}$. The student's removed energy is at most its target energy $\sum_i \lambda_i \rho_i \leq r^2 + \Gamma_{\mathrm{w}} n_{\mathrm{w}} \Xi_{\mathrm{w}}$, so also $\Gamma_{\mathrm{s}} \leq C/n_{\mathrm{s}} \leq C'/n_{\mathrm{w}}$.

Step 2: the student state is a noisy observation of the target. With $c^{\mathrm{s}} = F^{\mathrm{s}} F^{\mathrm{w}} \omega^2$ and $\rho^{\mathrm{s}} = (F^{\mathrm{s}})^2[(F^{\mathrm{w}})^2 \omega^2 + v^{\mathrm{w}} + \Gamma_{\mathrm{s}}/\lambda]$, one has

$$\lambda_i\Big(\omega_i^2 - \frac{(c_i^{\mathrm{s}})^2}{\rho_i^{\mathrm{s}}}\Big) = \frac{\lambda_i \omega_i^2 N_i}{\lambda_i \omega_i^2 + N_i}, \qquad N_i = \Gamma_{\mathrm{w}} + \Gamma_{\mathrm{s}}(1 + \tau_{\mathrm{w}}/\lambda_i)^2,$$

for $i \leq m$; for $i > m$ the term is $\lambda_i \omega_i^2$. Each head term is at most $\min(\lambda_i \omega_i^2, N_i)$.

Step 3: sum. Let $i_1 = \min(\tau_{\mathrm{w}}^{-1/a}, m) \asymp n_{\mathrm{w}}$. For $i \leq i_1$ the direction is observed and $\lambda_i \geq \tau_{\mathrm{w}}$, so $N_i \leq \bar{\Gamma} = \Gamma_{\mathrm{w}} + 4\Gamma_{\mathrm{s}} \leq C/n_{\mathrm{w}}$. Every other term, observed or not, is at most $\lambda_i \omega_i^2 = G i^{-s}$, with $G \leq r^2$. Hence

$$K_{\mathrm{lin}}^{\mathrm{s}} \leq \sum_{i \geq 1} \min(G i^{-s}, \bar{\Gamma}) + \sum_{i > i_1} G i^{-s} \leq C\, \bar{\Gamma}^{(s-1)/s} + C\, i_1^{1-s} \leq C\, n_{\mathrm{w}}^{-(s-1)/s}.$$

The same bound holds for the teacher with $\Gamma_{\mathrm{s}} = 0$. Theorem D.14 (ii) gives $\bar{\mathcal{E}}_s \to \sigma^2 C_{\mathrm{floor}} > 0$, so the calibration error $\bar{\mathcal{E}}_s - K^{\mathrm{s}}_{\mathrm{lin}}$ has the same limit. □

The lower bound $K^{\mathrm{s}}_{\mathrm{lin}} \geq K^{\mathrm{w}}_{\mathrm{lin}} \geq \sum_{i \leq m} \lambda_i \omega_i^2 \Gamma_{\mathrm{w}} / (\lambda_i \omega_i^2 + \Gamma_{\mathrm{w}})$, together with $\Gamma_{\mathrm{w}} \geq \sigma^2 / n_{\mathrm{w}}$, shows that the rate $n_{\mathrm{w}}^{-(a+b-1)/(a+b)}$ is exact for these formulas. It is the exponent of the Bayes bound (6.1).

Two clocks of forgetting. Linear ignorance never falls along a chain (Theorem 2.1), and Proposition J.3 drives the energy to zero when the condition number is bounded. The exact calculus shows that knowledge is lost too, on a separate and possibly much slower clock.

Proposition B.4 (Two clocks)**.** *Consider a chain of identical fresh full-dimensional stages, each either minimum-norm with $n < d$ or ridge with $\eta > 0$, with exact maps $(\bar{F}, W)$ as in Lemma 3.1, and a teacher state with $\rho \neq 0$. Every entry of W is positive, so W has a Perron root μ, and $\mu > \max_i W_{ii} \geq \max_i \bar{F}_i^2$. As $T \to \infty$, the energy $\sum_i \rho_i^{(T)}$ is of exact order μ^T, and the knowledge share of each direction satisfies $c_i^{(T)2} / \rho_i^{(T)} = O\big((\bar{F}_i^2/\mu)^T\big)$. Hence $\mu < 1$, $K_{\mathrm{lin}}(\beta^{(T)}) \to r^2$ and $\mathrm{CE}_{\mathrm{lin}}(\beta^{(T)}) \to 0$: every such chain is eventually uninformed.*

Proof. For $n < d$ or $\eta > 0$, each entry of the fitting operator of a Gaussian design is a nonconstant analytic function of the design, so $W_{ij} = \mathbb{E} S_{ij}^2 > 0$. A positive matrix is primitive, and $W^T \rho = \mu^T (v\, u^\top \rho)(1 + o(1))$ entrywise, with positive Perron vectors v, u and $u^\top \rho > 0$. The Perron root of a positive matrix of size at least two exceeds every diagonal entry, and $W_{ii} \geq \bar{F}_i^2$ by Lemma 3.1; also μ is at most the largest column sum, which is at most $\max_j \bar{F}_j < 1$. Since $c^{(T)} = \bar{F}^T \odot c$, the share $c_i^{(T)2} / \rho_i^{(T)}$ is $\bar{F}_i^{2T} c_i^2 / (\mu^T v_i u^\top \rho)(1 + o(1))$. Summing the shares against λ gives $K_{\mathrm{lin}} \to \sum_i \lambda_i \omega_i^2 = r^2$, and the risk tends to r^2 because $c^{(T)}, \rho^{(T)} \to 0$. □

Between the two clocks the chain is underconfident. When $\log(1/\mu) \gg \log(\mu / \bar{F}_i^2)$, direction i keeps its share c_i^2 / ρ_i of knowledge long after its energy has gone, so its calibration slope $g_i = (c_i^2/\rho_i)/c_i$ becomes large. In the calculus the stage map is $\mathrm{diag}(F^2) + uw^\top$ with $u, w > 0$ (the forward form of (J.1)), and the same argument applies. The two clocks can differ enormously. Table 1 evaluates them, with the ledger along the chain, for three chains whose stages each use $n = d/2$ fresh inputs.

Table 1: Two clocks of forgetting in the deterministic calculus. Half-lives are $\log 2 / \log(1/\mu)$ for energy and $\log 2 / \log(\mu / \max_i \bar{F}_i^2)$ for the knowledge of the top direction. Power law: $d = 1000$, $(a, b) = (1.2, 1)$, $(m, n_{\mathrm{w}}) = (300, 100)$, $\sigma^2 = 0.25$, $r^2 = 1$. Spiked: $d = 1000$, $k = 50$ spikes at $L = 100$ carrying all signal, same teacher budgets. Isotropic: $\Sigma = I$, signal on the teacher's head, same budgets. Evaluated spectrally and checked against direct iteration to 4×10^{-14} by `audit/check_ledger.py`.

Chain	Energy half-life	Knowledge half-life	Depth	Risk	K_{lin}	$\mathrm{CE}_{\mathrm{lin}}$
Power law	467.5	1.8×10^8	0	0.2353	0.0632	0.1721
			47	0.0665	0.0640	0.0025
			5000	0.9676	0.2096	0.7580
			10^8	1.0000	0.5471	0.4529
			10^9	1.0000	0.9859	0.0141
Spiked	31.5	1.8×10^4	0	0.3364	0.2267	0.1097
			20	0.2320	0.2318	0.0002
			200	0.8024	0.2371	0.5653
			10^4	1.0000	0.4732	0.5268
			10^5	1.0000	0.9825	0.0175
Isotropic	1.0	1.0	0	0.7917	0.7576	0.0341
			1	0.8958	0.8135	0.0823
			10	0.9998	0.9992	0.0006

In the two anisotropic chains the risk first falls towards the teacher's linear ignorance, reaching its minimum at depths 47 and 20, while the ignorance barely moves; the calibration error then rises

as energy drains away, and only much later does the knowledge go. Under isotropy the two clocks coincide, the first stage already increases the calibration error, and knowledge and energy vanish together. These are evaluations of the calculus at fixed parameters, not estimator theorems.

B.3 Equal-budget depth on two-level spectra

The certified depth advantage in (5.6) uses a 4000-level power law. To test whether a simpler geometry produces one, we optimized the deterministic formulas on two-level spectra with $d = 400$: a fraction f of coordinates at height h carries all signal, the rest sit at one, and $\sigma^2 = 0.25$, $r^2 = 1$. The single stage uses the whole budget N with its ridge optimized over a grid and a bounded search, and includes the minimum-norm and zero predictors. The two-stage chain optimizes the split and both ridges by Nelder–Mead from eight of 64 scored starting points. Table 2 lists every configuration.

Table 2: Best single-stage ridge risk $R_1^\star$ and best two-stage risk $R_2^\star$ at equal total budget, deterministic calculus, two-level spectra. Relative difference $(R_1^\star - R_2^\star)/R_1^\star$; positive favors depth. Evaluated by `audit/check_ledger.py`. Numerical optimization, not a certificate.

h	f	m/d	m/n_{w}	N/d	$R_1^\star$	$R_2^\star$	rel. diff.
3	0.05	0.5	2	0.5	0.4225	0.7397	−0.751
3	0.05	0.5	2	1.5	0.3266	0.3827	−0.172
3	0.05	0.8	3	1.0	0.3854	0.5532	−0.436
3	0.05	0.3	3	0.5	0.5135	0.7954	−0.549
3	0.15	0.5	2	0.5	0.5136	0.8050	−0.567
3	0.15	0.5	2	1.5	0.4180	0.4737	−0.133
3	0.15	0.8	3	1.0	0.4585	0.6264	−0.366
10	0.05	0.5	2	0.5	0.1956	0.3343	−0.709
10	0.05	0.5	2	1.5	0.1627	0.1705	−0.048
10	0.05	0.8	3	1.0	0.1589	0.1972	−0.241
10	0.05	0.3	3	0.5	0.2729	0.4131	−0.514
10	0.15	0.5	2	0.5	0.3524	0.5239	−0.487
10	0.15	0.5	2	1.5	0.3161	0.3267	−0.034
10	0.15	0.8	3	1.0	0.2823	0.3283	−0.163
30	0.05	0.5	2	0.5	0.1353	0.1668	−0.233
30	0.05	0.5	2	1.5	0.1186	0.1184	+0.002
30	0.05	0.8	3	1.0	0.0961	0.1022	−0.064
30	0.05	0.3	3	0.5	0.2267	0.2426	−0.070
30	0.15	0.5	2	0.5	0.3479	0.4090	−0.176
30	0.15	0.5	2	1.5	0.3217	0.3266	−0.015
30	0.15	0.8	3	1.0	0.2578	0.2671	−0.036

Only one configuration favors depth, by 0.16%. The ledger explains the pattern. A single ridge at the full budget can already realize a strong bulk filter with a small leak, whereas the first stage of a chain leaks with half the budget, and Theorem 2.1 forbids any later stage from recovering what it leaked. With the spikes far above the bulk, the leak into the signal directions is nearly unfiltered by later stages. These examples suggest that depth pays when a single resolvent is far from the teacher's calibration slope, as in the power-law certificate; we have no general criterion for it.

B.4 A second proof of the chain approximation in expectation

Lemma 3.1 makes expected chain risks exact linear algebra: $\mathbb{E}\,\mathcal{E}(\beta^{(t)}) = \sum_i \lambda_i(\omega_i^2 - 2c_i^{(t)} + \rho_i^{(t)})$ with $c^{(t)} = (\prod_{s\le t} \bar{F}^{(s)}) \odot c^{\mathrm{w}}$ and $\rho^{(t)} = W_t \cdots W_1 \rho^{\mathrm{w}}$. The calculus replaces $\bar{F}^{(s)}$ by $F^{(s)}$ and W_s by $\widehat{W}_s = \mathcal{T}_s^\top$, where $\mathcal{T}_s$ is the backward operator (J.1) (written there with eigenvalues c_j). Only a single-stage comparison is needed.

Lemma B.5 (Single-stage law in expectation)**.** *Under the bounds of Lemma G.3 for a noiseless stage, there is $\epsilon_d \to 0$, depending only on those bounds, such that for all $q, x \in \mathbb{R}^d$ and $\rho \in \mathbb{R}^d_{\ge 0}$,*

$$\Big|\sum_i q_i(\bar{F}_i - F_i)x_i\Big| \le \epsilon_d \|q\|_\infty \|x\|_1, \qquad \big|q^\top (W - \mathcal{T}^\top)\rho\big| \le \epsilon_d \|q\|_\infty \|\rho\|_1.$$

Proof. By homogeneity take $\|q\|_\infty \le 1$ and $\|x\|_1 = \|\rho\|_1 = 1$. For the quadratic bound, let ε be a Rademacher vector independent of the design and $u = \varepsilon \odot \sqrt{\rho}$, so $\|u\| = 1$. Lemma G.3 and (J.1) give $\sum_i q_i (Su)_i^2 - \sum_j (\mathcal{T}q)_j u_j^2 \to 0$ in probability, uniformly over the unit ball. Both terms are bounded by one, so the convergence holds in expectation, uniformly. Averaging over ε removes all cross terms and leaves $q^\top W \rho - (\mathcal{T}q)^\top \rho$. For the linear bound, apply the linear part of Lemma G.3 to $z = q \odot \sqrt{|x|}$ and $\beta = \operatorname{sign}(x) \odot \sqrt{|x|}$; its expectation is $\sum_i q_i \bar{F}_i x_i$ because $\mathbb{E}S$ is diagonal. (Sign invariance of S alone already removes the cross terms; the Rademacher vector is a convenience.) □

Proposition B.6 (Expected chain risks, uniformly in depth)**.** *Under the hypotheses of Theorem J.2,* $\sup_{t \ge 0} |\mathbb{E}\,\mathcal{E}(\beta^{(t)}) - \mathcal{R}_t^{(d)}| \to 0$.

Proof. Let $\hat{c}^{(t)}, \hat{\rho}^{(t)}$ be the calculus states, and telescope:

$$\rho^{(t)} - \hat{\rho}^{(t)} = \sum_{s=1}^{t} W_t \cdots W_{s+1} (W_s - \widehat{W}_s) \hat{\rho}^{(s-1)} + W_t \cdots W_1 (\rho^{\mathrm{w}} - \hat{\rho}^{\mathrm{w}}).$$

Test against λ. The vectors $q_s = W_{s+1}^\top \cdots W_t^\top \lambda$ satisfy $\|q_s\|_\infty \le M$, since every W is nonnegative with column sums at most one (Lemma 3.1). The calculus contracts: (J.3) and Lemma G.1 give every calculus filter $F_i \le \chi := M^2/(M^2 + \delta_0) < 1$ and hence $\|\hat{\rho}^{(s)}\|_1 \le \chi^s \|\hat{\rho}^{\mathrm{w}}\|_1$. Lemma B.5 bounds the sth term by $\epsilon_d M \chi^{s-1} \|\hat{\rho}^{\mathrm{w}}\|_1$, a summable sequence.

The teacher term is $q_0^\top (\rho^{\mathrm{w}} - \hat{\rho}^{\mathrm{w}})$ with q_0 depending on t, so it needs convergence uniformly over $\|q_0\|_\infty \le M$, that is, $\|\rho^{\mathrm{w}} - \hat{\rho}^{\mathrm{w}}\|_1 \to 0$. This follows from the teacher's aggregate law (Section G.2) applied with the deterministic weight $q = \operatorname{sign}(\rho^{\mathrm{w}} - \hat{\rho}^{\mathrm{w}})$, whose failure probability is uniform over $\|q\|_\infty \le 1$, together with uniform integrability (Lemma G.5) and exact prior averaging of the affine teacher formula. The linear terms telescope in the same way, with $|\prod_s \bar{F}^{(s)}| \le 1$, $\prod_s F^{(s)} \le \chi^s$, and the teacher's linear law applied to $z = \operatorname{sign}(c^{\mathrm{w}} - \hat{c}^{\mathrm{w}}) \odot \lambda$. The resulting bound is uniform in t. □

This proves convergence of expected risks. The statement in probability and in L^1, Theorem J.2, needs the conditioning argument of Section J.1. The expected version needs only the exact linear structure of Lemma 3.1, the single-stage law of Lemma G.3 applied one stage at a time, the teacher's law, and uniform integrability (Lemma G.5).

C Detailed setup and spectral calculus

Data. Covariates $x \sim \mathcal{N}(0, \Sigma)$ in $\mathbb{R}^d$ with $\Sigma = \operatorname{diag}(\lambda_1 \ge \cdots \ge \lambda_d > 0)$ (diagonal without loss of generality: all estimators below are rotation-equivariant), responses $y = \langle w^*, x \rangle + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$ independent of x. The ground truth w^* is random with independent coordinates,

$$\mathbb{E}[w_i^*] = 0, \qquad \mathbb{E}[(w_i^*)^2] = \omega_i^2, \tag{C.1}$$

independent of everything else.

Two concrete profiles play a special role: the *isotropic prior*, ω_i^2 constant across coordinates and normalized to signal power r^2, i.e. $\omega_i^2 = r^2 / \sum_j \lambda_j$ (so that $\mathbb{E}\|w^*\|^2 = r^2 d / \sum_j \lambda_j$, which coincides with r^2 only when the eigenvalues average to 1), and the *power-law family*

$$\lambda_i = i^{-a} \ (a > 1), \qquad \omega_i^2 = G\, i^{-b} \ (b \ge 0), \qquad G \text{ s.t. } \textstyle\sum_i \lambda_i \omega_i^2 = r^2, \tag{C.2}$$

so that in both families r^2 is the signal power $\mathbb{E}\,\langle w^*, x \rangle^2$ — throughout the paper r^2 denotes that, never $\mathbb{E}\|w^*\|^2$ — and b measures the alignment of the signal with the leading spectrum (the *source* exponent). The excess risk of a parameter w is $\mathcal{E}(w) = \|\Sigma^{1/2}(w - w^*)\|^2$, so that the prediction risk is $\mathcal{E}(w) + \sigma^2$.

The pipeline. Three estimators are trained in sequence:

1. *Weak teacher.* The teacher observes the sketched features $Sx \in \mathbb{R}^m$, $m < d$ (proportional: $m/d \to \psi$, Assumption C.1). Our main results take S to be the *spectral truncation sketch* $S = [\, I_m \;\; 0\,]$, i.e. the teacher sees the top-m principal coordinates; Section L treats the Gaussian sketch. On n_{w} labeled samples $(X, y) \in \mathbb{R}^{n_{\mathrm{w}} \times d} \times \mathbb{R}^{n_{\mathrm{w}}}$ it fits, with $X_H := XS^\top$,

$$\widehat{w}_{\mathrm{w}} \;=\; \arg\min_{v \in \mathbb{R}^m} \|v\| \quad \text{s.t.} \quad X_H v = y \tag{C.3}$$

(minimum-norm interpolation, MNI; the ridge variant replaces the constraint by the penalized objective with parameter λ_{w}).

Its effective d-dimensional parameter is $\beta_{\mathrm{w}} := S^\top \widehat{w}_{\mathrm{w}}$, i.e. $(\widehat{w}_{\mathrm{w}}, 0_{d-m})$.

2. *Strong student.* On n_{s} *fresh unlabeled* samples $\widetilde{X} \in \mathbb{R}^{n_{\mathrm{s}} \times d}$ the teacher produces pseudo-labels $\widehat{y} = \widetilde{X}\beta_{\mathrm{w}}$ — a *noiseless linear function* of the student's own inputs, the algebraic seed of everything that follows — and the student fits

$$\widehat{w}_{\mathrm{s}} \;=\; \arg\min_{w \in \mathbb{R}^d} \|w\| \quad \text{s.t.} \quad \widetilde{X}w = \widehat{y}, \qquad \text{i.e.} \quad \widehat{w}_{\mathrm{s}} = \widetilde{P}\,\beta_{\mathrm{w}}, \tag{C.4}$$

where $\widetilde{P}$ is the orthogonal projection onto the row span of $\widetilde{X}$: *the MNI student is exactly the random-data projection of its teacher's parameter.*

The ridge student is $\widehat{w}_{\mathrm{s}}(\lambda_{\mathrm{s}}) = (\widehat{\Sigma} + \lambda_{\mathrm{s}} I)^{-1} \widehat{\Sigma}\, \beta_{\mathrm{w}}$ with $\widehat{\Sigma} = \widetilde{X}^\top \widetilde{X}/n_{\mathrm{s}}$.

3. *Ceiling.* $\widehat{w}_{\mathrm{c}}$ is the same estimator as (C.4) trained on the true labels $\widetilde{y} = \widetilde{X}w^* + \widetilde{\xi}$ of the same fresh inputs.

A k-hop *chain* iterates step 2: for $t = 1, \ldots, k-1$, model $t+1$ interpolates pseudo-labels of model t on a fresh batch of size n_{t+1} (with optional per-hop ridge; in Section D.4 the batch sizes $(n_t)_{t \le k}$ share a total budget $\sum_t n_t = N$), the capability ladder of bootstrapped W2S supervision.

Assumption C.1 (Proportional regime)**.** $d \to \infty$ with $m/d \to \psi \in (0, 1)$, $\kappa_{\mathrm{w}} = m/n_{\mathrm{w}} \to \bar{\kappa_{\mathrm{w}}} > 1$, $\kappa_{\mathrm{s}} = d/n_{\mathrm{s}} \to \bar{\kappa_{\mathrm{s}}} > 1$ (both bounded away from 1); the empirical joint distribution of $(i/d,\ \lambda_i,\ d\,\omega_i^2)$ converges weakly with uniformly integrable second moments in the last two coordinates; the label noise is non-degenerate, $\sigma^2 > 0$; and the spectra are bounded: $\lambda_i \in [1/M, M]$ for some $M \ge 1$. (The index coordinate i/d matters because the sketch selects the *top* coordinates: under spectral ties, the value-pair law alone does not determine which signal mass the teacher sees.)

The boundedness clause covers the isotropic and spiked families but not the power law (C.2), whose "active" condition number genuinely grows with d. Hence a division of labor, worth stating once and for all. The convergence theorems — Theorem D.1 and Theorem D.9 — are proved under Assumption C.1, for spectra of bounded condition number, by the noise-floor extension and aggregate composition argument (Section G). The limit formulas they converge to are a separate, purely deterministic object: the channel calculus of Section C.1, which is defined for *any* summable spectrum and in particular along the power law.

Every power-law statement in this paper — the imitation rate, the noise plateau, the chain threshold, the tuned-ridge exponent — is an exact statement about that calculus, evaluated at $\lambda_i = i^{-a}$, and each says so where it is stated (Remarks J.4, I.4 and K.16). The proof-grade ledger opening Section F records what is established at which grade.

The metric $\mathrm{PGR} := (\mathcal{E}_{\mathrm{w}} - \mathcal{E}_{\mathrm{s}})/(\mathcal{E}_{\mathrm{w}} - \mathcal{E}_{\mathrm{c}})$ is the performance gap recovered (Burns et al., 2023), well defined when $\mathcal{E}_{\mathrm{w}} \neq \mathcal{E}_{\mathrm{c}}$. Since $\mathrm{PGR} - 1 = (\mathcal{E}_{\mathrm{c}} - \mathcal{E}_{\mathrm{s}})/(\mathcal{E}_{\mathrm{w}} - \mathcal{E}_{\mathrm{c}})$, in the standard situation $\mathcal{E}_{\mathrm{w}} > \mathcal{E}_{\mathrm{c}}$ (weak teacher, stronger ceiling) $\mathrm{PGR} > 1$ means the student on pseudo-labels beats the same architecture trained on *true* labels; when $\mathcal{E}_{\mathrm{w}} < \mathcal{E}_{\mathrm{c}}$ the inequality reverses, and super-ceiling behavior (Corollary D.5) is read from $\mathcal{E}_{\mathrm{s}} < \mathcal{E}_{\mathrm{c}}$ directly rather than from PGR.

C.1 The interpolation channel

All our formulas are generated by a single primitive. Consider one *stage*: an MNI or ridge fit with feature covariance $C = \mathrm{diag}(c_1, \ldots, c_p)$, sample size n, ridge $\lambda \ge 0$ (MNI: $\lambda = 0$, $n < p$), target parameter u, and independent label noise of variance s^2.

Definition C.2 (Channel parameters)**.** Let $\tau = \tau(C, n, \lambda) > 0$ be the unique solution of

$$\lambda \;=\; \tau\Big(1 - \tfrac{1}{n}\textstyle\sum_i \frac{c_i}{c_i+\tau}\Big), \qquad (\text{MNI: } \textstyle\sum_i \frac{c_i}{c_i+\tau} = n), \tag{C.5}$$

and set

$$F_i = \frac{c_i}{c_i+\tau}, \qquad \Xi = \frac{1}{n}\sum_i \frac{c_i^2}{(c_i+\tau)^2} \in (0,1), \qquad \Gamma = \frac{\tau^2 \sum_i c_i u_i^2/(c_i+\tau)^2 \;+\; s^2}{n\,(1-\Xi)}. \tag{C.6}$$

The *channel* maps the target profile (u_i^2) to the output law $\widehat{w}_i \approx F_i u_i + \sqrt{v_i}\, g_i$ with $v_i = c_i\Gamma/(c_i+\tau)^2$ and g_i i.i.d. standard Gaussians.

Here "$\approx$" is the distributional surrogate of Han & Xu (2023): for globally Lipschitz test functionals, estimator and surrogate are asymptotically exchangeable; quadratic functionals are handled separately by localization and tail control, uniformly over targets u in a norm ball and uniformly in λ down to 0. Lemma F.1 makes this precise, and Section F verifies that Definition C.2 is an exact restatement of their theorem, with the dictionary between their fixed point γ^2 and our Γ.

Aggregate stage risk follows by integrating the surrogate:

$$\mathbb{E}\,\|C^{1/2}(\widehat{w}-u)\|^2 \;\longrightarrow\; \underbrace{\frac{\tau^2\sum_i c_i u_i^2/(c_i+\tau)^2}{1-\Xi}}_{=:\,B\ (\text{bias})} \;+\; \underbrace{\frac{s^2\,\Xi}{1-\Xi}}_{=:\,V\ (\text{variance})}, \tag{C.7}$$

recovering the classical formulas (Hastie et al., 2022). Three features of Definition C.2 do the real work later: the fluctuation profile $v_i \propto c_i/(c_i+\tau)^2$ is *spectrum-colored* (what a downstream student can filter); the level Γ is driven by the stage's *own bias* as well as its label noise (what makes even noiseless pipelines stochastic); and the denominator $n(1-\Xi)$ diverges as the stage approaches its interpolation threshold (what a chain hop pays for existing).

C.2 The Wiener form of the risk

Everything this paper proves is a statement about *which filters a pipeline can reach*, measured against one fixed target. Making that sentence precise costs three definitions and buys most of the proofs.

Fix a teacher, i.e. a shrinkage profile F^{w} and a fluctuation profile v^{w}, and let a downstream model apply a further filter $F = (F_i)_{i\le d}$ coordinatewise, so that its parameter has mean $F_i F_i^{\mathrm{w}} w_i^*$ and fluctuation variance $F_i^2 v_i^{\mathrm{w}}$. The resulting risk is the quadratic

$$R(F) \;=\; \sum_i \lambda_i \Big[\big(1 - F_i F_i^{\mathrm{w}}\big)^2 \omega_i^2 \;+\; F_i^2\, v_i^{\mathrm{w}}\Big]. \tag{C.8}$$

This *filter objective* charges the downstream model for how it reshapes what the teacher emitted, not for the fluctuation it injects on its own account.

That second charge — the *toll* $\sum_i \lambda_i v_i^{\mathrm{s}}$ of Definition C.2, which is positive for a nonzero target in the stated overparameterized regime — is accounted separately, because their budget dependence is different and neither term should be presumed globally monotone: every population-level statement below is about (C.8), every finite-budget statement is (C.8) plus a toll.

Definition C.3 (Teacher data, Wiener filter, and weight)**.** Given $(F^{\mathrm{w}}, v^{\mathrm{w}})$ and the signal profile ω^2, set

$$c_i \;:=\; F_i^{\mathrm{w}}\,\omega_i^2 \quad (\text{signal passed}), \qquad \rho_i \;:=\; (F_i^{\mathrm{w}})^2\omega_i^2 + v_i^{\mathrm{w}} \quad (\text{energy emitted}), \tag{C.9}$$

and define the *Wiener filter* and the *Wiener weight*

$$g_i \;:=\; \frac{c_i}{\rho_i} \;=\; \frac{F_i^{\mathrm{w}}\,\omega_i^2}{(F_i^{\mathrm{w}})^2\omega_i^2 + v_i^{\mathrm{w}}}, \qquad \pi_i \;:=\; \lambda_i\rho_i \;\ge\; 0, \tag{C.10}$$

with the convention $g_i := 0$, $\pi_i := 0$ on the *inaccessible set* $\mathcal{N} := \{i : \rho_i = 0\}$ — the coordinates the teacher neither passes nor perturbs, i.e. $i > m$ for the truncation sketch, where no downstream filter has anything to act on. Write $\langle u, v\rangle_\pi := \sum_i \pi_i u_i v_i$ and $\|u\|_\pi^2 := \langle u, u\rangle_\pi$.

g is the optimal diagonal linear filter for the teacher's output: what a downstream model would apply if it could choose each coordinate freely, shrinking by the local signal-to-emitted-energy ratio. The weight π prices a coordinate's filtering error and is itself *spectrum-colored* — large exactly where the teacher emits energy.

Proposition C.4 (Wiener form of the risk)**.** *For every filter F,*

$$R(F) \;=\; \|F-g\|_\pi^2 \;+\; \mathrm{Irr}, \qquad \mathrm{Irr} \;:=\; \sum_i \lambda_i\Big(\omega_i^2 - \frac{c_i^2}{\rho_i}\Big) \;=\; \sum_{i\notin\mathcal{N}} \frac{\lambda_i\,\omega_i^2\,v_i^{\mathrm{w}}}{\rho_i} \;+\; \sum_{i\in\mathcal{N}} \lambda_i\omega_i^2, \tag{C.11}$$

an identity, at every finite d and for every teacher, with $c_i^2/\rho_i := 0$ on $\mathcal{N}$.

Proof in Section F.1: complete the square coordinatewise in $L^2(\pi)$.

Two readings are worth naming. The teacher itself is $F = \mathbf{1}$, so $\bar{\mathcal{E}}_w = \mathrm{Irr} + \|\mathbf{1}-g\|_\pi^2$, and for any downstream pipeline reaching filter F and paying toll $\mathfrak{T} = \sum_i \lambda_i v_i^{\mathrm{s}}$,

$$\bar{\mathcal{E}}_w - \bar{\mathcal{E}}_s \;=\; \|\mathbf{1}-g\|_\pi^2 \;-\; \|F-g\|_\pi^2 \;-\; \mathfrak{T}. \tag{C.12}$$

Weak-to-strong gain is approximation progress minus toll — and Irr, the optimal diagonal-filter risk, cancels. Every composed risk formula in this paper is thus a squared distance to one fixed target in one fixed weighted space, plus a residual no diagonal linear filter can remove. Three consequences organize the rest.

Definition C.5 (Reachable filters)**.** For a stage with covariance Σ and effective ridge $t \geq 0$ the reachable filter is the resolvent $\varphi_t := \big(\lambda_i/(\lambda_i+t)\big)_i$; write $\mathcal{R}_1 := \{\varphi_t : t \in [0,\infty]\}$ for the *single-step family* and

$$\mathcal{R}_k \;:=\; \Big\{ \textstyle\prod_{j=1}^k \varphi_{t_j} \;:\; t_1,\dots,t_k \in [0,\infty] \Big\} \tag{C.13}$$

for the *k-hop family*. $\mathcal{R}_1 \subset \mathcal{R}_2 \subset \cdots$, and $\varphi_0 = \mathbf{1}$, $\varphi_\infty = \mathbf{0}$.

Proposition C.6 (The three levers)**.** *Let R be as in (C.8), with g and π from Definition C.3; every $F \in \mathcal{R}_k$ has entries in $[0,1]$. Then:*

(i) (Gain is filtering.) *For every filter F, $R(F) < R(\mathbf{1})$ — the downstream model beats its teacher — if and only if $\langle \mathbf{1}-F,\ \mathbf{1}+F-2g\rangle_\pi > 0$. Two coordinatewise consequences: if $g_i \geq 1$ at every accessible i — as for a deterministic teacher channel with no fluctuation, where $v^{\mathrm{w}} = 0$ gives $g_i = 1/F_i^{\mathrm{w}} \geq 1$ — then* no $F \in [0,1]^d$*, hence no reachable F and no chain of any length, beats the teacher; if instead $\sum_i \rho_i(1-g_i) > 0$, the single-step family already beats it, $R(\varphi_t) < R(\mathbf{1})$ for all small $t > 0$.*
For a constant *Wiener filter $g \equiv \gamma$ with $\pi \neq 0$ the two cases are exhaustive and the threshold is exactly $\gamma = 1$.*

(ii) (First-order optimality.) *If F^* minimizes R over a family $\mathcal{F}$ that is differentiable in a parameter θ, at an* interior *θ, then $\langle F^*-g, \partial_\theta F^*\rangle_\pi = 0$: the residual is π-orthogonal to the tangent directions. Interiority is not decoration — at a boundary optimum only a one-sided inequality survives, and the inner product can be nonzero (Remark C.7 (a)) — but it is automatic wherever this paper invokes (ii) (Remark C.7 (b)).*

(iii) (Enlarging the family cannot hurt.) $\min_{\mathcal{R}_{k+1}} R \leq \min_{\mathcal{R}_k} R$.
A new hop enters at the boundary *$t_{k+1} = 0$, along $\partial_{t_{k+1}} F_i|_{t_{k+1}=0} = -F_i^*/\lambda_i$, so the first-order test is* one-sided*: the inequality is* strict *whenever $\langle F^*-g, F^*/\lambda\rangle_\pi > 0$ at some minimizer F^* of R over $\mathcal{R}_k$, and equality forces $\langle F^*-g, F^*/\lambda\rangle_\pi \leq 0$ at every such minimizer — an inequality, not π-orthogonality.*

Proof in Section F.1: Proposition C.4 plus Hilbert-space facts in $L^2(\pi)$.

Remark C.7 (Where the boundary bites, and where it cannot). **(a) Boundary optima are real.** Take any spectrum, a prior flat on the teacher's head ($\omega_i^2 = \omega^2$ for $i \leq m$; the isotropic prior qualifies) and a *ridge* teacher whose ridge exceeds $s_{\mathrm{w}}^2/(n_{\mathrm{w}}\omega^2)$, where $s_{\mathrm{w}}^2 = \sigma^2 + \sum_{i>m}\lambda_i\omega_i^2$ is its effective noise — label noise plus the energy the sketch discarded. Then $g_i > 1$ at every accessible i, so $\mathbf{1}$, i.e. $t^* = 0$, minimizes R over every $\mathcal{R}_k$ — a *boundary* optimum — while $\langle \mathbf{1}-g, \partial_t\varphi_t|_{t=0}\rangle_\pi = \sum_i \rho_i(g_i-1) > 0$.

So (ii) fails verbatim without interiority, and (iii) can hold with equality even though the residual is *not* π-orthogonal to the new direction.

(b) The teachers of this paper are safe. For an *MNI* teacher, $\sum_i [F_i^{\mathrm{w}}(1 - F_i^{\mathrm{w}})\omega_i^2 - v_i^{\mathrm{w}}] = -s_{\mathrm{w}}^2/\tau_{\mathrm{w}}$ exactly, so $\frac{d}{dt}R(\varphi_t)|_{t=0^+} = -2s_{\mathrm{w}}^2/\tau_{\mathrm{w}} < 0$ — this is (D.9) — and $\frac{d}{du}R(\varphi_{1/u})|_{u=0^+} = -2\sum_i \lambda_i^2 c_i < 0$ whenever the teacher passes some signal.

Neither endpoint of $[0, \infty]$ can then be optimal, so the single-step optimum is interior and (ii) applies as it is used in Section D.4. Proofs of (a) and (b) in Section F.1.

How to read the rest of the paper. Two forces act, and (C.11) separates them: a model reaches a filter $F \in \mathcal{R}_k$, is charged $\|F - g\|_\pi^2$ for missing the Wiener target, and pays a toll for the fluctuation it injects. Section D.2: under isotropy ($\Sigma = I_d$) the reachable filters degenerate to *constants*, the MNI student's being $F \equiv \alpha := n_{\mathrm{s}}/d$ for $n_{\mathrm{s}} \le d$, and *both* charges are quadratic in the budget — the filter objective is $r^2 - 2\alpha\sum_i c_i + \alpha^2 \sum_i \rho_i$, with an interior optimum of its own at $\alpha = \sum_i c_i / \sum_i \rho_i$, and the toll is exactly $\alpha(1-\alpha)\sum_i \rho_i$ — but their α^2 terms cancel identically, leaving the affine risk of Corollary D.3.

The cancellation uses nothing about the prior, so what kills weak-to-strong gain under isotropy is not that g is constant: it is that the only reachable filters are *scalars*, which can do no more than interpolate between copying the teacher and predicting zero, so that beating a *competent* teacher needs spectral anisotropy (Remark D.4). Section D.3: the student traverses the curve $t \mapsto \varphi_t$; more unlabeled data pushes $\varphi_t \to \mathbf{1}$ (perfect imitation, no filtering) while less filters harder at the price of a toll, so wherever Theorem D.6 (ii) supplies both boundary values — at finite d always, in the proportional limit under bounded condition number — any gain at all is bought at an interior budget (Corollary D.8). Section D.4 is Proposition C.6 (iii) for $\mathcal{R}_1 \subset \mathcal{R}_2$: a second hop improves the approximation whenever the residual at the single-step optimum pairs strictly negatively with the direction the hop opens — at an interior single optimum that pairing is the strictness criterion (D.10), in covariance form — and it must then beat the toll it pays.

The converse is not local: Theorem D.12 certifies an instance where the criterion is positive, so no infinitesimal second hop helps, and a distant two-hop configuration still wins; buying a global statement therefore costs the explicit conditions of Theorem D.11. Section D.5 scales the two charges against the budget. With a noiseless exact teacher ($F^{\mathrm{w}} = \mathbf{1}$ and $v^{\mathrm{w}} = 0$, so $c = \rho = \omega^2$, $g = \mathbf{1}$ and $\mathrm{Irr} = 0$) the risk is the approximation charge at the budget-forced filter $\varphi_{\tau_{\mathrm{s}}}$ plus its toll, and it decays at the rate of Theorem D.14 (i).

At $\sigma^2 > 0$ with the shape held fixed nothing decays: interpolation refuses to average label noise away, Ξ_{w} and Ξ_{s} tend to constants in $(0, 1)$, and the strictly positive floor of Theorem D.14 (ii) is the surviving noise transfer — the teacher's label noise passed through the student's filter, plus the noise the student re-emits on its own account (Remark K.13) — read at the filter the budget forces, not at $\arg\min_{\mathcal{R}_1}$.

D Detailed statements and extensions

D.1 The composition theorem

The teacher stage is itself a channel in disguise. Because Σ is diagonal and x is Gaussian, the head coordinates $x_H = (x_1, \dots, x_m)$ are independent of the tail, so for the teacher the tail signal is indistinguishable from label noise: $y = \langle w_H^*, x_H \rangle + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,\, \sigma^2 + A_m)$ independent of x_H, where

$$A_m \;=\; \sum_{i>m} \lambda_i (w_i^*)^2 \tag{D.1}$$

is the *approximation error* of the sketch — energy the teacher cannot even represent, returned to it as effective noise.

The teacher is thus the channel of Definition C.2 on $(\Sigma_H, n_{\mathrm{w}})$ with noise $s^2 = \sigma^2 + A_m$; write $(F^{\mathrm{w}}, v^{\mathrm{w}}, \tau_{\mathrm{w}}, \Xi_{\mathrm{w}})$ for its parameters and set $F_i^{\mathrm{w}} = 0$ for $i > m$. The student is the channel on (Σ, n_{s}) with target β_{w} and *zero* label noise; write $(F^{\mathrm{s}}, v^{\mathrm{s}}, \tau_{\mathrm{s}}, \Xi_{\mathrm{s}})$, its Γ_{s} evaluated at the conditional channel second-moment profile

$$\rho_i = (F_i^{\mathrm{w}})^2 (w_i^*)^2 + v_i^{\mathrm{w}}.$$

This is not the coordinatewise realized teacher energy $(\beta_{\mathrm{w}})_i^2$. Prior averaging also averages v^{w}, whose level depends on the realized signal profile.

Theorem D.1 (Exact W2S risks)**.** *Under Assumption C.1, also permitting $\sigma^2 = 0$, write $p_i := (w_i^*)^2$ for the* realized *signal profile and $\bar{\mathcal{E}}.[p]$ for the displayed limit with p_i in place of ω_i^2 throughout. The two readings are different statements.*

Quenched statement. *Conditionally on a realized signal sequence of bounded energy whose channel sums converge, the three risks converge in probability as $d \to \infty$ to $\bar{\mathcal{E}}_w[p]$, $\bar{\mathcal{E}}_s[p]$, $\bar{\mathcal{E}}_c[p]$.*

Annealed statement. *The finite-d limit formulas are affine in the realized profile, so their prior-averages are* exactly *the displayed ω_i^2-forms; hence $\mathbb{E}\,\mathcal{E}_{\mathrm{s}} \to \bar{\mathcal{E}}_s$ (and likewise for $\mathcal{E}_{\mathrm{w}}$, $\mathcal{E}_{\mathrm{c}}$) whenever the risks are uniformly integrable, and $\mathcal{E}_{\mathrm{s}} \xrightarrow{\mathbb{P}} \bar{\mathcal{E}}_s$ whenever in addition the realized channel sums concentrate — for which*

$$\sum_i \lambda_i^2 \operatorname{Var}\left((w_i^*)^2\right) \to 0$$

suffices, and which holds inside Assumption C.1 for every prior with a bounded fourth-moment ratio $\mathbb{E}(w_i^*)^4 = O(\omega_i^4)$ but not *for the whole variance-only class (C.1) (Remark G.4, where a prior inside Assumption C.1 is exhibited for which the annealed convergence in probability fails).*

The two readings evaluate one and the same functional — at ω_i^2 (annealed), at p_i (quenched), with $\bar{A}_\psi = \lim \sum_{i>m} \lambda_i \omega_i^2$ becoming $\lim A_m$; what differs is the mode of convergence, and the two clauses above each state their own. The functionals are

$$\begin{aligned} \bar{\mathcal{E}}_w &= \bar{A}_\psi + B_{\mathrm{w}} + V_{\mathrm{w}}, \qquad \bar{\mathcal{E}}_c = B_{\mathrm{c}} + V_{\mathrm{c}}, \\ \bar{\mathcal{E}}_s &= \underbrace{\sum_i \lambda_i \left(1 - F_i^{\mathrm{s}} F_i^{\mathrm{w}}\right)^2 \omega_i^2}_{\text{double-shrinkage bias (incl. tail loss)}} + \underbrace{\sum_i \lambda_i (F_i^{\mathrm{s}})^2\, v_i^{\mathrm{w}}}_{\text{teacher noise, filtered}} + \underbrace{\sum_i \lambda_i\, v_i^{\mathrm{s}}}_{\text{student self-noise}}, \end{aligned} \tag{D.2}$$

where (B, V) are the stage bias/variance of (C.7) for the teacher stage $(\Sigma_H, n_{\mathrm{w}}, \omega_H^2, \sigma^2 + \bar{A}_\psi)$ and the ceiling stage $(\Sigma, n_{\mathrm{s}}, \omega^2, \sigma^2)$; every displayed sum denotes the limit of the corresponding finite-d channel sum, which exists by Assumption C.1. The ridge variants hold with $\tau(\lambda)$ from (C.5), for the teacher and student stages and the supervised reference.

Proof idea. Condition on the teacher. A small-noise coupling extends the Gaussian surrogate theorem to noiseless targets, uniformly on each bounded target ball. Its linear and diagonal quadratic laws turn the student risk into a bounded-coefficient quadratic functional of the teacher. Applying the same laws at the teacher yields the displayed formula. Localization, prior concentration, and uniform integrability distinguish its conditional, prior-averaged, and expected-risk readings; see Section G. □

Why the three terms. The student shrinks every teacher coordinate by F_i^{s}: signal the teacher had is shrunk *twice* (term one); noise the teacher injected is shrunk *once more* (term two — the W2S gain lives here, because F^{s} suppresses exactly the spectrum-colored profile $v_i^{\mathrm{w}} \propto \lambda_i/(\lambda_i + \tau_{\mathrm{w}})^2$ the teacher emits); and the student's own interpolation injects fresh fluctuation (term three, the price of finitely many pseudo-labels). Every phenomenon below is an arithmetic consequence of this three-way competition.

Equivalently, in the Wiener coordinates of Definition C.3 — $c_i = F_i^{\mathrm{w}} \omega_i^2$, $\rho_i = (F_i^{\mathrm{w}})^2 \omega_i^2 + v_i^{\mathrm{w}}$, $g = c/\rho$, $\pi = \lambda\rho$, all built from the same signal profile as (D.2) (ω^2 in the annealed reading, p in the quenched) — Proposition C.4 at $F = F^{\mathrm{s}}$ collapses the first two terms into one squared distance:

$$\bar{\mathcal{E}}_s = \underbrace{\|F^{\mathrm{s}} - g\|_\pi^2}_{\text{missing the Wiener filter}} + \underbrace{\mathrm{Irr}}_{\text{optimal diagonal-filter risk}} + \underbrace{\mathfrak{T}}_{\text{toll for one's own fluctuation}}, \qquad \mathfrak{T} = \sum_i \lambda_i v_i^{\mathrm{s}}. \tag{D.3}$$

The first two terms are the filter objective (C.8) at the filter the budget forces on the student; the third is the toll that objective excludes by construction.

This is the form promised in Section C.2: an approximation problem in $L^2(\pi)$, a residual no diagonal linear filter can remove, and a price for finitely many pseudo-labels.

Section N reports reproducible simulations at specified spectra and budgets. These checks compare finite estimators with the deterministic formulas; they do not establish approximation outside the theorem's assumptions.

Remark D.2 (Reading PGR at the limit). Whenever the gap in the denominator is nonzero in the reading used ($\bar{\mathcal{E}}_w[p] \neq \bar{\mathcal{E}}_c[p]$ quenched, $\bar{\mathcal{E}}_w \neq \bar{\mathcal{E}}_c$ annealed), continuous mapping turns each convergence clause of Theorem D.1 into one for PGR, in the same reading and under the same hypothesis. Quenched: conditionally on w^*, on the profile-convergence event, $\mathrm{PGR} \xrightarrow{\mathbb{P}} (\bar{\mathcal{E}}_w[p] - \bar{\mathcal{E}}_s[p])/(\bar{\mathcal{E}}_w[p] - \bar{\mathcal{E}}_c[p])$. Annealed: the same ratio in the ω^2-forms is the limit in probability only when the realized channel sums concentrate — e.g. under (G.9).

It is *not* delivered by the annealed clause alone: that clause controls the *expected* risks, and PGR is a nonlinear functional of them, so nothing about $\mathbb{E}\,\mathrm{PGR}$ follows from it. In either reading the limit is an explicit function of $(\psi, \bar{\kappa_{\mathrm{w}}}, \bar{\kappa_{\mathrm{s}}}, \sigma^2, r^2)$ and the spectral profile. Isolated shapes with $\bar{\mathcal{E}}_w = \bar{\mathcal{E}}_c$ exist; there the ratio is undefined and the comparison is read from the risks directly.

D.2 Structural corollaries of the single hop

Corollary D.3 (Isotropic minimum-norm imitation)**.** *Let $\Sigma = I_d$ (any sketch) and let the student be the minimum-norm interpolator* (C.4)*. Then for every finite d and $n_{\mathrm{s}} \leq d$,*

$$\mathbb{E}\,\mathcal{E}_{\mathrm{s}} \;=\; \tfrac{n_{\mathrm{s}}}{d}\,\mathbb{E}\,\mathcal{E}_{\mathrm{w}} + \big(1 - \tfrac{n_{\mathrm{s}}}{d}\big) r^2, \tag{D.4}$$

exactly, and for every *prior* (C.1)*: only $r^2 = \mathbb{E}\|w^*\|^2$ enters, so isotropy of the prior is not needed. Hence $\mathbb{E}\mathcal{E}_{\mathrm{s}} \geq \min(\mathbb{E}\mathcal{E}_{\mathrm{w}},\, r^2)$, the expected risk is affine in n_{s}/d, and for $n_{\mathrm{s}} < d$ the student beats its teacher on average iff the teacher is worse than predicting zero (at $n_{\mathrm{s}} = d$ the two are equal, whatever the teacher).*

Remark D.4 (What isotropy permits). For $n_{\mathrm{s}} > d$ the MNI student is the exact clone (Theorem D.6); in general the factor in (D.4) is $\min(n_{\mathrm{s}}, d)/d$. So *in expectation* the MNI student is a convex combination of its teacher and the null predictor, and no interior optimum exists. *An MNI student cannot beat a competent teacher* ($\mathbb{E}\mathcal{E}_{\mathrm{w}} < r^2$) *without spectral anisotropy*; an incompetent one is beaten trivially, by partial shrinkage toward the null predictor.

Minimum norm is essential here, not cosmetic: at $\Sigma = I_d$ a *ridge* student with data to spare ($n_{\mathrm{s}} \to \infty$, so no capacity shortfall) beats every *minimum-norm* teacher with head budget $n_{\mathrm{w}} < m - 1$ — the hypothesis of Lemma H.1, and not droppable: the inverse-Wishart mean $I/(m - n_{\mathrm{w}} - 1)$ that produces the display below exists only for $n_{\mathrm{w}} < m - 1$, and at $n_{\mathrm{w}} = m - 1$ (an admissible MNI teacher, since $n_{\mathrm{w}} < m$) it diverges and the ratio is indeterminate.

For every such teacher some finite $\lambda_{\mathrm{s}} > 0$ strictly improves on $\mathbb{E}\mathcal{E}_{\mathrm{w}}$, and an explicit threshold on λ_{s} is available (Lemma H.1) — and the *infimum* over $\lambda_{\mathrm{s}} \geq 0$ of its limiting risk falls below $\mathbb{E}\mathcal{E}_{\mathrm{w}}$ by exactly $[\bar{s}_{\mathrm{w}}^2\, n_{\mathrm{w}}/(m - n_{\mathrm{w}} - 1)]^2/\mathbb{E}\|\beta_{\mathrm{w}}\|^2 > 0$, where $\bar{s}_{\mathrm{w}}^2 := \mathbb{E}[\sigma^2 + A_m] = \sigma^2 + \sum_{i>m} \lambda_i \omega_i^2$ is the *expected* effective teacher noise (Lemma H.1; the identity is annealed throughout, so the realized A_m of (D.1) may not stand in its place).

That infimum is *attained*, at a finite λ_{s}^*, exactly when the teacher's head carries signal, $\sum_{i \leq m} \omega_i^2 > 0$; when it does not, the optimum is the null predictor and is approached only as $\lambda_{\mathrm{s}} \to \infty$. Either way, explicit shrinkage needs neither anisotropy nor a capacity shortfall.

The identity (D.4) is a two-line projection argument (Section H): isotropy makes the student's projection blind to the distinction between teacher signal and teacher noise, so imitation can only interpolate between "copy the teacher" and "know nothing." All W2S action is therefore in how the spectrum *colors* the teacher's error: F^{s} passes head coordinates, where signal concentrates when $b > 0$, and blocks the tail, where the teacher's interpolation noise spreads. The next corollary makes that denoising quantitative.

Corollary D.5 (Super-ceiling: when pseudo-labels beat true labels)**.** *Fix the pipeline shape $(\psi, \bar{\kappa_{\mathrm{w}}}, \bar{\kappa_{\mathrm{s}}}, r^2)$ and the spectral profile. Then $\bar{\mathcal{E}}_s$ and $\bar{\mathcal{E}}_c$ are* exactly affine *in the label noise σ^2, with explicit slopes β_{s} (noise transferred through teacher and student filters; Section H) and $\beta_{\mathrm{c}} = \Xi_{\mathrm{s}}/(1 - \Xi_{\mathrm{s}})$*

(noise absorbed directly). Consequently the super-ceiling set $\{\sigma^2 \geq 0 : \bar{\mathcal{E}}_s < \bar{\mathcal{E}}_c\}$ is an interval determined by the signs of $\bar{\mathcal{E}}_s(0) - \bar{\mathcal{E}}_c(0)$ and $\beta_c - \beta_s$, giving exactly four regimes:

- Denoising *($\beta_s < \beta_c$ and $\bar{\mathcal{E}}_s(0) > \bar{\mathcal{E}}_c(0)$): pseudo-labels win iff $\sigma^2 > \sigma_\star^2$, an explicit threshold.*
- Dominance *($\beta_s < \beta_c$ and $\bar{\mathcal{E}}_s(0) \leq \bar{\mathcal{E}}_c(0)$): pseudo-labels win at every $\sigma^2 > 0$, and at $\sigma^2 = 0$ too when the zero-noise inequality is strict.*
- Never *($\beta_s > \beta_c$ and $\bar{\mathcal{E}}_s(0) > \bar{\mathcal{E}}_c(0)$): the student never beats the ceiling.*
- Small noise only *($\beta_s > \beta_c$ and $\bar{\mathcal{E}}_s(0) \leq \bar{\mathcal{E}}_c(0)$): pseudo-labels win iff $\sigma^2 < \sigma_\star^2$.*

On the degenerate boundary lines ($\beta_s = \beta_c$ or $\bar{\mathcal{E}}_s(0) = \bar{\mathcal{E}}_c(0)$) the comparison is decided by the other coordinate, with exact ties possible. Scope. *The intercepts $\bar{\mathcal{E}}_s(0), \bar{\mathcal{E}}_c(0)$ are values of the affine* extension *of the limiting formulas to $\sigma^2 = 0$, a point Assumption C.1 excludes; they enter only as coefficients of that extension, and every conclusion above is read at $\sigma^2 > 0$.*

All four regimes are realizable, and which one occurs is not readable off the teacher's own quality. In the *Denoising* regime, a well-trained weak teacher is a noise filter, and once true labels are noisy enough, imitating the filter beats consuming the labels — but a good ceiling does not imply it. On the two-level spectrum with 5% of coordinates at $\lambda = 10$ and 95% at $\lambda = 1$, isotropic prior, $(\psi, \bar{\kappa_w}, \bar{\kappa_s}) = (0.15, 2.5, 3)$, the zero-noise ceiling is $2.8\times$ better than the teacher ($\bar{\mathcal{E}}_c(0) = 0.51673$ against $\bar{\mathcal{E}}_w(0) = 1.44252$), and the regime is nonetheless *Never*: $\bar{\mathcal{E}}_s(0) = 1.20918$ with $\beta_s = 0.77859 > 0.58778 = \beta_c$, evaluated from the deterministic formulas.

That instance also clarifies the mechanism: *Never* is the sign $\beta_s > \beta_c$, which a small student budget produces against a perfectly well-conditioned teacher; a teacher near its interpolation threshold ($\Xi_w \to 1$, inflating $\beta_s \to \infty$) is one route to that sign, not the only one. *Dominance* is the surprise: two-level spectra with a generous teacher and a small student budget realize it, because the student fits the teacher's pre-shrunk, smaller-norm parameter with less interpolation self-noise than the ceiling pays to fit w^* itself — weak-to-strong training as pure implicit regularization, beating even noiseless supervision. *Small noise only* occurs e.g. on the two-level spectrum with 20% of coordinates at $\lambda = 100$ and 80% at $\lambda = 1$, isotropic prior, $(\psi, \bar{\kappa_w}, \bar{\kappa_s}) = (0.95, 1.5, 6)$: there $\bar{\mathcal{E}}_s(0) = 0.28232 < 0.29832 = \bar{\mathcal{E}}_c(0)$ while $\beta_s = 3.022 > 1.782 = \beta_c$, so pseudo-labels win exactly below $\sigma_\star^2 = 0.012893$. Slope formulas, the four cases, and an explicit dominance witness are in Section H.

D.3 The imitation sweet spot

Theorem D.6 (Imitation sweet spot)**.** *Consider the MNI student.*

(i) Clone plateau (finite d, exact). *If $n_s \geq d$ then $\widehat{w}_s = \beta_w$ almost surely: the student is an exact clone, $\mathcal{E}_s = \mathcal{E}_w$ identically, and no weak-to-strong gain is possible.*

(ii) Boundaries; the order of limits is part of the statement. *Read (D.2) as the channel calculus of Definition C.2 at a spectrum with $\lambda_i > 0$, with the imitation budget the only free variable and $\bar{\kappa_s} := d/n_s$.*

Then $\bar{\kappa_s} \mapsto \bar{\mathcal{E}}_s$ is continuous on $(1, \infty)$, with $\bar{\mathcal{E}}_s \to \bar{\mathcal{E}}_w$ as $\bar{\kappa_s} \downarrow 1$ (perfect imitation) and $\bar{\mathcal{E}}_s \to r^2$ as $\bar{\kappa_s} \to \infty$ (no imitation), in each of two readings: (ii-a) finite d, limit-formula reading*, in which d is held fixed and finite and n_s ranges over $(0, d)$ — valid for* every *spectrum, power laws included; and* (ii-b) proportional*, in which $\bar{\mathcal{E}}_s(\bar{\kappa_s})$ is the Theorem D.1 limit at fixed $\bar{\kappa_s}$ — valid under that theorem's regularity, ν compactly supported with $\lambda_{\min} > 0$.*

Neither scope clause is decorative: in the proportional reading at a power-law spectrum the $\bar{\kappa_s} \to \infty$ limit is not *r^2 but the noise floor of Theorem D.14 (ii), which is free of both r^2 and b (Remark I.4 carries the witness).*

(iii) Non-monotonicity, with a closed-form instance. *Consider the k-spike family with all signal on k diverging spikes, unit bulk, and a teacher head covering the spikes.*

In the domain $k < n_{\mathrm{w}} < m$, $2k < d$, and $k < n_{\mathrm{s}} < d$, the limiting weak-to-strong gain has the closed form

$$\bar{\mathcal{E}}_w - \bar{\mathcal{E}}_s \;=\; \sigma^2\,\frac{(d-n_{\mathrm{s}})\,(n_{\mathrm{w}}-k)\,(n_{\mathrm{s}}-2k)}{(d-k)\,(n_{\mathrm{s}}-k)\,(m-n_{\mathrm{w}})}, \tag{D.5}$$

It is strictly positive iff $\sigma^2 > 0$ and $n_{\mathrm{s}} > 2k$, and vanishes at both ends ($n_{\mathrm{s}} \downarrow 2k$ and $n_{\mathrm{s}} \uparrow d$); at $\sigma^2 = 0$ the gain is identically zero, so the label noise that Assumption C.1 keeps strictly positive is what makes the "iff" bite.

The maximizing budget is

$$n_{\mathrm{s}}^* \;=\; k + \sqrt{k\,(d-k)}, \tag{D.6}$$

which lies strictly inside $(2k, d)$ exactly when $d > 2k$ (Section I.4).

Proof idea. (i) is linear algebra: with $n_{\mathrm{s}} \geq d$ independent Gaussian rows, $\widetilde{X}$ has full column rank, and the interpolation constraint $\widetilde{X}w = \widetilde{X}\beta_{\mathrm{w}}$ has the unique solution β_{w}. (ii) requires care only at $\bar{\kappa}_{\mathrm{s}} \downarrow 1$, where $\tau_{\mathrm{s}} \to 0$ and $\Xi_{\mathrm{s}} \to 1$ simultaneously; the fixed point ties the two rates together so that the self-noise term converges to zero energy against every fixed spectral weight. (iii) In the spiked family every channel quantity is a root of a quadratic.

The order of limits is $d \to \infty$ first at fixed spike *fraction* $\pi = k/d$, height L and ratios (a fixed finite k would violate Assumption C.1's uniform-integrability clause), then $L \to \infty$, which collapses the formulas to (D.5) — scale-invariant (degree 0 in $(k, n_{\mathrm{w}}, n_{\mathrm{s}}, m, d)$), hence written in finite-symbol notation and read at the π-level. There teacher *bulk* overfitting passes to the student damped by $F_{\mathrm{bulk}}^{\mathrm{s}\,2} < 1$ while spike signal passes undamped — the student filters exactly what the teacher got wrong — and continuity in $1/L$ transfers the strict inequality to finite spikes (already at $L = 100$ in a numerical illustration, Remark I.2: $\bar{\mathcal{E}}_s = 0.0712 < 0.1316 = \bar{\mathcal{E}}_w$, a 46% reduction). Full proof in Section I. □

Remark D.7 (Reading the spiked-family optimizer). The condition $n_{\mathrm{s}} > 2k$ says that *any* imitation budget beyond twice the signal dimension buys a strict improvement over the teacher. In this family the gain is *strictly unimodal*: with $t := n_{\mathrm{s}} - k$ and $D := d - k$ the n_{s}-dependent factor of (D.5) is $(D-t)(t-k)/t = (D+k) - t - Dk/t$, whose derivative $Dk/t^2 - 1$ is positive below $t = \sqrt{Dk}$ and negative above it.

So n_{s}^* of (D.6) is the unique maximizer and *every* enlargement past it strictly hurts — a property of this closed form, not of the general risk curve (Corollary D.8). In the sparse-signal regime $k/d \to 0$, $n_{\mathrm{s}}^* \approx \sqrt{k\,d}$, the geometric mean of signal and ambient dimension: the optimal degree of imitation scales as $\sqrt{k/d}$, not k/d.

Corollary D.8 (Any gain forces non-monotonicity)**.** *Fix a reading in which Theorem D.6 (ii) supplies its two boundary values — (ii-a) at finite d, or (ii-b) in the proportional limit under bounded condition number. Let $\bar{\mathcal{E}}_w < r^2$ — a competent teacher, beating the null risk — and suppose some budget achieves strict weak-to-strong gain, $\bar{\mathcal{E}}_s(\bar{\kappa}_{\mathrm{s}}^{\,\circ}) < \bar{\mathcal{E}}_w$. Since $\bar{\mathcal{E}}_w$ is exactly the clone value (Theorem D.6 (i)), such a budget makes the risk curve dip strictly below* both *of its boundary values, so by Theorem D.6 (ii):*

(a) *$\bar{\mathcal{E}}_s$ is* non-monotone *in n_{s}, and its infimum over $\bar{\kappa}_{\mathrm{s}} \in (1, \infty)$ is attained — at an interior budget.*
Any gain at all is therefore bought at an interior budget.

(b) *Let n_{s}^* be the* largest *optimal budget (the minimizer set is compact and interior, so this is well defined). Then $\bar{\mathcal{E}}_s(n_{\mathrm{s}}) > \bar{\mathcal{E}}_s(n_{\mathrm{s}}^*)$ for* every *$n_{\mathrm{s}} \in (n_{\mathrm{s}}^*, d]$: no enlargement of the imitation set past the optimum ever recovers it, up to and including the exact clone at $n_{\mathrm{s}} = d$.*

What is not *claimed is unimodality — and we do not merely withhold it: high-precision evaluation of (D.2) at finite d produces spectra whose risk curve has* two *interior local minima, both strictly below the clone value, so we in fact expect unimodality to* fail *in general.*
The re-descent between the two dips is of order 10^{-4}, too fine to certify without interval arithmetic, which is why we record an expectation here and not a counterexample. The argument itself gives no uniqueness of the interior minimizer and no monotone increase at every enlargement past it, only

(b). (In the spiked family of Theorem D.6 (iii) the closed form (D.5) is strictly unimodal, so there the stronger reading holds — Remark D.7.) For an incompetent teacher, $\bar{\mathcal{E}}_w > r^2$, gain can occur monotonically, with the optimum at the no-imitation boundary; the hypothesis is not removable. Neither is the reading: the whole argument is the two boundary values of Theorem D.6 (ii), so it is available exactly where they are. At a power-law spectrum that means the finite-d reading (ii-a). In the proportional limit there the small-budget boundary value is the r^2-free noise floor rather than r^2, interiority does not follow, and it can genuinely fail: at $(a, b) = (1.05, 1)$, $\psi = 0.3$, $\bar{\kappa}_{\mathrm{w}} = 3$, $\sigma^2 = 1/4$, $r^2 = 1$ the proportional-limit curve $\bar{\mathcal{E}}_s(\bar{\kappa}_{\mathrm{s}})$ decreases strictly from 0.1731 at $\bar{\kappa}_{\mathrm{s}} = 2$ to 0.0338 at $\bar{\kappa}_{\mathrm{s}} = 10^3$, entirely below the clone value $\bar{\mathcal{E}}_w = 0.2087$ — a strict gain with no interior optimum in sight.
The same instance read at finite $d = 2 \cdot 10^5$ behaves as the corollary says: $\bar{\mathcal{E}}_s$ falls to 0.0416 at $n_{\mathrm{s}} \approx 730$ and climbs back to 0.8808 at $n_{\mathrm{s}} = 1$, on its way to $r^2 = 1$.

Takeaway. For the minimum-norm student the entire W2S effect is a failure of imitation, optimally dosed: it can surpass its teacher because $n_{\mathrm{s}} < d$ prevents the interpolation constraints from uniquely identifying the teacher's function, and the projection that loses part of the teacher *is* the regularizer that filters the teacher's noise. A ridge student buys the same regularization explicitly and needs no capacity shortfall (Remark D.4); what the two share is that *some* filter must sit between teacher and student.

Contrast the known non-monotonicities, which live on an iterative learner *stopping-time* axis (Medvedev et al., 2025) — an axis that in single-pass SGD is itself the pseudo-label count (Wu et al., 2026c); here the student is a fixed interpolation rule with no stopping knob at all, and the resource that must not be too large is the *unlabeled sample size*. Pseudo-label imitation has an optimal data budget; "more unlabeled data" is not a free action.

D.4 Distillation chains: when ladders beat leaps

A k-hop chain trains model $t{+}1$ by (ridge-regularized) interpolation of model t's pseudo-labels on a fresh batch of size n_{t+1} (matching the indexing of Section C), $\sum_{t \le k} n_t = N$. Chains compose channels:

Theorem D.9 (Chain recursion)**.** *Under Assumption C.1, also permitting $\sigma^2 = 0$, with per-hop ratios fixed ($d/n_t \to \bar{\kappa}_t > 1$, ridges $\lambda_t \ge 0$ fixed, k fixed), define the chain state $(\alpha^{(t)}, \nu^{(t)})$ by the teacher initialization $\alpha_i^{(0)} = F_i^{\mathrm{w}} \mathbf{1}\{i \le m\}$, $\nu_i^{(0)} = v_i^{\mathrm{w}} \mathbf{1}\{i \le m\}$ and the recursion*

$$\alpha_i \;\leftarrow\; F_i^{(t)} \alpha_i, \qquad \nu_i \;\leftarrow\; \big(F_i^{(t)}\big)^2 \nu_i + v_i^{(t)}, \tag{D.7}$$

where hop t's channel $(\tau_t, \Xi_t, F^{(t)})$ is that of (Σ, n_t, λ_t) and $v^{(t)}$ is computed at the realized *profile $\rho_i^{(t-1)} = (\alpha_i^{(t-1)})^2 p_i + \nu_i^{(t-1)}$, $p_i := (w_i^*)^2$, with zero label noise (the quenched convention of Theorem D.1; replacing p_i by ω_i^2 throughout gives the annealed recursion, and the two agree under the moment condition of Remark G.4).*

Then the recursion computes the limits of all risk-relevant aggregate functionals *of the chain. For a deterministic signal sequence, assume bounded Euclidean norms. For random signals, apply the conditional bounds on signal- and teacher-norm balls, then average and remove localization as in (G.7). Accordingly "deterministic" below permits w^*-*measurable *vectors. The excess risk follows by taking $(q, u) = (\lambda, \lambda \odot w^*)$ (Section J.1). For every $D, U > 0$ and* each *such pair with $\|q\|_\infty \le D$, $\|u\| \le U$, the failure probabilities are uniform over these two bounded classes after norm localization. There is no assertion of a single event serving all pairs simultaneously:*

$$\sum_i q_i \big(\beta_i^{(t)}\big)^2 \xrightarrow{\mathbb{P}} \sum_i q_i \, \rho_i^{(t)}, \qquad u^\top \beta^{(t)} \xrightarrow{\mathbb{P}} \sum_i u_i \, \alpha_i^{(t)} w_i^*, \qquad t = 0, 1, \ldots, k \tag{D.8}$$

In particular the hop-t excess risk converges,

$$\mathcal{E}^{(t)} \xrightarrow{\mathbb{P}} \bar{\mathcal{E}}^{(t)} = \sum_i \lambda_i \big[(1 - \alpha_i^{(t)})^2 p_i + \nu_i^{(t)}\big],$$

and hop $t{+}1$'s realized channel inputs (the Γ-quadratic) converge to the recursion's $(\rho^{(t)}, \Gamma^{(t+1)})$.

Thus a chain applies the *product filter* $\prod_t \lambda_i/(\lambda_i+\tau_t)$ to the teacher — composed Tikhonov shrinkages — but, in contrast to the design-reuse self-distillation settings in which every stage is refit at the same n design points (Lecoiu et al., 2026; Garg et al., 2026), each hop must be paid for with a slice of the data budget, and the fluctuation $v^{(t)}$ it *injects* is derived from that slice: budget-coupled, and spectrum-colored so that the next hop can filter it. A phase transition resolves the competition between richer filters and thinner slices.

The population comparison below assumes that the single-step optimum is attained in the interior. That is a statement about the *global* minimum of R_1 over $[0,\infty]$, not about a stationary point, and two conditions supply it. Differentiate the single-ridge population risk R_1 at $t=0$: for a *ridgeless* teacher the bias parts cancel by (G.1), which gives $\Gamma_{\mathrm{w}} S_2 - \tau_{\mathrm{w}} S_1 = s_{\mathrm{w}}^2/\tau_{\mathrm{w}}$ through $n_{\mathrm{w}}(1-\Xi_{\mathrm{w}}) = \tau_{\mathrm{w}} S_2$, with $S_1 := \sum_{i\le m} \lambda_i \omega_i^2/(\lambda_i+\tau_{\mathrm{w}})^2$ and $S_2 := \sum_{i\le m} \lambda_i/(\lambda_i+\tau_{\mathrm{w}})^2$, leaving

$$R_1'(0) \;=\; -\,2\,\frac{s_{\mathrm{w}}^2}{\tau_{\mathrm{w}}}, \qquad s_{\mathrm{w}}^2 = \sigma^2 + A_m \qquad \text{(MNI teacher)}, \tag{D.9}$$

so exogenous teacher noise, $s_{\mathrm{w}}^2 > 0$ — guaranteed by Assumption C.1 — rules out $t^* = 0$.

Ruling out $t^* = \infty$ is a second and independent condition: $R_1(t) - R_1(\infty) = -2t^{-1}\sum_i \lambda_i^2 F_i^{\mathrm{w}} \omega_i^2 + O(t^{-2})$, so the teacher must *pass* some signal, $\sum_i \lambda_i^2 F_i^{\mathrm{w}} \omega_i^2 > 0$. Under both, the minimum of R_1 over the compactified $[0,\infty]$ is attained in the interior; both hold throughout the power-law family below, where $\omega_i^2 > 0$ on the head.

The implication runs only this way. (D.9) is genuinely ridgeless: with teacher ridge $\lambda_{\mathrm{w}} > 0$ its left side is strictly the larger, by $2\lambda_{\mathrm{w}} n_{\mathrm{w}}(\tau_{\mathrm{w}}^2 S_1 + s_{\mathrm{w}}^2)/[\tau_{\mathrm{w}}(\tau_{\mathrm{w}}^2 S_2 + \lambda_{\mathrm{w}} n_{\mathrm{w}})] > 0$. A teacher whose head carries no signal has $R_1' < 0$ on all of $[0,\infty)$ — an infimum only at $t=\infty$ — however large s_{w}^2 may be.

And $s_{\mathrm{w}}^2 > 0$ is not necessary: the channel level Γ_{w} of (C.6) is fed by the teacher's own *bias* as well as by its label noise, and bias-only fluctuation already produces a strict interior optimum. What a chain filters is the teacher's *fluctuation*, whichever of the two sources emitted it. One further distinction: this optimum is not the sweet spot of Theorem D.6.

There the variable is the budget n_{s} and the objective carries the student's own toll; here it is the ridge t at a fixed budget, with no toll. The first does not follow from the second — Remark I.3 exhibits $s_{\mathrm{w}}^2 > 0$ and an interior optimum of R_1 alongside no sweet spot at all.

Theorem D.10 (Chains vs. the best single step)**.** *Fix a teacher and a total unlabeled budget N; compare the best single ridge student ($R_1^* = \min_\lambda \bar{\mathcal{E}}_s$) with the best two-hop chain (R_2^*, optimized over the split and both ridges); write $R_1^{\mathrm{pop},*}, R_2^{\mathrm{pop},*}$ for their population ($N\to\infty$) counterparts. Assume the single-step population risk R_1 attains its infimum over $[0,\infty]$ at an interior t^*, so that $R_1(t^*) = R_1^{\mathrm{pop},*}$; put $F_i^* = \lambda_i/(\lambda_i+t^*)$, and define the* filter residuals

$$\Delta_i := F_i^{\mathrm{w}}(1 - F_i^* F_i^{\mathrm{w}})\,\omega_i^2 - F_i^* v_i^{\mathrm{w}}.$$

(i) Population level. $R_2^{\mathrm{pop},*} \le R_1^{\mathrm{pop},*}$ *always, and* $R_k^*(N) \ge R_k^{\mathrm{pop},*}$ *at every budget; the excess over the population value is at most an explicit* splitting cost, *which vanishes as N grows whenever the per-hop effective dimension* $\mathrm{ed}(\tau) := \sum_j \lambda_j^2/(\lambda_j+\tau)^2$ *is $o(N)$ at the relevant ridges — true at fixed d, and along the power law (C.2) at fixed $d/N > 1$ exactly when those ridges stand far above the interpolation scale, $\tau N^a \to \infty$;* false *for a spectrum bounded above and below in the proportional regime, where a noiseless minimum-norm hop at $\Sigma = I$, $N/d \to \alpha$ pays the N-independent toll $\alpha(1-\alpha)\|u\|^2$ (Proposition J.20, Remark J.22).*

Vanishing is not the same as negligible: along the power law at the tuned *ridges the splitting cost is $\Theta(N^{-1+1/(a+b)})$ for $b < a+1$, $\Theta(N^{-1+1/(2a+1)})$ for $b > a+1$ (past the Tikhonov saturation point), and $\Theta\big(N^{-1+1/(2a+1)}(\log N)^{1/(2a+1)}\big)$ at the saturation boundary $b = a+1$ — in no case $O(1/N)$ — and it stays a fixed fraction of $R_1^{\mathrm{pop},*}$ at every d, so the population gap is approached in absolute but not in relative terms (Proposition J.25, Remark J.21).*

The boundary factor is literally the same $(\log N)^{1/(2a+1)}$ that Theorem D.14 (iii) carries, and for the same reason: $b = a+1$ makes $a+b = 2a+1$, the logarithmic case of the

underlying power-law sums. If the strictness criterion

$$\sum_i F_i^*(1-F_i^*)\,\Delta_i \;<\; 0 \qquad\Longleftrightarrow\qquad \mathrm{Cov}_\pi\big(\Delta,\; 1/\lambda\big) \;<\; 0, \quad \pi_i \propto (F_i^*)^2, \tag{D.10}$$

holds at that t^*, then the two-hop family wins strictly, $R_2^{\mathrm{pop},*} < R_1^{\mathrm{pop},*}$.

Globality of t^* is not removable: an instance meeting every standing hypothesis has a strictly negative criterion at a local *interior optimum sitting 42% above $R_1^{\mathrm{pop},*}$, so the argument controls the wrong quantity (Remark J.7).*

(ii) Finite budgets: a chain-wins certificate. *At the fully specified instance $d = 4000$, $m = 1200$, $n_{\mathrm{w}} = 400$, $N = 2000$, $\sigma^2 = 0.25$, $r^2 = 1$, $(a, b) = (1.2, 2.0)$ — i.e. $(\psi, \bar{\kappa_{\mathrm{w}}}, N/d) = (0.3, 3, 0.5)$ — the* annealed *recursion values satisfy $R_2^* \le 0.016302 < 0.017217 \le R_1^*$, a rigorous margin. (This is a statement about the deterministic recursion of Theorem D.9 at that instance — in its annealed reading, ω_i^2 in place of the realized p_i, which is the reading the certificate computes and the one the population analysis uses; the two readings are distinct along the power law, Remark G.4 (b) and Remark J.19 — certified by interval arithmetic, not about a finite-sample random risk.)*

Here Δ_i is positive where the single filter under-fits the teacher's signal and negative where it under-kills the teacher's noise, so criterion (D.10) says the residuals at the single-ridge optimum correlate negatively with $1/\lambda$: the tail is under-filtered because one knob must serve head and tail at once.

Proof idea. (i) The population objective is a weighted least-squares fit of the Wiener profile by the filter family; the single-ridge family $\{\lambda/(\lambda+t)\}$ is the $t_2 = 0$ slice of the two-hop family $\{\lambda^2/((\lambda+t_1)(\lambda+t_2))\}$, so first-order perturbation in t_2 at the single optimum decides strictness; the optimality identity $\sum_i (F_i^*)^2\Delta_i = 0$ turns the derivative into the covariance form (D.10). (ii) The mechanism is a filter-order jump: a single ridge rolls off as λ^1 in the tail, a composed filter as λ^2, so on aligned signals the single knob must over-shrink the head to control the tail — reminiscent of escaping a one-parameter filter family's rigidity in classical regularization (Engl et al., 1996, Chs. 4–5), though Remark J.28 delimits how this does and does *not* correspond to classical qualification. The certificate combines interval-enclosed fixed points with the population lower bound on every single-ridge student. Full proof in Section J. □

Sufficiency in Theorem D.10 (i) asks nothing beyond the interior global single-step optimum: no transversality, no alignment restriction. Its converse — does criterion ≥ 0 imply that *no* two-hop configuration beats the best single ridge? — splits the alignment axis into three, and fails in general.

Theorem D.11 (Global necessity: a trichotomy)**.** *Population level, power-law family; for (ii)–(iv) assume the single-ridge risk attains its infimum at an interior t^* — automatic for the MNI teacher (C.3) here, by the two conditions stated with (D.9), and not automatic for a ridge teacher.*

(i) $b = 0$: unconditional tie. $\min_Q \psi = \inf_{e_2=0} \psi$*: no two-hop chain, physical or relaxed, ever strictly beats the best single ridge.*

Moreover at every interior single optimum the pattern is $(-\,\mathrm{head},\, +\,\mathrm{tail})$ *and the criterion is* strictly positive *— the $b = 0$ failure of (D.10) is a theorem, for every overparametrized teacher ($\tau_{\mathrm{w}} > 0$; Assumption C.1).*

(ii) $b \ge a$: chains always win. *At every interior single optimum the pattern is exactly* $(+\,\mathrm{head},\, -\,\mathrm{tail})$ *and the criterion is strictly negative: the two-hop chain strictly wins for* all $b \ge a$.

The criterion-failing region is therefore contained in $\{b < a\}$, and necessity is vacuous on $b \ge a$.

(iii) $b \ge 2a$: equal-ridge rigidity. *ψ has no critical point in the open quadrant; consequently every minimizer of ψ over the physical region D lies on the parabola $t_1 = t_2$: the optimal two-hop chain has exactly equal ridges $t_1^* = t_2^* > 0$.*

(iv) $0 < b < a$: conditional necessity. *Under the transversality/no-saddle conditions* (SC1)–(SC3) *of Definition J.12 — per-instance semialgebraic conditions, sampled at scale with*

zero violations on the criterion-failing region, not exhaustively verified — criterion ≥ 0 at t^* implies the global tie $\min_Q \psi = R_1^{\mathrm{pop},*}$. Combined with sufficiency (Theorem D.10 (i)): under (SC)*, strict global win $\iff$ criterion < 0.*

Theorem D.12 (No unconditional converse)**.** *At the power-law instance $d = 4000$, $m = 1200$, $n_{\mathrm{w}} = 400$, $\sigma^2 = 0.25$, $r^2 = 1$, $(a, b) = (2.0, 0.0675)$ — all values below referring to the* centered *population objective ψ of Section J.4, which drops a filter-independent constant, so the margin is centering-invariant — the transversal derivative at the single optimum $t^* = 2.9057934773 \times 10^{-3}$ is $D_0 := \partial_{e_2}\psi|_{(t^*,0)} = +0.3667620 > 0$, so criterion (D.10) is positive.*

Nevertheless, the interior point $(e_1, e_2) = (2.9043477340 \times 10^{-3},\ 1.8160576582 \times 10^{-9})$ — the physical two-hop chain $(t_1, t_2) = (2.90372 \times 10^{-3},\ 6.25424 \times 10^{-7})$, both hops strictly positive — satisfies

$$\psi(e_1, e_2) - R_1^{\mathrm{pop},*} \;=\; -2.20937164288882 \times 10^{-10} \;<\; 0, \tag{D.11}$$

Thus criterion ≥ 0 does not *imply global optimality of the single ridge, and Theorem D.10 (i) cannot be upgraded to an unconditional iff: any unconditional necessity theorem on $0 < b < a$ must carry an (SC3)-type hypothesis.*

Every quantity above is enclosed by interval arithmetic with directed rounding, so the sign of (D.11) is a rigorous computer-assisted verification, not a floating-point evaluation.

The trichotomy's two unconditional ends and the certified failure of the naive converse are proved in Section J.4, which also states *(SC1)–(SC3)* (Definition J.12) and their status (Remark J.13: semialgebraic per instance, sampled at scale, not proved as general inequalities), and locates the sliver in which the converse fails (Remark J.18).

Remark D.13 (The phase boundary $b^*(a; d)$, and what fuels a chain). In the power-law family, criterion (D.10) provably *holds* for every $b \geq a$ and provably *fails* at $b = 0$ (Section J.4). Between the two we *select* a boundary rather than assume one. The criterion is a function of the whole parameter list $(a, b, \psi, \bar{\kappa_{\mathrm{w}}}, \sigma^2, r^2, d)$ — the ambient dimension included, since both t^* and the teacher channel move with d — so at each fixed $(\psi, \bar{\kappa_{\mathrm{w}}}, \sigma^2, r^2, d)$ put

$$b^*(a; d) \;:=\; \inf\big\{\, b \geq 0 \;:\; \text{criterion} < 0 \text{ at every } b' \in [b, a] \,\big\} \;\in\; [0, a],$$

the smallest source exponent past which the criterion never again fails below $b = a$.

This is well defined and $\leq a$ at every d by the two provable ends; that it is the *unique* sign change — that the criterion is negative on all of $(b^*(a; d), \infty)$ and positive on $[0, b^*(a; d))$ — is observed numerically and not proved, so every statement below about b^* is at numerical grade. At $d = 4000$, with $\psi = 0.3$, $\bar{\kappa_{\mathrm{w}}} = 3$, $\sigma^2 = 1/4$, $r^2 = 1$, one finds $b^*(a; 4000) \in (0, \frac{1}{2})$ throughout — peaking at ≈ 0.30 near $a \approx 1.1$ and decaying to 0.002 by $a = 3$ (Section J.5).

The d-dependence is real, and it is a decay. At the same shape parameters:

$b^*(a; d)$	d=1000	2000	4000	8000	16000	32000
$a = 1.2$	0.32276	0.30471	0.28567	0.26684	0.24885	0.23203
$a = 1.5$	0.27169	0.23423	0.20159	0.17350	0.14943	0.12883
$a = 2.0$	0.13222	0.09481	0.06761	0.04800	0.03397	0.02398
$a = 3.0$	0.01042	0.00435	0.00178	0.00072	0.00029	0.00012

Over these five doublings the decay is a clean power of d, with fitted exponents -0.095, -0.215, -0.493, -1.288 at $a = 1.2, 1.5, 2.0, 3.0$.

If it persists, $b^*(a; d) \to 0$ and the "chains lose at small b" phase *closes*: at any fixed $b > 0$ the criterion is eventually negative and Theorem D.10 (i) makes the two-hop chain a strict population winner. That would be a stronger and cleaner statement than the one the table displays, and we do not have it: the two provable ends are d-free and pin b^* only to $[0, a]$, and we have no asymptotic analysis of the criterion at fixed small $b > 0$. We record the decay as an observation and the limit as an open problem.

The same d-dependence places Theorem D.12 exactly. That theorem is rigorous at its own fully specified instance, but the instance ($d = 4000$, $(a, b) = (2, 0.0675)$) sits a distance 1.1×10^{-4} in b

below $b^*(2; 4000) = 0.06761$, and at the same (a, b) the transversal derivative D_0 it certifies moves from $+0.3667$ at $d = 4000$ to -1.95×10^2 at $d = 8000$ and -3.38×10^4 at $d = 128000$. Nothing about the counterexample is claimed uniformly in d, and its statement fixes every parameter for precisely this reason.

The crossing marks the *local*-criterion sign change: it is the exact global phase boundary in the second-order (supercritical) regime, numerically $a \lesssim 1.54$, and over-estimates the true boundary by a sliver of width $\approx 0.22\%$ in b (at $a = 2$, $d = 4000$) where the transition is first-order — a bisection estimate, not a bound (Remark J.18). Deep in the aligned regime the optimum is rigid: for $b \geq 2a$ the optimal two-hop chain provably has equal ridges $t_1^* = t_2^*$. For a teacher that emits no fluctuation at all ($v^{\mathrm{w}} \equiv 0$) the criterion degenerates to zero (Remark J.9): a chain wins by filtering what the teacher emitted, and there is then nothing to filter.

A chain enlarges the population filter family and adds a finite-data cost. The local criterion gives a sufficient population improvement; it does not alone determine whether splitting a fixed budget helps. The certificate supplies a separate positive instance. Negative values from numerical optimization do not certify global failure of every chain.

D.5 Weak-to-strong scaling laws

Theorem D.14 (Three scaling regimes)**.** *Power-law family (C.2), $a > 1$; the statements are exact asymptotics of the limiting risk formulas (cf. the division of labor after Assumption C.1).*

(i) Imitation rate. *With a noiseless exact teacher, and with the student's overparametrization ratio bounded away from the interpolation boundary — $n_{\mathrm{s}} < d$ and $\liminf_d d/n_{\mathrm{s}} > 1$, the clause Assumption C.1 already imposes, here also allowing $d/n_{\mathrm{s}} \to \infty$ (sparse data) — the MNI student obeys*

$$\bar{\mathcal{E}}_s(n_{\mathrm{s}}) \asymp n_{\mathrm{s}}^{-\min(a+b-1,\, 2a)}$$

(times $\log n_{\mathrm{s}}$ at the boundary $b = a+1$), with constants depending only on (a, b, r^2) and that bound: the source rate up to the interpolation ceiling $2a$.

The hypothesis is not technical — at $d = n_{\mathrm{s}}+1$ the exponent is strictly larger, $\bar{\mathcal{E}}_s \asymp n_{\mathrm{s}}^{-(a+b)}$ for $b < a+1$ (Remark K.12).

(ii) Noise plateau. *With $\sigma^2 > 0$ and fixed ratios $(\psi, \bar{\kappa_{\mathrm{w}}}, \bar{\kappa_{\mathrm{s}}})$, scaling all of $(m, n_{\mathrm{w}}, n_{\mathrm{s}}, d)$ proportionally drives $\bar{\mathcal{E}}_s$ to an explicit strictly positive limit*

$$\sigma^2\, C_{\mathrm{floor}}(a, \psi, \bar{\kappa_{\mathrm{w}}}, \bar{\kappa_{\mathrm{s}}}) :$$

interpolating pipelines do not obey a decaying scaling law at fixed shape.

This is a limit, not a finite-d lower bound, and the approach is not one-sided: in the power-law families we evaluate $\bar{\mathcal{E}}_s$ falls to $\sigma^2 C_{\mathrm{floor}}$ from above *at small source exponents and rises to it from below at large ones (Remark K.13).*

(iii) No scaling penalty for weak supervision. *With $\sigma^2 > 0$, optimally tuned ridge at both stages, generous capacity $m = \Theta(n_{\mathrm{w}})$ and an imitation budget that dominates the label budget, $d/n_{\mathrm{s}} \to \bar{\kappa_{\mathrm{s}}} \in (1, \infty)$ with $n_{\mathrm{s}}/n_{\mathrm{w}} \to \infty$ (the regime in which the student is a near-clone and its self-noise is lower order; the proof uses the bound away from 1, not merely $n_{\mathrm{s}} = \Theta(d)$, and Remark K.15 records where),*

$$\bar{\mathcal{E}}_s^*(n_{\mathrm{w}}) \asymp n_{\mathrm{w}}^{-(a+b-1)/(a+b)}$$

for $b < a+1$ (saturating at $n_{\mathrm{w}}^{-2a/(2a+1)}$ for $b > a+1$, with an extra $(\log n_{\mathrm{w}})^{1/(2a+1)}$ factor at the boundary $b = a+1$) — the same exponent as optimally-ridged supervised regression on the n_{w} true labels.

For the next corollary, take the Gaussian prior $w^* \sim \mathcal{N}(0, \mathrm{diag}(\omega_i^2))$, the prior class (C.1) with Gaussian marginals. The Bayes rate is prior-specific: discrete priors with the same variances can be learned faster.

Corollary D.15 (The exponent is information-theoretically optimal)**.** *Conditional on the teacher's labeled sample, the student's unlabeled data is independent of w^*; hence $w^* \to (X, y) \to \widehat{w}_{\mathrm{s}}$ is a Markov chain, and* no *weak-to-strong pipeline — whatever the student, its budget, or the number of hops — can beat the Bayes risk of n_{w} labeled examples, which is $\Omega\big(n_{\mathrm{w}}^{-(a+b-1)/(a+b)}\big)$ (the matching upper bound is the supervised tuned ridge; the same exponent is attained, with matching upper and lower bounds for one-pass sketched SGD in the regime $b < 1$, by Lin et al., 2024). The rate in Theorem D.14 (iii) is therefore optimal for $b < a+1$.*

This lower bound is *unconditional* in the convergence sense: it needs no asymptotic theory, only the Markov structure and the Gaussian prior.

Remark D.16 (Two structural facts behind Theorem D.14). (a) For a noiseless teacher the student's risk collapses to $B_0/(1-\Xi_{\mathrm{s}})$ — pure bias inflated by a single scalar — and in the sparse-data regime of (i) ($n_{\mathrm{s}}/d \to 0$), $\Xi_{\mathrm{s}} \to 1 - 1/a$ exactly, so the inflation factor tends to a. (b) The plateau level in (ii) is *independent of the signal profile b*: a property of the covariance decay and the sampling ratios alone, while b controls only the (vanishing) imitation part.

Numerical evaluations of the deterministic formulas give fitted slopes: fitted slopes $-1.057/-2.019/-4.019$ against predictions $-1/-2/-4$ across $(a, b) \in \{(1.5, 0.5), (2, 1), (2, 3.5)\}$ for (i) — the last in the $2a$-saturation regime — risk flat to 1.4% across an $8\times$ proportional scale-up for (ii), and fitted exponent -0.667 against $-(a+b-1)/(a+b) = -2/3$ for both the W2S pipeline and its supervised reference in (iii). Proofs in Section K.

Reading (ii) and (iii) together: *under the displayed resource regime, the regularized formulas preserve the supervised exponent, whereas fixed-shape interpolation retains a positive noise limit.* One further consequence of (D.2): as a function of teacher capacity m at fixed n_{w}, the student inherits its teacher's model-size double descent through the filtered-noise term — the formulas can transfer large teacher interpolation variance to the student; the effect depends on the downstream regularization. The capacity formulas and the double-descent picture follow from (D.2) by the same three-term competition.

E Notation and proof routes

E.1 Notation

The notation of the main text is collected in Section A.2; this subsection lists what the detailed derivations add. The main text uses $\mathcal{R}_s^{(d)}$ for a finite-dimensional deterministic formula. The detailed statements use $\bar{\mathcal{E}}_s$ for the same formula or its specified limit; each theorem states which limit is taken. Conditional formulas use the realized profile $p_i = (w_i^*)^2$ and may be random with respect to the signal prior. Prior-averaged formulas use $\omega_i^2 = \mathbb{E}p_i$. The teacher parameter β_{w}, input designs, and risks are random. A stage's τ, Ξ, F depend only on covariance, sample size and ridge; its Γ, v also depend on the target profile. They are not samplewise coordinates of the fitted estimator.

Symbols	Meaning in the detailed derivations
$\bar{\mathcal{E}}_w, \bar{\mathcal{E}}_s, \bar{\mathcal{E}}_c$	Deterministic risk formulas of teacher, student and supervised reference, or their specified limits.
p_i	Realized signal profile $(w_i^*)^2$, with $\omega_i^2 = \mathbb{E}p_i$.
Irr	Linear ignorance of the teacher in the calculus (Proposition C.4); includes head uncertainty and tail signal.
δ_0, R	Distance from interpolation and a fixed localization radius.

The detailed derivations retain several local conventions: c_i denotes a stage eigenvalue in Definition C.2 and a signal cross-moment in Definition C.3; ψ denotes the dimension ratio or the centered chain objective; π_i denotes the Wiener weight or the explicitly renormalized criterion weight. The spike count and the chain length are both denoted by k in their respective sections, and in Section G a is an added noise variance, not a power-law exponent. These conventions never identify the objects.

E.2 Proof routes

Ledger. Section B proves the knowledge–calibration decomposition and its monotonicity by conditioning, the exact second-order calculus by sign symmetry, and the consequences in the calculus by direct computation. Its second proof of the chain approximation, in expectation, uses the exact calculus, Lemma G.3, the teacher's law and uniform integrability.

Composition. Lemma F.1 states the imported positive-noise surrogate. Lemma G.2 removes its noise floor by coupling and ordered limits, and Lemma G.3 extracts linear and quadratic functionals. Section G conditions on the teacher, evaluates those aggregates, and proves the prior-averaging and integrability steps.

Single-stage consequences. Section H proves the isotropic identity and ridge comparison. Section I establishes cloning, boundary values, and the spiked gain and optimizer. The information bounds have short proofs in the main text and further access-model details in Section M.

Chains and scaling. Section J.1 iterates the aggregate one-stage estimates. Sections J.3, J.4 and J.6 separate the population perturbation argument, global sign analysis, and finite-budget comparisons. The two computer-assisted inequalities use the accompanying interval certificates. Section K supplies the power-law sum estimates and their rate consequences.

E.3 Scope of the supporting results

Probabilistic composition is proved for a bounded spectral window. The chain approximation is uniform over all depths when every stage stays a fixed distance from the interpolation threshold (Theorem J.2). The constants depend on that window and cannot be carried to power laws by substitution. Power-law rates concern the deterministic formulas. A numerical optimizer supplies a candidate, whereas interval certificates establish specified global comparisons. Replication transfers a finite-template gap to actual estimators in Theorem J.30. The Gaussian Bayes bound applies to actual estimators independently of the deterministic-equivalent theorem; matching a formula exponent to its lower bound does not by itself prove estimator-level rate attainment.

F The interpolation channel: statement and provenance

Lemma F.1 (Channel surrogate; restatement of Han & Xu (2023))**.** *Consider a stage as in Definition C.2 with Gaussian design, covariance* $C = \mathrm{diag}(c_i)$, $c_i \in [1/M, M]$, $p/n \to \kappa > 1$ *bounded away from one, ridge* $\lambda \in [0, N]$, *noise level* $s^2 \in [1/K_s, K_s]$ *for a constant* $K_s \geq 1$ *(the noise floor: the* noiseless *stage* $s^2 = 0$ *is outside this lemma's scope — its asymptotic extension is Lemma G.2), and target* u *with* $\|u\| \leq R$.
Let $\widehat{w}(\lambda)$ *be the ridge estimator (*$\lambda = 0$*: MNI). Then for every globally L-Lipschitz function* $f : \mathbb{R}^p \to \mathbb{R}$, *and every* $\epsilon \in (0, 1/2]$,

$$\sup_{\|u\| \leq R} \Pr\Big(\sup_{\lambda \in [0,N]} \big| f(\widehat{w}(\lambda)) - \mathbb{E}_g f\big(W(\lambda, g)\big) \big| \geq \epsilon \Big) \;\leq\; C\, p\, e^{-p\,\epsilon^4/C},$$

$$W(\lambda, g)_i := F_i u_i + \sqrt{v_i}\, g_i,$$

with (F, v) *the channel parameters of Definition C.2 at* $\tau(\lambda)$.

Provenance and the Γ–γ^2 dictionary. We use version 2 of arXiv:2307.02044. Han & Xu (2023, Thm. 2.2 and eq. (2.2)) state the surrogate as

$$\widehat{w} \approx (C + \tau I)^{-1} \big(Cu + p^{-1/2} \gamma\, C^{1/2} g \big)$$

with

$$\gamma^2 = \kappa \big[s^2 + \tau^2 \langle u, C(C + \tau I)^{-2} u \rangle \big] / (1 - \kappa\chi) \quad \text{and} \quad \chi = p^{-1} \operatorname{tr}[C^2 (C + \tau I)^{-2}].$$

Since $\kappa\chi = \Xi$ and $\kappa/p = 1/n$, the noise coordinate variance is

$$c_i \gamma^2 / (p(c_i + \tau)^2) = c_i \Gamma / (c_i + \tau)^2 = v_i$$

with Γ as in (C.6): Definition C.2 gives their effective parameters in channel coordinates. Their objective normalizes the squared loss by the feature dimension; ours normalizes it by the sample size. Thus their ridge parameter is $(n/p)\lambda$, while their $\gamma^2 = p\Gamma$.

Their Thm. 2.2 (Gaussian designs) carries the exponential tail above and the sup over targets in a ball; their Thm. 2.3 extends to sub-Gaussian designs at polynomial rates for typical targets — we consume only the Gaussian form.

The aggregate identity (C.7) follows from the linear and quadratic functionals in Lemma G.3; consistency of the per-coordinate profile $v_i \propto c_i/(c_i+\tau)^2$ with the aggregate variance $V = s^2\Xi/(1-\Xi)$ is the computation

$$\sum_i c_i v_i = \Gamma \sum_i c_i^2/(c_i+\tau)^2 = \Gamma n \Xi.$$

Self-consistency check (used repeatedly). For any target profile,

$$\begin{aligned}\sum_i c_i\big[(1-F_i)^2u_i^2 + v_i\big] &= \tau^2 \sum_i \tfrac{c_i u_i^2}{(c_i+\tau)^2} + \big[\tau^2 \sum_i \tfrac{c_i u_i^2}{(c_i+\tau)^2} + s^2\big]\tfrac{\Xi}{1-\Xi} \\ &= B + V :\end{aligned}$$

the channel reproduces its own aggregate.

F.1 The Wiener form of the risk: proofs of Propositions C.4 and C.6

Proof of Proposition C.4. On $\mathcal{N}$ the summand of (C.8) is $\lambda_i\omega_i^2$ whatever F_i is ($F_i^{\mathrm{w}} = v_i^{\mathrm{w}} = 0$ there), matching both sides. Off $\mathcal{N}$, expand (C.8) in F_i and use c_i, ρ_i of Definition C.3:

$$\begin{aligned}R(F) &= \sum_i \lambda_i[\omega_i^2 - 2F_i c_i + F_i^2\rho_i] \\ &= \sum_i \lambda_i\rho_i\big(F_i - \tfrac{c_i}{\rho_i}\big)^2 + \sum_i \lambda_i\big(\omega_i^2 - \tfrac{c_i^2}{\rho_i}\big),\end{aligned}$$

completing the square coordinatewise. The first sum is $\|F-g\|_\pi^2$; in the second,

$$\omega_i^2 - c_i^2/\rho_i = \omega_i^2(\rho_i - (F_i^{\mathrm{w}})^2\omega_i^2)/\rho_i = \omega_i^2 v_i^{\mathrm{w}}/\rho_i.$$

□

Proof of Proposition C.6. All three are Proposition C.4 plus Hilbert-space facts in $L^2(\pi)$. Note first that every $F \in \mathcal{R}_k$ has entries in $[0,1]$: each φ_t, $t \in [0,\infty]$, does, and $[0,1]$ is closed under products.

(i) By Proposition C.4,

$$R(F) - R(\mathbf{1}) = \|F-g\|_\pi^2 - \|\mathbf{1}-g\|_\pi^2 = \langle F-\mathbf{1}, F+\mathbf{1}-2g\rangle_\pi = \sum_i \pi_i(F_i-1)(F_i+1-2g_i),$$

and the criterion in (i) is the negation of this quantity being ≥ 0. If $g_i \geq 1$ at every accessible i and $F \in [0,1]^d$, each summand is a product of $F_i - 1 \leq 0$ and $F_i + 1 - 2g_i \leq F_i - 1 \leq 0$, hence $R(F) \geq R(\mathbf{1})$: no such F — in particular no $F \in \mathcal{R}_k$, for any k — beats the teacher. (A noiseless teacher is of this kind: $v^{\mathrm{w}} = 0$ gives $\rho_i = (F_i^{\mathrm{w}})^2\omega_i^2$ and $g_i = 1/F_i^{\mathrm{w}} \geq 1$ at every accessible i.) In the other direction, $\partial_t\varphi_{t,i}|_{t=0} = -1/\lambda_i$ and $\pi_i/\lambda_i = \rho_i$, so

$$\frac{d}{dt}R(\varphi_t)\Big|_{t=0^+} = 2\langle \mathbf{1}-g, \partial_t\varphi_t|_{t=0}\rangle_\pi = -2\sum_i \rho_i(1-g_i), \tag{F.1}$$

which is < 0 exactly when $\sum_i \rho_i(1-g_i) > 0$; in that case $R(\varphi_t) < R(\mathbf{1})$ for all small $t > 0$.

For a constant $g \equiv \gamma$ the two cases are $\gamma \geq 1$ and $\gamma < 1$, because (F.1) is then $-2(1-\gamma)\sum_i \rho_i$ and $\sum_i \rho_i > 0$ whenever $\pi \neq 0$. The threshold is $\gamma = 1$: the summand $\pi_i(1-F_i)(1+F_i-2\gamma)$ is ≤ 0 *iff* $\gamma \geq \frac{1}{2}(1+F_i)$, so no coordinate gains when $\gamma \geq \frac{1}{2}(1+\max_i F_i)$, and no $F \in [0,1]^d$ gains when $\gamma \geq 1$.

(ii) $\partial_\theta \|F-g\|_\pi^2 = 2\langle F-g, \partial_\theta F\rangle_\pi$, which vanishes at an interior minimizer. At a boundary minimizer only $\langle F^*-g, \partial_\theta F^*\rangle_\pi \geq 0$ along admissible directions survives, and Remark C.7 (a) realizes a strict inequality inside the model.

(iii) $\varphi_0 = \mathbf{1}$ gives $\mathcal{R}_k \subset \mathcal{R}_{k+1}$, hence $\min_{\mathcal{R}_{k+1}} R \leq \min_{\mathcal{R}_k} R$; both minima are attained, because $(t_1,\dots,t_k) \mapsto \prod_j \varphi_{t_j}$ is continuous on the compact $[0,\infty]^k$ (with $\varphi_\infty = \mathbf{0}$) and R is continuous on $[0,1]^d$. Let F^* minimize R over $\mathcal{R}_k$ and let the new hop be $F(s) := F^*\varphi_s \in \mathcal{R}_{k+1}$, $s \geq 0$, so that $\partial_s F_i(s)|_{s=0} = -F_i^*/\lambda_i$ and

$$\frac{d}{ds} R(F(s))\Big|_{s=0^+} = 2\langle F^*-g, \partial_s F(s)|_{s=0}\rangle_\pi = -2\langle F^*-g, F^*/\lambda\rangle_\pi.$$

If $\langle F^*-g, F^*/\lambda\rangle_\pi > 0$ this is negative and $R(F(s)) < R(F^*)$ for small $s > 0$: the inclusion is strict.

Conversely, if $\min_{\mathcal{R}_{k+1}} R = \min_{\mathcal{R}_k} R$, every $\mathcal{R}_k$-minimizer also minimizes over $\mathcal{R}_{k+1}$; as s ranges over $s \geq 0$ only, the one-sided derivative above must be ≥ 0, i.e. $\langle F^*-g, F^*/\lambda\rangle_\pi \leq 0$. Both halves are one-sided, and neither converse holds: Remark C.7 (a) gives equality with $\langle F^*-g, F^*/\lambda\rangle_\pi < 0$ *strictly*, and Theorem D.12 certifies an instance in which $\langle F^*-g, F^*/\lambda\rangle_\pi < 0$ at the single-ridge optimum — so no infinitesimal second hop helps — and yet a distant two-hop configuration strictly beats every single ridge. Whether equality holds is therefore a global question, settled under explicit conditions in Theorem D.11. □

Proof of Remark C.7. Write $B := \sum_{i\leq m} \lambda_i/(\lambda_i+\tau_{\mathrm{w}})^2$ and $A := \sum_{i\leq m} \lambda_i\omega_i^2/(\lambda_i+\tau_{\mathrm{w}})^2$, so that $\Gamma_{\mathrm{w}} = (\tau_{\mathrm{w}}^2 A + s_{\mathrm{w}}^2)/(n_{\mathrm{w}}(1-\Xi_{\mathrm{w}}))$ by (C.6), with $s_{\mathrm{w}}^2 = \sigma^2 + A_m$ the teacher's effective noise (Section G.1). Since $F_i^{\mathrm{w}}(1-F_i^{\mathrm{w}}) = \lambda_i\tau_{\mathrm{w}}/(\lambda_i+\tau_{\mathrm{w}})^2$ and $v_i^{\mathrm{w}} = \lambda_i\Gamma_{\mathrm{w}}/(\lambda_i+\tau_{\mathrm{w}})^2$,

$$\sum_i \left[F_i^{\mathrm{w}}(1-F_i^{\mathrm{w}})\omega_i^2 - v_i^{\mathrm{w}}\right] = \tau_{\mathrm{w}} A - \Gamma_{\mathrm{w}} B, \tag{F.2}$$

and at an accessible i, $g_i > 1$ is $v_i^{\mathrm{w}} < F_i^{\mathrm{w}}(1-F_i^{\mathrm{w}})\omega_i^2$, i.e. $\Gamma_{\mathrm{w}} < \tau_{\mathrm{w}}\omega_i^2$.

Subtracting $\sum_{i\leq m}(F_i^{\mathrm{w}})^2 = n_{\mathrm{w}}\Xi_{\mathrm{w}}$ from $\sum_{i\leq m} F_i^{\mathrm{w}}$ gives $\tau_{\mathrm{w}} B = \sum_{i\leq m} F_i^{\mathrm{w}} - n_{\mathrm{w}}\Xi_{\mathrm{w}}$, so the two forms of the fixed point (C.5) ($\sum_{i\leq m} F_i^{\mathrm{w}} = n_{\mathrm{w}}$ for MNI, and $= n_{\mathrm{w}}(1-\lambda_{\mathrm{w}}/\tau_{\mathrm{w}})$ for ridge) give

$$n_{\mathrm{w}}(1-\Xi_{\mathrm{w}}) = \tau_{\mathrm{w}} B \quad \text{(MNI)}, \qquad n_{\mathrm{w}}(1-\Xi_{\mathrm{w}}) = \tau_{\mathrm{w}} B + n_{\mathrm{w}}\lambda_{\mathrm{w}}/\tau_{\mathrm{w}} \quad \text{(ridge)}. \tag{F.3}$$

(a) A prior flat on the head gives $A = \omega^2 B$, so by (F.3) the condition $\Gamma_{\mathrm{w}} < \tau_{\mathrm{w}}\omega^2$ reads $\tau_{\mathrm{w}}^2\omega^2 B + s_{\mathrm{w}}^2 < \tau_{\mathrm{w}}\omega^2(\tau_{\mathrm{w}} B + n_{\mathrm{w}}\lambda_{\mathrm{w}}/\tau_{\mathrm{w}})$, i.e. exactly $s_{\mathrm{w}}^2 < n_{\mathrm{w}}\lambda_{\mathrm{w}}\omega^2$. Under that hypothesis $g_i > 1$ for every $i \leq m$; since $R(F) = \sum_i \pi_i (F_i - g_i)^2 + \mathrm{Irr}$ and $F_i \in [0,1]$, every accessible summand is minimized at $F_i = 1$, and the inaccessible ones ($\pi_i = 0$) do not depend on F.

As $\mathbf{1} = \varphi_0 \in \mathcal{R}_k$, the minimum over *every* $\mathcal{R}_k$ is attained at $F^* = \mathbf{1}$, i.e. at the boundary point $t_1 = \cdots = t_k = 0$, and $\min_{\mathcal{R}_{k+1}} R = \min_{\mathcal{R}_k} R = R(\mathbf{1})$. By (F.1), $\langle \mathbf{1}-g, \partial_t \varphi_t|_{t=0}\rangle_\pi = \sum_i \rho_i(g_i - 1) > 0$, and the pairing of (iii) is its negative, $\langle \mathbf{1}-g, \mathbf{1}/\lambda\rangle_\pi = \sum_i \rho_i(1-g_i) < 0$.

(b) For an MNI teacher (F.3) turns $\Gamma_{\mathrm{w}} B$ into $(\tau_{\mathrm{w}}^2 A + s_{\mathrm{w}}^2)/\tau_{\mathrm{w}}$, so (F.2) collapses to $-s_{\mathrm{w}}^2/\tau_{\mathrm{w}}$; since $\rho_i(1-g_i) = \rho_i - c_i = v_i^{\mathrm{w}} - F_i^{\mathrm{w}}(1-F_i^{\mathrm{w}})\omega_i^2$, the derivative (F.1) is twice (F.2), i.e. $-2s_{\mathrm{w}}^2/\tau_{\mathrm{w}} < 0$ — which is (D.9). At the other endpoint put $u = 1/t$: $\varphi_{1/u,i} = \lambda_i u/(1+\lambda_i u)$ has $\partial_u \varphi_{1/u,i}|_{u=0} = \lambda_i$, so $\frac{d}{du} R(\varphi_{1/u})|_{u=0^+} = 2\langle \mathbf{0}-g, \lambda\rangle_\pi = -2\sum_i \pi_i g_i \lambda_i = -2\sum_i \lambda_i^2 c_i$, which is < 0 as soon as the teacher passes some signal ($c \neq 0$). Neither endpoint of the compact interval $[0,\infty]$ can then minimize the continuous function $t \mapsto R(\varphi_t)$, so its minimizer is interior. □

G Proof of the composition theorem (Theorem D.1)

Theorem D.1 (Exact W2S risks). *Under Assumption C.1, also permitting $\sigma^2 = 0$, write $p_i := (w_i^*)^2$ for the* realized *signal profile and $\bar{\mathcal{E}}[p]$ for the displayed limit with p_i in place of ω_i^2 throughout. The two readings are different statements.*

Quenched statement. *Conditionally on a realized signal sequence of bounded energy whose channel sums converge, the three risks converge in probability as $d \to \infty$ to $\bar{\mathcal{E}}_w[p]$, $\bar{\mathcal{E}}_s[p]$, $\bar{\mathcal{E}}_c[p]$.*

Annealed statement. *The finite-d limit formulas are affine in the realized profile, so their prior-averages are* exactly *the displayed ω_i^2-forms; hence $\mathbb{E}\,\mathcal{E}_{\rm s} \to \bar{\mathcal{E}}_s$ (and likewise for $\mathcal{E}_{\rm w}$, $\mathcal{E}_{\rm c}$) whenever the risks are uniformly integrable, and $\mathcal{E}_{\rm s} \xrightarrow{\mathbb{P}} \bar{\mathcal{E}}_s$ whenever in addition the realized channel sums concentrate — for which*

$$\sum_i \lambda_i^2 \operatorname{Var}\left((w_i^*)^2\right) \to 0$$

suffices, and which holds inside Assumption C.1 for every prior with a bounded fourth-moment ratio $\mathbb{E}(w_i^*)^4 = O(\omega_i^4)$ but not *for the whole variance-only class (C.1) (Remark G.4, where a prior inside Assumption C.1 is exhibited for which the annealed convergence in probability fails).*

The two readings evaluate one and the same functional — at ω_i^2 (annealed), at p_i (quenched), with $\bar{A}_\psi = \lim \sum_{i>m} \lambda_i \omega_i^2$ becoming $\lim A_m$; what differs is the mode of convergence, and the two clauses above each state their own. The functionals are

$$\begin{aligned} \bar{\mathcal{E}}_w &= \bar{A}_\psi + B_{\rm w} + V_{\rm w}, \qquad \bar{\mathcal{E}}_c = B_{\rm c} + V_{\rm c}, \\ \bar{\mathcal{E}}_s &= \underbrace{\sum_i \lambda_i \left(1 - F_i^{\rm s} F_i^{\rm w}\right)^2 \omega_i^2}_{\text{double-shrinkage bias (incl. tail loss)}} + \underbrace{\sum_i \lambda_i (F_i^{\rm s})^2\, v_i^{\rm w}}_{\text{teacher noise, filtered}} + \underbrace{\sum_i \lambda_i\, v_i^{\rm s}}_{\text{student self-noise}}, \end{aligned} \tag{D.2}$$

where (B, V) are the stage bias/variance of (C.7) for the teacher stage $(\Sigma_H, n_{\rm w}, \omega_H^2, \sigma^2 + \bar{A}_\psi)$ and the ceiling stage $(\Sigma, n_{\rm s}, \omega^2, \sigma^2)$; every displayed sum denotes the limit of the corresponding finite-d channel sum, which exists by Assumption C.1. The ridge variants hold with $\tau(\lambda)$ from (C.5), for the teacher and student stages and the supervised reference.

The proof uses one distributional input, the Gaussian-design theorem of Han & Xu (2023, Thm. 2.2), restated in Lemma F.1. We first extend its asymptotic conclusion to zero label noise. The extension requires a small-noise stability estimate, not a rate uniform in the noise floor. Linear and diagonal quadratic functionals then form a class that is closed under conditioning on an independent next stage.

G.1 A SINGLE-STAGE LAW INCLUDING ZERO NOISE

Write $X = ZC^{1/2} \in \mathbb{R}^{n\times p}$, where Z has independent standard Gaussian entries, $C = \operatorname{diag}(c_i)$, and

$$c_i \in [M^{-1}, M], \qquad p/n \in [1+\delta_0, K], \qquad \eta \in [0, \eta_0], \qquad \|u\| \le R, \qquad s^2 \in [0, S].$$

All five bounds are fixed as $n \to \infty$. Let $\widehat{b}$ fit $Xu + s\zeta$, with $\zeta \sim \mathcal{N}(0, I_n)$ independent of Z, using ridge η or minimum norm at $\eta = 0$. The parameters F, τ, Ξ, v are those of Definition C.2, with target u and noise variance s^2.

Lemma G.1 (Fixed-point bounds). *Let $B_C = \sum_i c_i/(c_i+\tau)^2$. Then*

$$n(1-\Xi) = \frac{n\eta}{\tau} + \tau B_C, \qquad \tau \ge \frac{\delta_0}{M}, \qquad 1 - \Xi \ge \frac{\delta_0}{M^2 + \delta_0}. \tag{G.1}$$

Consequently $\max_i v_i \le C_0/n$ and $\sum_i v_i \le C_0$, uniformly under the bounds above. Here and below a constant may depend on the fixed bounds, but not on the target or its coordinates.

Proof. Subtract $\sum_i F_i^2$ from n, and use $\sum_i F_i = n(1-\eta/\tau)$ and $F_i(1-F_i) = \tau c_i/(c_i+\tau)^2$. Also $n \ge \sum_i F_i \ge p/(1+M\tau)$, which proves the lower bound on τ. Since $\Xi \le \max_i F_i \le M/(M+\tau)$, the gap follows. The numerator of Γ is at most $MR^2 + S$; hence $\Gamma = O(n^{-1})$. Finally, $\sum_i v_i = \Gamma B_C \le (MR^2+S)/\tau$. □

Lemma G.2 (Removing the label-noise floor). *For every fixed $L < \infty$ and $\epsilon > 0$,*

$$\Pr\left(\left|h(\widehat{b}) - \mathbb{E}_g h(b^{\rm se})\right| > \epsilon\right) \longrightarrow 0, \qquad b_i^{\rm se} = F_i u_i + \sqrt{v_i}\, g_i, \tag{G.2}$$

uniformly over the bounded parameters above and globally L-Lipschitz functions $h : \mathbb{R}^p \to \mathbb{R}$. Uniformity concerns the probability bound for each such choice; it does not assert one event for all functions or targets. In particular, $s^2 = 0$ is allowed.

Proof. **1. Couple a positive-noise problem to the given one.** For a fixed $a > 0$, add independent noise $\sqrt{a}\, z$ to the labels, and denote the resulting estimator by $\widehat{b}_a$. Put $G = XX^\top$. The singular-value formula for ridge gives, also at $\eta = 0$,

$$\mathbb{E}\|\widehat{b}_a - \widehat{b}\|^2 \le a\,\mathbb{E}\,\mathrm{tr}(G^{-1}) \le aM\,\mathbb{E}\,\mathrm{tr}((ZZ^\top)^{-1}) = aM\frac{n}{p-n-1} \le C_1 a.$$

The last inequality holds for all sufficiently large n. For completeness, a diagonal entry of $(ZZ^\top)^{-1}$ is the reciprocal of the squared length of one Gaussian row projected off the other $n-1$ rows. That length has law χ^2_{p-n+1}, whose reciprocal mean is $1/(p-n-1)$. This proves the trace identity without a spectral limit.

2. Couple the two Gaussian surrogates. The fixed point does not depend on label noise. Increasing s^2 by a adds

$$\Delta v_i = \frac{ac_i}{n(1-\Xi)(c_i+\tau)^2}$$

to its coordinate variances. Couple the surrogates using the same g. The inequality $(\sqrt{x+y}-\sqrt{x})^2 \le y$ and (G.1) give

$$\mathbb{E}_g\|b_a^{\mathrm{se}} - b^{\mathrm{se}}\|^2 \le \sum_i \Delta v_i = \frac{aB_C}{n(1-\Xi)} \le \frac{a}{\tau} \le C_2 a. \tag{G.3}$$

3. Take the limits in their specified order. At each fixed $a > 0$, the perturbed problem satisfies the noise-floor assumption of Lemma F.1: $s^2 + a \in [a, S+a]$. Its other hypotheses hold uniformly because $\|C\| \le M$, $\mathrm{tr}(C^{-1})/p \le M$, and the aspect ratio stays away from one. A target ball of radius R is reduced to the unit ball by scaling the labels and target by $\max\{R, 1\}$. Thus, for fixed a, the probability of $|h(\widehat{b}_a) - \mathbb{E}_g h(b_a^{\mathrm{se}})| > \epsilon/3$ tends to zero uniformly. The surrogate expectation changes by at most $L\sqrt{C_2 a}$, and Markov's inequality bounds the estimator perturbation probability by $9L^2C_1 a/\epsilon^2$. Choosing a small enough for the deterministic surrogate difference, taking $n \to \infty$, and then $a \downarrow 0$ proves (G.2). The imported constant may diverge as $a \downarrow 0$; no step needs it to remain bounded. The argument proves convergence, without an explicit rate for a dimension-dependent choice of a. □

Lemma G.3 (Aggregate single-stage law)**.** *For deterministic weights q and vectors z with $\|q\|_\infty \le D$ and $\|z\| \le U$, under the same bounds,*

$$\sum_i q_i \widehat{b}_i^2 - \sum_i q_i(F_i^2 u_i^2 + v_i) \xrightarrow{\mathbb{P}} 0, \tag{G.4}$$

$$z^\top \widehat{b} - \sum_i z_i F_i u_i \xrightarrow{\mathbb{P}} 0. \tag{G.5}$$

The failure probabilities are uniform over the stated bounded choices. The conclusions also hold conditionally when these choices are measurable with respect to a sigma-algebra independent of the new design and noise.

Proof. The linear claim is Lemma G.2 with $h(b) = z^\top b$. For $q_i \ge 0$, apply it to the globally $\sqrt{D}$-Lipschitz function $h(b) = (\sum_i q_i b_i^2)^{1/2}$. For $Y = \sum_i q_i (b_i^{\mathrm{se}})^2$, independence of the Gaussian coordinates gives

$$\mathrm{Var}_g(Y) = \sum_i q_i^2(4F_i^2u_i^2v_i + 2v_i^2) = O(n^{-1}), \qquad \mathbb{E}_g Y = O(1).$$

Moreover $|\mathbb{E}_g\sqrt{Y} - \sqrt{\mathbb{E}_g Y}| \le (\mathbb{E}_g|Y - \mathbb{E}_g Y|)^{1/2} = O(n^{-1/4})$. Squaring the convergent, tight square roots proves the quadratic claim. For signed weights, apply this argument to q_+ and q_- separately. Conditioning is legitimate because the bound is uniform over the fixed target ball, not because of a simultaneous event over that ball. □

G.2 Composition by conditioning

1. Fix the target and reduce the teacher problem. Work first with a deterministic signal w^* of bounded Euclidean norm, and write $p_i = (w_i^*)^2$. The Gaussian head and tail designs are independent, so, exactly,

$$y = X_H w_H^* + \varepsilon, \qquad \varepsilon \mid w^* \sim \mathcal{N}(0, s_{\mathrm{w}}^2 I), \qquad s_{\mathrm{w}}^2 = \sigma^2 + \sum_{i>m} \lambda_i p_i, \tag{G.6}$$

with ε independent of X_H. The variance is bounded and may equal zero. Applying Lemma G.3 on the head, and padding by zeros, gives the teacher mean profile $F_i^{\mathrm{w}} w_i^*$ and second-moment profile $\rho_i = (F_i^{\mathrm{w}})^2 p_i + v_i^{\mathrm{w}}$ on every required aggregate.

2. Control the random target handed to the student. The deterministic part of a ridge fit is a contraction. The inverse-Gram bound in Lemma G.2 therefore gives $\mathbb{E}[\|\beta_{\mathrm{w}}\|^2 \mid w^*] \le C(1+\|w^*\|^2)$, uniformly in dimension. In particular,

$$\lim_{B\to\infty} \sup_d \Pr(\|\beta_{\mathrm{w}}\| > B \mid w^*) = 0. \tag{G.7}$$

On $\{\|\beta_{\mathrm{w}}\| \le B\}$, condition on all teacher data. The next design is independent, so Lemma G.3 applies to its now-fixed target β_{w}. Apply its quadratic part with $q_i = \lambda_i$ and its linear part with $z_i = \lambda_i w_i^*$. First let $d \to \infty$ at fixed B, then remove localization. This proves

$$\mathcal{E}_{\mathrm{s}} - \Phi_s(\beta_{\mathrm{w}}) \xrightarrow{\mathbb{P}} 0, \qquad \Phi_s(b) = \sum_i \lambda_i \big[(F_i^{\mathrm{s}} b_i - w_i^*)^2 + v_i^{\mathrm{s}}(b)\big].$$

3. Evaluate the teacher functional. Since the student's fluctuation level is linear in b_i^2,

$$\Phi_s(b) = \sum_i \lambda_i p_i - 2\sum_i \lambda_i F_i^{\mathrm{s}} w_i^* b_i + \sum_i Q_i b_i^2,$$
$$Q_i = \lambda_i \left[(F_i^{\mathrm{s}})^2 + \frac{\Xi_s}{1-\Xi_s}(1-F_i^{\mathrm{s}})^2\right]. \tag{G.8}$$

The weights Q_i and the norm of the linear coefficient are bounded by Lemma G.1. A second application of Lemma G.3, now to the teacher, replaces b_i^2 and $b_i w_i^*$ by ρ_i and $F_i^{\mathrm{w}} p_i$ in these sums. Expanding (G.8) yields

$$\Phi_s(\beta_{\mathrm{w}}) - \sum_i \lambda_i \big[(1 - F_i^{\mathrm{s}} F_i^{\mathrm{w}})^2 p_i + (F_i^{\mathrm{s}})^2 v_i^{\mathrm{w}} + v_i^{\mathrm{s}}(\rho)\big] \xrightarrow{\mathbb{P}} 0.$$

This is the claimed finite-dimensional deterministic equivalent. Convergence of its channel sums names the limiting value, but is not needed for approximation by the finite-dimensional formula.

4. Teacher, reference, and ridge variants. Each single-stage risk is a linear combination of the two aggregate functionals and the fixed signal energy. The same argument therefore proves the teacher and supervised-reference formulas. Every preceding step allows fixed nonnegative ridges and $\sigma^2 \ge 0$; the reference may also be ridged. No separate pseudoinverse trace equivalent is needed.

G.3 Random signals and expected risks

Remark G.4 (Conditional and prior-averaged statements). The preceding proof is uniform on each bounded signal ball. For a random signal with tight norm, first restrict to such a ball and then remove the restriction. This proves approximation by the formula evaluated at the *realized* profile $p_i = (w_i^*)^2$. Each risk formula is affine in this profile; its expectation is exactly the formula evaluated at ω_i^2. Its coefficients are bounded under the spectral and aspect-ratio bounds, by (G.8) and the teacher formula. Independence of the signal coordinates therefore makes

$$V_d := \sum_i \lambda_i^2 \operatorname{Var}((w_i^*)^2) \longrightarrow 0 \tag{G.9}$$

a sufficient condition for concentration about the prior-averaged formula. The bounded fourth-moment ratio and $\sup_{d,i} d\omega_i^2 < \infty$ in Theorem 3.2 give $V_d = O(d^{-1})$.

Second moments alone do not give this conclusion. With $\Sigma = I$ and independent $w_i^* = \pm 1$ of probability $1/(2d)$ each, and zero otherwise, one has $\omega_i^2 = 1/d$ but $\|w^*\|^2 \Rightarrow \mathrm{Poisson}(1)$. For a minimum-norm supervised reference with $n_{\mathrm{s}}/d \to \alpha \in (0,1)$, the realized formula is $(1-\alpha)\|w^*\|^2 + \sigma^2\alpha/(1-\alpha)$, which does not concentrate at its expectation. This prior satisfies Assumption C.1 but not the bounded fourth-moment ratio in the main theorem.

Nor does Gaussianity itself imply (G.9). For $\lambda_i = i^{-a}$ and $\omega_i^2 = Gi^{-b}$,

$$V_d = 2G^2 \sum_{i \le d} i^{-2(a+b)} \longrightarrow 2G^2\zeta(2a+2b) > 0. \tag{G.10}$$

Failure of a sufficient condition does not prove that a particular risk fails to concentrate. Power-law spectra need their own approximation and concentration arguments. Also, variance-only profiles do not determine almost-sure source exponents: choosing independent nonzero probabilities with summable total can leave only finitely many nonzero coordinates almost surely. Our power-law rates remain statements about the prior-averaged deterministic formulas.

Lemma G.5 (Uniform integrability under fourth moments)**.** *If $\sup_d \mathbb{E}\|w^*\|^4 < \infty$ and σ^2 is bounded, the teacher, reference, and all downstream chain risks have uniformly bounded second moments for all sufficiently large dimensions under the bounded-spectrum proportional assumptions. The bound for the chain is uniform over its length.*

Proof. Conditional on w^*, the teacher noise variance is at most $\sigma^2 + M\|w^*\|^2$. For $W = ZZ^\top$ and $p - n > 3$, the row-projection calculation used above also gives $\mathbb{E}(W^{-1})_{aa}^2 = [(p-n-1)(p-n-3)]^{-1}$. Thus Cauchy–Schwarz yields

$$\mathbb{E}(\operatorname{tr} W^{-1})^2 \le \frac{n^2}{(p-n-1)(p-n-3)} = O(1).$$

Conditional Gaussian fourth moments and $\operatorname{tr}(W^{-2}) \le (\operatorname{tr} W^{-1})^2$ imply

$$\mathbb{E}\|X^\dagger(s\zeta)\|^4 \le 3M^2 s^4\, \mathbb{E}(\operatorname{tr} W^{-1})^2 = O(s^4).$$

Ridge noise amplification is no larger than minimum-norm amplification. Combining this with contraction of the signal part proves $\sup_d \mathbb{E}\|\beta_{\mathrm{w}}\|^4 < \infty$. Every noiseless downstream fit satisfies $\|\beta^{(t)}\| \le \|\beta_{\mathrm{w}}\|$ pointwise. Finally, $\mathcal{E}(\beta^{(t)})^2 \le 8M^2(\|\beta_{\mathrm{w}}\|^4 + \|w^*\|^4)$. The reference follows by the same single-stage argument. □

The prior conditions in Theorem 3.2 imply $\sup_d \mathbb{E}\|w^*\|^4 < \infty$: expand the fourth power of the norm, using independence and the bounded fourth-moment ratio. Consequently its risk approximations hold in expectation as well as in probability. More general variance-only priors retain the explicit uniform-integrability condition in Theorem D.1.

Remark G.6 (Scope of the argument). The small-noise argument takes dimension to infinity at each fixed added variance, and only then lets that variance vanish. It establishes neither an explicit convergence rate nor a bound uniform over deteriorating spectra. Both the inverse-Gram bound and the imported surrogate depend on the lower spectral bound. It therefore does not transfer the power-law formula rates to random estimators, and supplies no error bound relative to a vanishing risk. Independent batches are equally essential: each conditioning step uses a target independent of the design being fitted.

H Structural corollaries

H.1 Isotropic minimum-norm imitation: proof of Corollary D.3

With $\Sigma = I_d$, decompose $\widehat{w}_{\mathrm{s}} - w^* = \widetilde{P}(\beta_{\mathrm{w}} - w^*) - (I - \widetilde{P})w^*$. The two terms are orthogonal (ranges of $\widetilde{P}$ and $I - \widetilde{P}$), so

$$\mathcal{E}_{\mathrm{s}} = \|\widetilde{P}(\beta_{\mathrm{w}} - w^*)\|^2 + \|(I - \widetilde{P})w^*\|^2.$$

For $n_s \le d$, $\widetilde{P}$ is a uniformly random rank-n_s orthogonal projection independent of (β_w, w^*) (rotational invariance of the isotropic Gaussian design; for $n_s > d$ it is the identity and the student is the exact clone), so $\mathbb{E}\|\widetilde{P}z\|^2 = \frac{n_s}{d}\|z\|^2$ for any independent z, giving $\mathbb{E}\mathcal{E}_s = \frac{n_s}{d}\mathbb{E}\|\beta_w - w^*\|^2 + (1 - \frac{n_s}{d})\mathbb{E}\|w^*\|^2$, which is (D.4).

The conclusions are statements about expectations: $\mathbb{E}\mathcal{E}_s \ge \min(\mathbb{E}\mathcal{E}_w, r^2)$ and (expected gain $>$ $0 \iff \mathbb{E}\mathcal{E}_w > r^2$) are immediate from the affine identity. Pathwise the inequality can fail (a realized projection nearly aligned with $\beta_w - w^*$ makes $\mathcal{E}_s$ small on that event); only the annealed statement is claimed. □

Two scope notes. First, the prior entered only through $\mathbb{E}\|w^*\|^2 = r^2$; the profile (ω_i^2) never appears, so (D.4) holds at $\Sigma = I_d$ for *every* prior (C.1), not only the isotropic one (checked at $\omega_i^2 \propto i^{-2}$: $\mathbb{E}\mathcal{E}_s$ matches the identity to 0.4 standard errors over 2×10^4 trials). Second, it used the student's *minimum-norm* rule through $\widehat{w}_s = \widetilde{P}\beta_w$, and that hypothesis is not removable.

Lemma H.1 (At $\Sigma = I_d$ a ridge student beats every *minimum-norm* teacher)**.** *Let $\Sigma = I_d$, let the teacher be MNI on the head with $n_w < m - 1$, and let the student be ridge with parameter $\lambda_s \ge 0$. Write*

$$\mathsf{A} := \mathbb{E}\|\beta_w\|^2, \qquad \mathsf{B} := \mathbb{E}\langle \beta_w, w^*\rangle, \qquad \bar{s}_w^2 := \mathbb{E}\big[\sigma^2 + A_m\big] = \sigma^2 + \textstyle\sum_{i>m}\omega_i^2 \;>\; 0,$$

the last being the expected *effective teacher noise: the realized $s_w^2 = \sigma^2 + A_m$ of (D.1) is a function of w^* and has no place in an identity between expectations (Remark H.2).*
Then, as $n_s \to \infty$,

$$\mathbb{E}\mathcal{E}_w - \inf_{\lambda_s \ge 0}\lim_{n_s\to\infty}\mathbb{E}\mathcal{E}_s \;=\; \frac{(\mathsf{A}-\mathsf{B})^2}{\mathsf{A}} \;=\; \frac{\big[\bar{s}_w^2\, n_w/(m - n_w - 1)\big]^2}{\mathsf{A}} \;>\; 0. \tag{H.1}$$

The infimum is attained *iff $\mathsf{B} > 0$, i.e. iff the teacher head band carries signal ($\sum_{i\le m}\omega_i^2 > 0$), and then at $\lambda_s^* = \mathsf{A}/\mathsf{B} - 1 > 0$; if $\mathsf{B} = 0$ — permitted by (C.1), all signal in the tail — it is an infimum only, approached as $\lambda_s \to \infty$, the optimum being the null predictor.*
In either case some finite *ridge already beats the teacher: with $c = (1+\lambda_s)^{-1}$, $\mathbb{E}\mathcal{E}_w - \lim_{n_s\to\infty}\mathbb{E}\mathcal{E}_s = (1-c)\,[\mathsf{A}(1+c) - 2\mathsf{B}]$, and the threshold is explicit rather than asymptotic: since $\mathsf{A} > \mathsf{B}$, the right-hand side is > 0 exactly for $c \in (\max\{0,\, 2c^* - 1\},\, 1)$ with $c^* := \mathsf{B}/\mathsf{A} \in [0, 1)$ — that is, for* every *$\lambda_s > 0$ when $c^* \le \frac{1}{2}$, and for $0 < \lambda_s < 2(1-c^*)/(2c^* - 1)$ when $c^* > \frac{1}{2}$. (*Every *teacher means every teacher this lemma admits: MNI on the head band with $n_w < m - 1$, at $\Sigma = I_d$, with $\bar{s}_w^2 > 0$.*
A ridge *teacher is outside its scope — the inverse-Wishart step that evaluates $\mathsf{A} - \mathsf{B}$ below uses the minimum-norm rule — and nothing is claimed against one here.)*

Proof. **Step 1: take the student limit inside the expected risk.** $\widehat{\Sigma} \to I_d$ almost surely, so $\widehat{w}_s = (\widehat{\Sigma} + \lambda_s I)^{-1}\widehat{\Sigma}\beta_w \to c\,\beta_w$ with $c = (1+\lambda_s)^{-1} \in (0, 1]$, and hence $\mathcal{E}_s = \|\widehat{w}_s - w^*\|^2 \to \|c\beta_w - w^*\|^2$ almost surely.

Passing to expectations is dominated convergence, and the envelope is deterministic: $\widehat{\Sigma} \succeq 0$ puts every eigenvalue of $(\widehat{\Sigma} + \lambda_s I)^{-1}\widehat{\Sigma}$ in $[0, 1]$, so $\|(\widehat{\Sigma} + \lambda_s I)^{-1}\widehat{\Sigma}\|_{op} \le 1$ for every draw and every n_s, whence $\|\widehat{w}_s\| \le \|\beta_w\|$ and $\mathcal{E}_s \le 2\|\beta_w\|^2 + 2\|w^*\|^2$. That envelope does not depend on n_s and is integrable, $\mathbb{E}[2\|\beta_w\|^2 + 2\|w^*\|^2] = 2\mathsf{A} + 2r^2 < \infty$ — A is evaluated below and is finite exactly because $n_w < m - 1$ — so $\lim_{n_s\to\infty}\mathbb{E}\mathcal{E}_s = \mathbb{E}\|c\beta_w - w^*\|^2$.

Step 2: minimize the scalar quadratic. At $\Sigma = I_d$,

$$\mathbb{E}\|c\beta_w - w^*\|^2 = c^2\mathsf{A} - 2c\mathsf{B} + \mathbb{E}\|w^*\|^2$$

is a strictly convex quadratic in c, with unconstrained minimizer $c^* = \mathsf{B}/\mathsf{A}$ and value $\mathbb{E}\|w^*\|^2 - \mathsf{B}^2/\mathsf{A}$. Since c ranges over $(0, 1]$ and $0 \le c^* < 1$ (shown below), that value is attained iff $c^* > 0$, i.e. iff $\mathsf{B} > 0$, and is otherwise the limit as $c \downarrow 0$; either way it is the infimum.

Subtracting it from $\mathbb{E}\mathcal{E}_w = \mathbb{E}\|\beta_w - w^*\|^2 = \mathsf{A} - 2\mathsf{B} + \mathbb{E}\|w^*\|^2$ leaves $\mathsf{A} - 2\mathsf{B} + \mathsf{B}^2/\mathsf{A} = (\mathsf{A}-\mathsf{B})^2/\mathsf{A}$, the first equality.

Step 3: evaluate the teacher moments. For the second, write $\widehat{w}_{\mathrm{w}} = Pw_H^* + X_H^+\varepsilon$ with $P = X_H^+X_H$ the projection onto the n_{w}-dimensional row span and ε the effective noise, which conditionally on w^* is $\mathcal{N}(0, s_{\mathrm{w}}^2 I)$ and independent of X_H; and X_H is independent of w^*.

The cross term vanishes because $\mathbb{E}[\varepsilon \mid X_H, w^*] = 0$, so

$$\mathsf{B} = \mathbb{E}\langle Pw_H^*, w_H^*\rangle = \tfrac{n_{\mathrm{w}}}{m}\sum_{i\le m}\omega_i^2$$

and

$$\mathsf{A} \;=\; \mathsf{B} + \mathbb{E}\|X_H^+\varepsilon\|^2 \;=\; \mathsf{B} + \mathbb{E}\big[s_{\mathrm{w}}^2\big]\;\mathbb{E}\,\mathrm{tr}[(X_HX_H^\top)^{-1}] \;=\; \mathsf{B} + \bar{s}_{\mathrm{w}}^2\,\frac{n_{\mathrm{w}}}{m-n_{\mathrm{w}}-1},$$

the middle step because $s_{\mathrm{w}}^2 = \sigma^2 + A_m$ is w^*-measurable while $X_H \perp w^*$, so the two factors separate — this is exactly where a realized value may not stand in for an expectation — and the last by the inverse-Wishart mean $\mathbb{E}[(X_HX_H^\top)^{-1}] = I_{n_{\mathrm{w}}}/(m-n_{\mathrm{w}}-1)$, valid for $n_{\mathrm{w}} < m-1$.

Hence $\mathsf{A}-\mathsf{B} = \bar{s}_{\mathrm{w}}^2\, n_{\mathrm{w}}/(m-n_{\mathrm{w}}-1) > 0$, which gives the second equality, $c^* = \mathsf{B}/\mathsf{A} < 1$, and strict positivity. □

Remark H.2 (Why the expectation cannot be moved inside). A_m is random under (C.1) and (H.1) is quadratic in $\bar{s}_{\mathrm{w}}^2$, so putting the realized s_{w}^2 there is not a harmless abbreviation. Witness: $\Sigma = I_d$ with a single tail coordinate ($d = m+1$) taking the value 0 with probability $\frac{1}{2}$ and $\pm\sqrt{2}$ with probability $\frac{1}{4}$ each — mean 0, variance 1, hence a member of (C.1) — independent of a nondegenerate head signal, and $\sigma^2 = 1$.

Then $s_{\mathrm{w}}^2 \in \{1, 3\}$ with $\bar{s}_{\mathrm{w}}^2 = 2$, so with $c := n_{\mathrm{w}}/(m-n_{\mathrm{w}}-1)$ the numerator of (H.1) is $4c^2$, whereas the realized reading would give c^2 or $9c^2$ according to the draw. Every expectation in (H.1) is over the prior as well as the design. The version conditional on w^* is equally true, but then A, B and $\mathbb{E}\mathcal{E}_{\mathrm{w}}$ must *all* be read conditionally, whereas Corollary D.3, with which the lemma is contrasted, is annealed.

At $d = 40$, $m = 36$, $n_{\mathrm{w}} = 20$, $\sigma^2 = 0.01$ and the isotropic prior ($\bar{s}_{\mathrm{w}}^2 = \sigma^2 + (d-m)r^2/d = 0.11$, $\mathsf{B} = 0.5$, $\mathsf{A} = 0.64667$) this reads: the teacher is competent ($\mathbb{E}\mathcal{E}_{\mathrm{w}} = 0.6467 < r^2 = 1$) and the optimally shrunk clone beats it by 0.033265 at $\lambda_{\mathrm{s}}^* = 0.29333$ in the abundant-student-data limit. The student has ample capacity to represent its teacher. Thus "beating a competent teacher requires spectral anisotropy", and "the student wins only because $n_{\mathrm{s}} < d$", are statements about the minimum-norm student, not about the pipeline: with explicit regularization neither survives.

H.2 Super-ceiling threshold: proof of Corollary D.5

None of the fixed points $(\tau_{\mathrm{w}}, \Xi_{\mathrm{w}}, \tau_{\mathrm{s}}, \Xi_{\mathrm{s}})$ depend on σ^2, so the σ^2-dependence of (D.2) enters only linearly, through Γ_{w} (hence v^{w}), and through Γ_{s} via the profile $\rho = (F^{\mathrm{w}})^2\omega^2 + v^{\mathrm{w}}$. Both $\bar{\mathcal{E}}_s$ and $\bar{\mathcal{E}}_c$ are therefore exactly affine in σ^2 (confirmed to machine precision numerically). Differentiating,

$$\beta_{\mathrm{c}} = \frac{\Xi_{\mathrm{s}}}{1-\Xi_{\mathrm{s}}}, \qquad \beta_{\mathrm{s}} = \frac{1}{n_{\mathrm{w}}\,(1-\Xi_{\mathrm{w}})}\sum_{i\le m}\frac{\lambda_i^2}{(\lambda_i+\tau_{\mathrm{s}})^2(\lambda_i+\tau_{\mathrm{w}})^2}\Big(\lambda_i^2 + \tau_{\mathrm{s}}^2\tfrac{\Xi_{\mathrm{s}}}{1-\Xi_{\mathrm{s}}}\Big), \tag{H.2}$$

where the two summands are the noise passed through both filters and the noise re-emitted by the student's own interpolation.

Writing $\bar{\mathcal{E}}_s - \bar{\mathcal{E}}_c = [\bar{\mathcal{E}}_s(0) - \bar{\mathcal{E}}_c(0)] - (\beta_{\mathrm{c}} - \beta_{\mathrm{s}})\,\sigma^2$, the super-ceiling set is the sub-interval of $[0,\infty)$ (possibly empty, bounded, or everything) determined by the two signs, which yields exactly the four regimes of Corollary D.5 (with boundary $\sigma_\star^2 = (\bar{\mathcal{E}}_s(0) - \bar{\mathcal{E}}_c(0))/(\beta_{\mathrm{c}} - \beta_{\mathrm{s}})$ in regimes (i) and (iv)). □

Remark H.3 (The zero-noise intercepts). The $\sigma^2 = 0$ intercepts are the affine continuation: the convergence theorems are invoked at $\sigma^2 > 0$, and the intercepts are their $\sigma^2 \downarrow 0$ limits, by continuity of the limit formulas.

Remark H.4 (The zero-noise sign is *not* determined by teacher imperfection). It is tempting to argue that an imperfect teacher forces $\bar{\mathcal{E}}_s(0) > \bar{\mathcal{E}}_c(0)$; this is *false*, and the witness below refutes it.

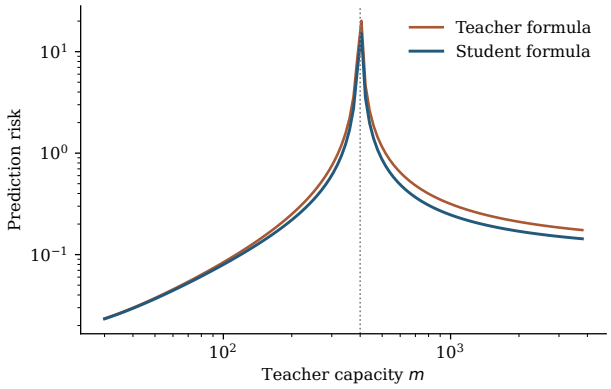


Figure 4: Deterministic teacher and student risk formulas vs. teacher capacity m (a=1.5, b=1, n_{w}=400, n_{s}=1000, d=4000, σ^2=0.25). The teacher's interpolation-threshold divergence at $m = n_{\mathrm{w}}$ passes to the student through the retained teacher fluctuation term. The curves are formula evaluations, not Monte Carlo measurements.

Explicit regime-(ii) witness (bounded condition number 100, inside Assumption C.1): two-level spectrum with 5% of coordinates at $\lambda = 100$ and 95% at $\lambda = 1$, isotropic prior, $\psi = 0.9$, $\bar{\kappa_{\mathrm{w}}} = 3$, $\bar{\kappa_{\mathrm{s}}} = 20$. There $\bar{\mathcal{E}}_s(0) = 0.3911 < 0.4497 = \bar{\mathcal{E}}_c(0)$ while the teacher is far from perfect ($\bar{\mathcal{E}}_w(0) = 0.1485$), as evaluated from the deterministic formulas. Since also $\beta_{\mathrm{s}} = 0.442 < 0.757 = \beta_{\mathrm{c}}$, the student beats the ceiling at *every* noise level.

Mechanism: with a small budget the ceiling pays heavy interpolation self-noise to fit the full-norm w^*, while the student fits the teacher's pre-shrunk β_{w}, whose smaller norm lowers the self-noise by more than the inherited teacher error costs — weak-to-strong training as pure implicit regularization, beating even noiseless supervision.

Two further possible implications — that every regime-(ii) instance has $\bar{\mathcal{E}}_c(0) > \bar{\mathcal{E}}_w(0)$, and that $\bar{\mathcal{E}}_c(0) \le \bar{\mathcal{E}}_w(0)$ forces $\bar{\mathcal{E}}_s(0) > \bar{\mathcal{E}}_c(0)$ — are both false, and a single formula instance disproves both. Two-level spectrum with 44% of coordinates at $\lambda = 20$ and 56% at $\lambda = 1$, signal shares $(0.03, 0.97)$ on the two blocks, $\psi = 0.60$, $\bar{\kappa_{\mathrm{w}}} = 2.4$, $\bar{\kappa_{\mathrm{s}}} = 4.7$: there $\bar{\mathcal{E}}_w(0) = 2.0096 > \bar{\mathcal{E}}_c(0) = 1.5070 > \bar{\mathcal{E}}_s(0) = 1.4148$ with $\beta_{\mathrm{s}} = 0.4507 < 0.6586 = \beta_{\mathrm{c}}$ — regime (ii), with the ceiling *better* than the teacher and the student below both. A general characterization of the zero-noise sign remains open.

Near the teacher's interpolation threshold, $1 - \Xi_{\mathrm{w}} \to 0$ inflates $\beta_{\mathrm{s}} \to \infty$ and the attenuation condition must fail. A moderate Ξ_{w} does not rescue it. Each factor $\lambda_i^2/(\lambda_i + \tau_{\mathrm{s}})^2 \cdot \lambda_i^2/(\lambda_i + \tau_{\mathrm{w}})^2 \le 1$ damps the transfer, and the condition does hold across most of our scanned families (e.g. $\beta_{\mathrm{s}} = 0.863 < 0.929 = \beta_{\mathrm{c}}$ at the power-law configuration $a = 1.5$, $b = 1$, $\psi = 0.3$, $\bar{\kappa_{\mathrm{w}}} = 3$, $\sigma^2 = 0.25$, $r^2 = 1$, read at $\bar{\kappa_{\mathrm{s}}} = 4$), but it can fail with the teacher far from its threshold: the regime-(iii) instance displayed after Corollary D.5 has $\Xi_{\mathrm{w}} = 0.5496$ and $\beta_{\mathrm{s}} = 0.77859 > 0.58778 = \beta_{\mathrm{c}}$. The sign of $\beta_{\mathrm{c}} - \beta_{\mathrm{s}}$ is a joint property of both budgets, not a teacher-side diagnostic.

H.3 Teacher capacity formulas and the transferred double descent

As a function of m at fixed n_{w}: on the underparametrized branch ($m < n_{\mathrm{w}}$) the teacher stage is ordinary least squares, $\bar{\mathcal{E}}_w = \bar{A}_\psi + \sigma^2_{\mathrm{eff}} \frac{m/n_{\mathrm{w}}}{1-m/n_{\mathrm{w}}}$-type with $\bar{A}_\psi$ decreasing and the variance increasing in m — the classical U-shape; at $m = n_{\mathrm{w}}$ the variance factor $1/(1 - \Xi_{\mathrm{w}})$ diverges; on the overparametrized branch (C.7) applies.

Figure 4 illustrates the subsequent decline of both risk formulas in one configuration, consistent with the regularizing effect of overparametrization studied in benign overfitting (Bartlett et al., 2020). The student inherits the spike through the passed-noise term $\sum_i \lambda_i (F_i^{\mathrm{s}})^2 v_i^{\mathrm{w}} \propto \Gamma_{\mathrm{w}} \uparrow \infty$. This illustration does not establish monotonicity for general spectra or equality of teacher and student capacity optima.

I Proof of the sweet-spot theorem (Theorem D.6)

Theorem D.6 (Imitation sweet spot)**.** *Consider the MNI student.*

(i) Clone plateau (finite d, exact). *If $n_s \geq d$ then $\widehat{w}_s = \beta_w$ almost surely: the student is an exact clone, $\mathcal{E}_s = \mathcal{E}_w$ identically, and no weak-to-strong gain is possible.*

(ii) Boundaries; the order of limits is part of the statement. *Read* (D.2) *as the channel calculus of Definition C.2 at a spectrum with $\lambda_i > 0$, with the imitation budget the only free variable and $\bar{\kappa}_s := d/n_s$.*

Then $\bar{\kappa}_s \mapsto \bar{\mathcal{E}}_s$ is continuous on $(1, \infty)$, with $\bar{\mathcal{E}}_s \to \bar{\mathcal{E}}_w$ as $\bar{\kappa}_s \downarrow 1$ (perfect imitation) and $\bar{\mathcal{E}}_s \to r^2$ as $\bar{\kappa}_s \to \infty$ (no imitation), in each of two readings: (ii-a) finite d, limit-formula reading*, in which d is held fixed and finite and n_s ranges over $(0, d)$ — valid for* every *spectrum, power laws included; and* (ii-b) proportional*, in which $\bar{\mathcal{E}}_s(\bar{\kappa}_s)$ is the Theorem D.1 limit at fixed $\bar{\kappa}_s$ — valid under that theorem's regularity, ν compactly supported with $\lambda_{\min} > 0$.*

Neither scope clause is decorative: in the proportional reading at a power-law spectrum the $\bar{\kappa}_s \to \infty$ limit is not *r^2 but the noise floor of Theorem D.14 (ii), which is free of both r^2 and b (Remark I.4 carries the witness).*

(iii) Non-monotonicity, with a closed-form instance. *Consider the k-spike family with all signal on k diverging spikes, unit bulk, and a teacher head covering the spikes.*

In the domain $k < n_w < m$, $2k < d$, and $k < n_s < d$, the limiting weak-to-strong gain has the closed form

$$\bar{\mathcal{E}}_w - \bar{\mathcal{E}}_s \;=\; \sigma^2 \, \frac{(d - n_s)\,(n_w - k)\,(n_s - 2k)}{(d - k)\,(n_s - k)\,(m - n_w)}, \tag{D.5}$$

It is strictly positive iff $\sigma^2 > 0$ and $n_s > 2k$, and vanishes at both ends ($n_s \downarrow 2k$ and $n_s \uparrow d$); at $\sigma^2 = 0$ the gain is identically zero, so the label noise that Assumption C.1 keeps strictly positive is what makes the "iff" bite.

The maximizing budget is

$$n_s^* \;=\; k + \sqrt{k\,(d - k)}, \tag{D.6}$$

which lies strictly inside $(2k, d)$ exactly when $d > 2k$ (Section I.4).

Part (i) is finite-d linear algebra and needs no import. Parts (ii)–(iii) are statements about the composed limit formula (D.2) (Theorem D.1) — the true proportional limit — and are written conditionally on it, as flagged in Remark I.4. Throughout, a stage with eigenvalues $c = (c_i)$, n samples, target second moments u_i^2 and noise s^2 has the channel scalars of Definition C.2, and the composed W2S risk is (D.2). Only the student channel (τ_s, F^s, v^s) depends on κ_s; the teacher outputs are held fixed.

I.1 Master fixed-point identity

Lemma I.1 (Master identity)**.** *Let ν be the limiting spectral distribution of Σ and write $m_\nu(\tau) = \int \frac{\lambda}{\lambda+\tau}\, d\nu$, $\chi(\tau) = \int \frac{\lambda^2}{(\lambda+\tau)^2}\, d\nu$, $\zeta(\tau) = \int \frac{\lambda}{(\lambda+\tau)^2}\, d\nu$. The ridgeless student fixed point is $m_\nu(\tau_s) = 1/\kappa_s$, $\Xi_s = \kappa_s \chi(\tau_s)$, and*

$$1 - \Xi_s \;=\; \kappa_s\, \tau_s\, \zeta(\tau_s) \;=\; \kappa_s \big(m_\nu(\tau_s) - \chi(\tau_s)\big) \;>\; 0. \tag{I.1}$$

Proof. The finite-d identity (G.1) reads $1 - \Xi_s = \frac{\tau_s}{n_s} \sum_i \frac{\lambda_i}{(\lambda_i + \tau_s)^2}$; dividing the sum by d and using $n_s = d/\kappa_s$ sends it to $\kappa_s \tau_s \zeta(\tau_s)$.

The middle equality is the algebraic identity $\chi = m_\nu - \tau\zeta$ (clear coordinatewise: $\frac{\lambda}{\lambda+\tau} - \frac{\tau\lambda}{(\lambda+\tau)^2} = \frac{\lambda^2}{(\lambda+\tau)^2}$). Positivity holds for $\tau_s > 0$, which in turn proves $\Xi_s \in (0, 1)$ whenever the stage is strictly overparametrized. □

This is the exact resolution of the "$0 \times \infty$" encountered at $\kappa_s \downarrow 1$ in Part (ii): the self-noise $B_0^s \, \Xi_s/(1-\Xi_s)$ with $B_0^s = \tau_s^2 P(\tau_s)$, $P(\tau) := \lim \sum_i \lambda_i \rho_i/(\lambda_i+\tau)^2$, becomes by (I.1)

$$\text{self-noise} = \tau_s \, P(\tau_s) \, \chi(\tau_s)/\zeta(\tau_s). \tag{I.2}$$

I.2 Part (i): the clone plateau (exact, finite d)

Pseudo-labels $\widehat{y} = \widetilde{X}\beta_w$ lie exactly in the column space of $\widetilde{X}$ (noiseless, linear).

MNI student. If $n_s \geq d$ and $\operatorname{rank}(\widetilde{X}) = d$ (a.s. for Gaussian $\widetilde{X}$), the system $\widetilde{X}v = \widetilde{X}\beta_w$ has kernel $\{0\}$, so its solution set is the singleton $\{\beta_w\}$; both the minimum-norm interpolator and OLS return $v = \beta_w$. Hence $\widehat{w}_s = \beta_w$ pathwise and $\mathcal{E}_s = \|\Sigma^{1/2}(\beta_w - w^*)\|^2 = \mathcal{E}_w$ identically — W2S gain is impossible once the student can fit its teacher. (Verified to machine precision: $\max_i |(\widehat{w}_s - \beta_w)_i| = 6 \times 10^{-15}$ at $d = 40$, $n_s = 60$.)

Ridge student. $\widehat{w}_s = \widehat{\Sigma}(\widehat{\Sigma} + \lambda_s I)^{-1}\beta_w$ is a strict shrinkage of β_w; as $n_s \to \infty$ ($\geq d$), $\widehat{\Sigma} \to \Sigma$ and $\widehat{w}_s \to F^{(\lambda_s)}\beta_w$ with $F_i^{(\lambda_s)} = \lambda_i/(\lambda_i + \lambda_s)$ — the deterministic *shrunk clone*, whose risk can lie above or below $\mathcal{E}_w$. We use MNI for (ii)–(iii) precisely because its only lever is the sample size n_s.

I.3 Part (ii): continuity and the two boundary limits

Part (ii) has two readings and they are *not* interchangeable; which one is in force is decided by the order in which $d \to \infty$ and n_s-variation are taken. Throughout, $\bar{\mathcal{E}}_s$ means (D.2) evaluated in the channel calculus of Definition C.2 with the student budget the only free variable, $\bar{\kappa_s} = d/n_s$, and the teacher outputs held fixed.

(ii-a) Finite d, every spectrum. Fix $d < \infty$ and $\lambda_1, \ldots, \lambda_d > 0$, and let n_s range over the reals in $(0, d)$; all sums below are finite, so no convergence theorem is needed.

Continuity. $N(\tau) := \sum_{i \leq d} \lambda_i/(\lambda_i + \tau)$ is continuous and strictly decreasing on $(0, \infty)$ with $N(0^+) = d$ and $N(\infty) = 0$, so $n_s \mapsto \tau_s = N^{-1}(n_s)$ is a continuous strictly decreasing bijection $(0, d) \to (0, \infty)$, i.e. $\bar{\kappa_s} \mapsto \tau_s$ is a continuous increasing bijection $(1, \infty) \to (0, \infty)$. By (G.1),

$$n_s \, (1 - \Xi_s) \;=\; \tau_s \, S(\tau_s), \qquad S(\tau) := \sum_{i \leq d} \frac{\lambda_i}{(\lambda_i + \tau)^2} \;>\; 0, \tag{I.3}$$

so $1 - \Xi_s > 0$ for every $\tau_s > 0$ and no denominator in (D.2) vanishes; each summand is a rational function of τ_s with nonvanishing denominator. Hence $\bar{\kappa_s} \mapsto \bar{\mathcal{E}}_s$ is continuous on $(1, \infty)$.

Limit $\bar{\kappa_s} \downarrow 1$ ($n_s \uparrow d$, $\tau_s \to 0^+$). Write $P(\tau) := \sum_{i \leq d} \lambda_i \rho_i/(\lambda_i + \tau)^2$, so that $\Gamma_s = \tau_s^2 P(\tau_s)/(n_s(1 - \Xi_s))$; by (I.3) this is exactly

$$\Gamma_s \;=\; \frac{\tau_s \, P(\tau_s)}{S(\tau_s)} \;\longrightarrow\; \frac{0 \cdot P(0)}{S(0)} \;=\; 0, \tag{I.4}$$

because $P(0) = \sum_i \rho_i/\lambda_i$ and $S(0) = \sum_i \lambda_i^{-1}$ are finite and $S(0) > 0$ — this is where finiteness of d and $\lambda_i > 0$ are used, and it is the exact resolution of the "$0 \times \infty$" at $\Xi_s \to 1$. Hence $v_i^s \to 0$ and the self-noise $\sum_i \lambda_i v_i^s \to 0$, while $F_i^s \to 1$ for each i makes the first two terms of (D.2) tend to $\sum_i \lambda_i (1 - F_i^w)^2 \omega_i^2 + \sum_i \lambda_i v_i^w = \bar{\mathcal{E}}_w$. So $\bar{\mathcal{E}}_s \to \bar{\mathcal{E}}_w$, matching Part (i) across the boundary.

Limit $\bar{\kappa_s} \to \infty$ ($n_s \downarrow 0$, $\tau_s \to \infty$). Now $F_i^s \to 0$ for each i, so the bias term tends to $\sum_i \lambda_i \omega_i^2 = r^2$ and the passed-noise term to 0. For the self-noise, with $E(\tau) := \sum_{i \leq d} \lambda_i^2/(\lambda_i + \tau)^2$, (I.4) gives

$$\sum_i \lambda_i v_i^s \;=\; \Gamma_s \, E(\tau_s) \;=\; \frac{1}{\tau_s} \cdot \frac{\left[\tau_s^2 P(\tau_s)\right] \left[\tau_s^2 E(\tau_s)\right]}{\tau_s^2 S(\tau_s)},$$

and as $\tau \to \infty$ the three bracketed quantities tend to $\sum_i \lambda_i \rho_i$, $\sum_i \lambda_i^2$ and $\sum_i \lambda_i$, all finite with the last positive; so the self-noise is $O(1/\tau_s) \to 0$. Hence $\bar{\mathcal{E}}_s \to r^2$, the null-predictor risk. □

(ii-b) The proportional limit, under bounded condition number. Assume the Theorem D.1 regularity: ν compactly supported with $\lambda_{\min} > 0$, and a finite limiting profile ($\sum \lambda_i \omega_i^2 \to r^2$, $\sum \lambda_i \rho_i < \infty$). Since m_ν is continuous, strictly decreasing, $m_\nu(0) = 1$, $m_\nu(\infty) = 0$, the map $\kappa_{\rm s} \mapsto \tau_{\rm s} = m_\nu^{-1}(1/\kappa_{\rm s})$ is a continuous increasing bijection $(1, \infty) \to (0, \infty)$; χ, ζ, P are continuous, and by (I.1) $1 - \Xi_{\rm s} > 0$ is bounded away from 0 on compacts, so no denominator vanishes and $\kappa_{\rm s} \mapsto \bar{\mathcal{E}}_s$ is continuous on $(1, \infty)$.

Limit $\kappa_{\rm s} \downarrow 1$ (the delicate one: $\tau_{\rm s} \to 0$ and $\Xi_{\rm s} \to 1$ together). $m_\nu(\tau_{\rm s}) = 1/\kappa_{\rm s} \to 1 = m_\nu(0)$ forces $\tau_{\rm s} \to 0^+$. By (I.2), self-noise $= \tau_{\rm s}\,[P\chi/\zeta] \to 0$ since $P(0) < \infty$, $\chi(0) = 1$, $\zeta(0) = \int \lambda^{-1}\, d\nu \in (0, \infty)$: the divergence $1/(1 - \Xi_{\rm s})$ is exactly cancelled by $B_0^{\rm s} \sim \tau_{\rm s}^2 P$, leaving a single power $\tau_{\rm s} \to 0$. For the bias and passed-noise terms, $F_i^{\rm s} \uparrow 1$ monotonically, so by monotone/dominated convergence

$$\sum_i \lambda_i (1 - F_i^{\rm s} F_i^{\rm w})^2 \omega_i^2 + \sum_i \lambda_i (F_i^{\rm s})^2 v_i^{\rm w} \to \sum_i \lambda_i (1 - F_i^{\rm w})^2 \omega_i^2 + \sum_i \lambda_i v_i^{\rm w} = \bar{\mathcal{E}}_w$$

(the last equality is $\bar{A}_\psi + B_{\rm w} + V_{\rm w}$).

Hence $\bar{\mathcal{E}}_s(\kappa_{\rm s}) \to \bar{\mathcal{E}}_w$, matching Part (i) across the boundary.

Limit $\kappa_{\rm s} \to \infty$ ($\tau_{\rm s} \to \infty$). Now $F_i^{\rm s} \to 0$ for each i; $\chi(\tau_{\rm s}) \to 0$ while $\kappa_{\rm s} = 1/m_\nu(\tau_{\rm s}) \sim \tau_{\rm s} / \int \lambda\, d\nu \to \infty$, so $\Xi_{\rm s} = \kappa_{\rm s}\chi(\tau_{\rm s}) \to 0$ and $1 - \Xi_{\rm s} \to 1$. Thus self-noise $\to 0$ and passed-noise $\to 0$, while bias $\to \sum_i \lambda_i \omega_i^2 = r^2$ by dominated convergence. Hence $\bar{\mathcal{E}}_s(\kappa_{\rm s}) \to r^2$, the null-predictor risk.

Why (ii-b)'s hypothesis cannot be dropped. Every step of the $\kappa_{\rm s} \to \infty$ argument runs through $\tau_{\rm s} \to \infty$, which the LSD supplies only when $\lambda_{\min} > 0$. Along a power law the proportional limit sends $n_{\rm s} = d/\bar{\kappa_{\rm s}} \to \infty$ as well, hence $\tau_{\rm s} \asymp n_{\rm s}^{-a} \to 0$ and $F^{\rm s} \to 1$ — the opposite regime — and the $\bar{\kappa_{\rm s}} \to \infty$ limit is then *not* r^2 at all. That is not a gap in the argument but a false conclusion, and Remark I.4 exhibits it.

Isotropic guide and the monotonicity contrast. For $\Sigma = I$ the exact projection identity (D.4) gives $\bar{\mathcal{E}}_s = \kappa_{\rm s}^{-1} \bar{\mathcal{E}}_w + (1 - \kappa_{\rm s}^{-1}) r^2$, *linear in* $\kappa_{\rm s}^{-1}$, hence *monotone* — no interior optimum. An isotropic spectrum admits gain only in the degenerate regime $\bar{\mathcal{E}}_w > r^2$, and even then monotonically: **anisotropy is necessary for a genuine sweet spot**, which motivates Part (iii).

I.4 Part (iii): a provable non-monotone instance in the spiked family

Spectrum and prior (two atoms). Fix a spike fraction $\pi \in (0, 1)$; let $k = \pi d$ spike coordinates have $\lambda = L > 1$ and carry all the signal ($\omega_i^2 = r^2/(kL)$ on spikes, so $\sum_{\rm spikes} \lambda_i \omega_i^2 = r^2$), and let the $d - k$ bulk coordinates have $\lambda = 1$, zero signal. Taking $k = \pi d$ keeps this inside the Theorem D.1 framework: the LSD is $\nu = \pi \delta_L + (1 - \pi)\delta_1$ and $d\,\omega_{\rm spike}^2 = r^2/(\pi L)$ stays finite (a genuine BBP-type two-atom model).

With $m > k$ the head H sees all k spikes plus $m - k$ bulk coordinates, so $A_m = 0$ and $s_{\rm w}^2 = \sigma^2$. Mechanism: the overparametrized teacher sprays fluctuation over its $m - k$ signal-free bulk directions; the student, with $\tau_{\rm s} > 0$, shrinks bulk by $F_{\rm bulk}^{\rm s} = 1/(1 + \tau_{\rm s}) < 1$ (denoising) while preserving spikes ($F_{\rm spike}^{\rm s} \approx 1$).

All fixed points are roots of quadratics. Clearing denominators in the MNI conditions $kL/(L + \tau) + (\#\text{bulk})/(1 + \tau) = n$:

$$\begin{aligned} &\text{teacher:}\ \ n_{\rm w}\tau_{\rm w}^2 + [n_{\rm w}(L + 1) - kL - (m - k)]\tau_{\rm w} - (m - n_{\rm w})L = 0, \\ &\text{student:}\ \ n_{\rm s}\tau_{\rm s}^2 + [n_{\rm s}(L + 1) - kL - (d - k)]\tau_{\rm s} - (d - n_{\rm s})L = 0, \end{aligned} \tag{I.5}$$

each with a unique positive root (the MNI map is strictly monotone).

Infinite-spike ($L \to \infty$) collapse. Dividing each quadratic by L and letting $L \to \infty$ gives the clean roots

$$\tau_{\mathrm{w}} \to \mu_{\mathrm{w}} - 1, \;\; \mu_{\mathrm{w}} := \tfrac{m-k}{n_{\mathrm{w}}-k}; \qquad \tau_{\mathrm{s}} \to \mu_{\mathrm{s}} - 1, \;\; \mu_{\mathrm{s}} := \tfrac{d-k}{n_{\mathrm{s}}-k},$$

valid for $n_{\mathrm{w}} > k$, $n_{\mathrm{s}} > k$ (an infinite spike is captured by k sample equations for free; the residual bulk is isotropic MNI with effective overparametrization μ, so $\tau = \mu - 1$).

Splitting every channel sum into spike ($L/(L+\tau) \to 1$, $L^2/(L+\tau)^2 \to 1$, $L/(L+\tau)^2 \to 0$) and bulk ($\lambda = 1$) blocks and passing to the limit termwise:

$$\bar{\mathcal{E}}_w = \Gamma_{\mathrm{w}}\big[k + \tfrac{n_{\mathrm{w}}-k}{\mu_{\mathrm{w}}}\big], \quad \Gamma_{\mathrm{w}} = \tfrac{\sigma^2 \mu_{\mathrm{w}}}{(\mu_{\mathrm{w}}-1)(n_{\mathrm{w}}-k)},$$
$$\bar{\mathcal{E}}_s = k\Gamma_{\mathrm{w}} + \tfrac{(n_{\mathrm{w}}-k)\Gamma_{\mathrm{w}}}{\mu_{\mathrm{w}}\mu_{\mathrm{s}}^2} + \Gamma_{\mathrm{s}}\big[k + \tfrac{n_{\mathrm{s}}-k}{\mu_{\mathrm{s}}}\big], \quad \Gamma_{\mathrm{s}} = \tfrac{(\mu_{\mathrm{s}}-1)(n_{\mathrm{w}}-k)\Gamma_{\mathrm{w}}}{\mu_{\mathrm{w}}\mu_{\mathrm{s}}(n_{\mathrm{s}}-k)}.$$

Step 1: subtract the two limiting risks. The spike teacher-variance $k\Gamma_{\mathrm{w}}$ passes through unchanged and cancels in the gain, leaving (bulk denoising) minus (student self-noise).

After substituting Γ_{s} and simplifying (elementary, independently verified with a symbolic solver),

$$\bar{\mathcal{E}}_w - \bar{\mathcal{E}}_s \;=\; \sigma^2 \, \frac{(d-n_{\mathrm{s}})(n_{\mathrm{w}}-k)(n_{\mathrm{s}}-2k)}{(d-k)(n_{\mathrm{s}}-k)(m-n_{\mathrm{w}})}, \tag{I.6}$$

which *reproduces* (D.5) *of the main text*. In the valid regime $k < n_{\mathrm{w}} < m$ and $k < n_{\mathrm{s}} < d$ every factor except $(n_{\mathrm{s}} - 2k)$ is strictly positive, so

$$\bar{\mathcal{E}}_w - \bar{\mathcal{E}}_s > 0 \iff n_{\mathrm{s}} > 2k, \qquad = 0 \text{ at } n_{\mathrm{s}} = 2k \text{ and } n_{\mathrm{s}} = d,$$

and the gain vanishes at both ends.

The formula is scale-invariant (degree 0), so it depends only on the ratios; every $n_{\mathrm{s}} \in (2k, d)$ yields strict gain.

Step 2: optimize the budget. Maximizing over n_{s} is elementary. Substituting $x = n_{\mathrm{s}} - k > 0$,

$$\frac{(d-n_{\mathrm{s}})(n_{\mathrm{s}}-2k)}{n_{\mathrm{s}}-k} \;=\; \frac{\big((d-k)-x\big)(x-k)}{x} \;=\; d - x - \frac{k(d-k)}{x},$$

which is strictly concave on $(0, \infty)$ (second derivative $-2k(d-k)/x^3 < 0$) with derivative $-1 + k(d-k)/x^2$, hence has the unique maximizer $x = \sqrt{k(d-k)}$, i.e.

$$n_{\mathrm{s}}^* \;=\; k + \sqrt{k(d-k)},$$

which is (D.6), with maximal gain $\sigma^2(n_{\mathrm{w}}-k)\big(\sqrt{d-k} - \sqrt{k}\big)^2/[(d-k)(m-n_{\mathrm{w}})]$.

Step 3: check that the optimizer is interior. Since $\sqrt{k(d-k)} > k \iff d > 2k \iff \sqrt{k(d-k)} < d-k$, the maximizer lies strictly inside $(2k, d)$ exactly when $d > 2k$ — which Theorem D.6 (iii) assumes — so n_{s}^* is interior there, and $n_{\mathrm{s}}^* \approx \sqrt{kd}$ when $k/d \to 0$.

Certified instance. Take $d = 4000$, $k = 10$, $m = 1200$, $n_{\mathrm{w}} = 400$, $\sigma^2 = 1/4$, $r^2 = 1$, and $\kappa_{\mathrm{s}} = 2$ (i.e. $n_{\mathrm{s}} = 2000 > 2k = 20$). Then (I.6) is a product of six manifestly positive integers times σ^2:

$$\bar{\mathcal{E}}_w - \bar{\mathcal{E}}_s = \tfrac{1}{4} \cdot \frac{2000 \cdot 390 \cdot 1980}{3990 \cdot 1990 \cdot 800} = +0.060783,$$

so $\bar{\mathcal{E}}_s(2) = 0.13141 - 0.06078 = 0.07063 < 0.13141 = \bar{\mathcal{E}}_w$; the only nontrivial check is the single inequality $2000 > 20$.

At these fixed ratios, continuity of the fixed points and of the risk gives

$$G_L \longrightarrow G_\infty > 0 \quad \text{as } L \to \infty. \tag{I.7}$$

Thus there exists a finite L_0 such that every $L \geq L_0$ has positive gain. This conclusion needs no sampled bound on a derivative.

Remark I.2 (Finite-spike numerical illustration). Direct formula evaluation at $L = 100$ gives approximately $G_{100} = 0.0605$ and $\bar{\mathcal{E}}_s = 0.0712 < 0.1316 = \bar{\mathcal{E}}_w$ for the displayed ratios. These rounded values are numerical illustrations. No uniform derivative bound or directed-rounding certificate for all $L \geq 50$ is asserted.

I.5 COROLLARY: SWEET SPOT $\iff$ GAIN, AND THE GAIN-FREE OVERSHOOT WITNESS

Combine Part (i) ($\bar{\mathcal{E}}_s(\kappa_s = 1) = \bar{\mathcal{E}}_w$) with the Part (ii) boundaries ($\bar{\mathcal{E}}_s(1^+) = \bar{\mathcal{E}}_w$, $\bar{\mathcal{E}}_s(\infty) = r^2$), in the meaningful regime $\bar{\mathcal{E}}_w < r^2$ (teacher beats the null). Everything below is read in a *fixed* reading of Part (ii) — (ii-a) at finite d, or (ii-b) proportional under bounded condition number — because the two boundary values are the entire content of the argument; at a power-law spectrum that means (ii-a), and Remark I.4 records what goes wrong otherwise.

(i) Gain ⇒ non-monotone, with an attained interior optimum (rigorous). Suppose $\bar{\mathcal{E}}_s(\kappa_s^\circ) < \bar{\mathcal{E}}_w$ for some $\kappa_s^\circ \in (1, \infty)$.

By Theorem D.6 (ii) the map $\kappa_s \mapsto \bar{\mathcal{E}}_s(\kappa_s)$ is continuous on $(1, \infty)$ with boundary limits $\bar{\mathcal{E}}_w$ (at $\kappa_s \downarrow 1$) and r^2 (at $\kappa_s \to \infty$), so it extends continuously to the compact $[1, \infty]$. That extension attains its minimum $\mu \le \bar{\mathcal{E}}_s(\kappa_s^\circ) < \bar{\mathcal{E}}_w < r^2$, and because μ is strictly below *both* endpoint values the minimizer set $S := \{\kappa_s \in [1, \infty] : \bar{\mathcal{E}}_s(\kappa_s) = \mu\}$ is a nonempty compact subset of the *open* interval $(1, \infty)$: the optimum is attained, and attained in the interior.

A monotone function stays between its endpoint limits, so $\bar{\mathcal{E}}_s$ is non-monotone.

(ii) No enlargement past the optimum ever recovers it. Let $\kappa_s^* := \min S$ — the smallest optimal κ_s, i.e. the *largest* optimal budget $n_s^* := d/\kappa_s^*$. For every $\kappa_s \in [1, \kappa_s^*)$, i.e. every budget $n_s \in (n_s^*, d]$, we have $\kappa_s \notin S$ and therefore $\bar{\mathcal{E}}_s(\kappa_s) > \mu = \bar{\mathcal{E}}_s(\kappa_s^*)$; at $n_s = d$ this reads $\bar{\mathcal{E}}_w > \mu$, the clone value of Theorem D.6 (i). In particular $n_s \mapsto \bar{\mathcal{E}}_s$ strictly increases somewhere on (n_s°, d).

What (ii) does not say. It yields neither uniqueness of the interior minimizer nor monotonicity of $\bar{\mathcal{E}}_s$ on (n_s^*, d); nothing in the argument excludes a second interior local minimum beyond n_s^*, and we do not claim there is none. Unimodality is available only where a closed form supplies it — as in the spiked family of Theorem D.6 (iii), where the n_s-dependent factor of (D.5) is $(D - t)(t - k)/t = (D + k) - t - Dk/t$ in the variables $t := n_s - k$, $D := d - k$, with derivative $Dk/t^2 - 1$, hence strictly increasing below $t = \sqrt{Dk}$ and strictly decreasing above it, so that the gain peaks exactly at (D.6) and every enlargement past it strictly hurts.

(iii) The exact equivalence. By continuity, [$\exists \kappa_s$: gain] $\iff$ [$\inf_{\kappa_s > 1} \bar{\mathcal{E}}_s < \bar{\mathcal{E}}_w$] $\iff$ [the interior optimum beats the clone] — strict gain is exactly the risk curve *dipping below its clone value*. This forces non-monotonicity via (i).

Remark I.3 (The literal converse "non-monotone ⇒ gain" is false: an explicit witness). The equivalence is with *dip-below-clone*, not bare non-monotonicity. Consider $d = 200$ with five atoms $\lambda \in \{120, 2, 1, 0.4, 0.2\}$, coordinate counts $\{12, 9, 62, 101, 16\}$, signal shares $\{0.084, 0.791, 0.046, 0.020, 0.060\}$ (most signal on the $\lambda = 2$ atom), head $m = 116$ ($\kappa_w = 2.9$), $n_w = 40$, $\sigma^2 = 0.034$.

Here $\bar{\mathcal{E}}_w = 0.6830 < r^2 = 1$, yet $\bar{\mathcal{E}}_s(\kappa_s)$ *rises* from $\bar{\mathcal{E}}_w$ to a peak $\bar{\mathcal{E}}_s = 1.43$ at $\kappa_s \approx 15$ (above the null r^2) before descending toward r^2 — strongly non-monotone, but $\bar{\mathcal{E}}_s > \bar{\mathcal{E}}_w$ for every $\kappa_s > 1$, so there is *no* W2S gain. That the absence of boundary gain is not a grid artifact follows from the analytic initial slope

$$\left.\frac{d\bar{\mathcal{E}}_s}{d\tau_s}\right|_{\tau_s = 0} = 2\sum_i (1 - F_i^w) F_i^w \omega_i^2 - 2\sum_i v_i^w + d\,\frac{\sum_i \rho_i/\lambda_i}{\sum_i 1/\lambda_i} = +0.0287 > 0, \tag{I.8}$$

so $\bar{\mathcal{E}}_s$ leaves the clone *upward*; an exhaustive scan over every integer $n_s \in [2, 199]$ confirms $\min \bar{\mathcal{E}}_s = \bar{\mathcal{E}}_w$ is attained only in the $\kappa_s \downarrow 1$ limit.

Such witnesses are rare (they need slope$_0 > 0$: a well-determined teacher with little filterable variance yet signal placed to force overshoot — about 0.1% of a randomized spectral screen), which is why unfiltered sweeps see overshoot only alongside a (often microscopic) gain. "Non-monotone" and "gain" are genuinely different phenomena.

Remark I.4 (Conditional status, and the order of limits). Part (i) is unconditional finite-d linear algebra. Parts (ii)–(iii) and Corollary D.8 are statements about the composed limit formula (D.2), and Part (ii) comes in the two readings of Section I.3: *(ii-a)* at finite d, where (D.2) is the channel calculus of Definition C.2 with n_s a real parameter in $(0, d)$ and both boundary limits hold for *every* spectrum with $\lambda_i > 0$; and *(ii-b)* in the proportional limit, where they hold when ν is compactly supported with $\lambda_{\min} > 0$ — which is also where Theorem D.1 identifies (D.2) with the exact proportional

limit of the random risk. The spiked family satisfies Assumption C.1, so the limiting calculation and finite-height existence argument (I.6)–(I.7) are covered by Theorem D.1.

At a power-law spectrum, read Parts (ii)–(iii) and Corollary D.8 in the finite-d reading (ii-a) (Section G.3). The proportional reading is not merely unproved there: its $\bar{\kappa_s} \to \infty$ clause is *false*. Taking $d \to \infty$ first at fixed $\bar{\kappa_s}$ sends $n_s = d/\bar{\kappa_s} \to \infty$, hence $\tau_s \asymp n_s^{-a} \to 0$ (Lemma K.9 (a)) and $F^s \to 1$, and the limit is the noise floor $\sigma^2 C_{\mathrm{floor}}(a, \psi, \bar{\kappa_w}, \bar{\kappa_s})$ of Theorem D.14 (ii) — which contains neither r^2 nor b (Remark D.16 (b), Remark K.13). The two statements are in flat contradiction, and it is Part (ii) that gives way.

Witness, at exactly the power-law example ($a = 1.5$, $b = 1$, $\psi = 0.3$, $\bar{\kappa_w} = 3$, $\sigma^2 = 1/4$): the proportional value of $\bar{\mathcal{E}}_s$ is 0.135595 at $r^2 = 1$ and 0.135606 at $r^2 = 4$ ($\bar{\kappa_s} = 100$, $n_s = 8000$, $d = 8 \cdot 10^5$), and 0.138485 against 0.138577 at $\bar{\kappa_s} = 10^3$ ($n_s = 2000$, $d = 2 \cdot 10^6$) — four-digit agreement across a fourfold change in r^2, where Part (ii) read proportionally would demand the values 1 and 4.

Read at finite d the same instance behaves exactly as (ii-a) says: at $d = 8 \cdot 10^5$ and $r^2 = 4$, $\bar{\mathcal{E}}_s$ runs $0.4211,\ 2.6270,\ 3.8318,\ 3.9828,\ 3.9983,\ 3.9998$ at $n_s = 10,\ 1,\ 10^{-1},\ 10^{-2},\ 10^{-3},\ 10^{-4}$, converging to $r^2 = 4$; over the last four entries the deficit $r^2 - \bar{\mathcal{E}}_s$ falls by a clean factor of ten per decade ($1.682 \cdot 10^{-1}$, $1.715 \cdot 10^{-2}$, $1.718 \cdot 10^{-3}$, $1.719 \cdot 10^{-4}$), i.e. the approach is linear in n_s. This is the $\lambda_{\min} \to 0$ frontier of Remark G.6 at the one place in this paper where it changes a conclusion rather than a constant.

I.6 The cost of holding the filter fixed

The interior optimum in Theorem 4.2 concerns a fixed minimum-norm rule. The following identity separates a change in its mean filter from a change in the statistical cost of realizing that filter.

Proposition I.5 (More samples at the same effective filter)**.** *Fix a positive finite spectrum and teacher profiles (ρ, c). For $t > 0$, put*

$$F_i(t) = \frac{\lambda_i}{\lambda_i + t}, \qquad D_j(t) = \sum_i F_i(t)^j \quad (j = 1, 2), \qquad B(t) = \sum_i \lambda_i \rho_i (1 - F_i(t))^2.$$

A sample budget n realizes this mean filter with nonnegative ridge exactly when $n \geq D_1(t)$, at ridge $\eta_n(t) = t[1 - D_1(t)/n]$. Its deterministic risk is

$$\mathcal{R}_n(t) = R(F(t)) + \frac{B(t) D_2(t)}{n - D_2(t)}. \tag{I.9}$$

For $n_2 > n_1 \geq D_1(t)$,

$$\mathcal{R}_{n_1}(t) - \mathcal{R}_{n_2}(t) = \frac{B(t) D_2(t)(n_2 - n_1)}{(n_1 - D_2(t))(n_2 - D_2(t))} \geq 0. \tag{I.10}$$

Consequently the optimally ridge-tuned risk in this calculus is nonincreasing in the sample budget. The inequality is strict at a fixed $t > 0$ whenever the teacher has nonzero second-moment energy.

Proof. The ridge identity is the fixed-point equation solved for η. Since $0 < F_i(t) < 1$, one has $D_2(t) < D_1(t) \leq n$, so the denominator is positive. Also $\Xi = D_2(t)/n$ and $\Gamma = B(t)/[n - D_2(t)]$. Summing the student fluctuation variances against λ_i therefore gives $B(t)D_2(t)/[n - D_2(t)]$. Subtracting this expression at the two budgets proves (I.10). The feasible set of effective thresholds expands with n, and the risk at each common threshold decreases. Taking infima proves monotonicity, with the zero and identity filters included by their continuous limits when feasible. □

This is an exact identity of the deterministic calculus. It is not a finite-sample ordering of two randomly fitted ridge estimators. It explains why the minimum-norm budget optimum does not recommend discarding available samples when the regularization parameter can be retuned: a larger batch can preserve the old mean filter while lowering its fluctuation cost.

J PROOFS FOR CHAINS (THEOREMS D.9 AND D.10)

A model β in the Σ-eigenbasis is summarized by its surrogate state $\beta_i \approx \alpha_i w_i^* + \sqrt{\nu_i}\, g_i$ ($g \perp w^*$): a mean coefficient α_i and a fluctuation variance ν_i, with excess risk $\bar{\mathcal{E}}(\beta) = \sum_i \lambda_i[(1-\alpha_i)^2\omega_i^2 + \nu_i]$ and per-coordinate second moment $\rho_i = \alpha_i^2\omega_i^2 + \nu_i$. The chain composition, the strictness criterion, the global-necessity trichotomy with its certified counterexample (Section J.4), and the finite-budget certificates are proved in turn; Section J.7 records the qualification analogy at the level clearly marked heuristic.

J.1 AGGREGATE RECURSION AND ITS STABILITY IN DEPTH

Theorem D.9 (Chain recursion)**.** *Under Assumption C.1, also permitting $\sigma^2 = 0$, with per-hop ratios fixed ($d/n_t \to \bar{\kappa}_t > 1$, ridges $\lambda_t \geq 0$ fixed, k fixed), define the chain state $(\alpha^{(t)}, \nu^{(t)})$ by the teacher initialization $\alpha_i^{(0)} = F_i^{\mathrm{w}}\mathbf{1}\{i \leq m\}$, $\nu_i^{(0)} = v_i^{\mathrm{w}}\mathbf{1}\{i \leq m\}$ and the recursion*

$$\alpha_i \;\leftarrow\; F_i^{(t)}\alpha_i, \qquad \nu_i \;\leftarrow\; \big(F_i^{(t)}\big)^2\nu_i + v_i^{(t)}, \tag{D.7}$$

where hop t's channel $(\tau_t, \Xi_t, F^{(t)})$ is that of (Σ, n_t, λ_t) and $v^{(t)}$ is computed at the realized *profile $\rho_i^{(t-1)} = (\alpha_i^{(t-1)})^2 p_i + \nu_i^{(t-1)}$, $p_i := (w_i^*)^2$, with zero label noise (the quenched convention of Theorem D.1; replacing p_i by ω_i^2 throughout gives the annealed recursion, and the two agree under the moment condition of Remark G.4).*

Then the recursion computes the limits of all risk-relevant aggregate functionals *of the chain. For a deterministic signal sequence, assume bounded Euclidean norms. For random signals, apply the conditional bounds on signal- and teacher-norm balls, then average and remove localization as in (G.7). Accordingly "deterministic" below permits w^*-*measurable *vectors. The excess risk follows by taking $(q, u) = (\lambda, \lambda \odot w^*)$ (Section J.1). For every $D, U > 0$ and* each *such pair with $\|q\|_\infty \leq D$, $\|u\| \leq U$, the failure probabilities are uniform over these two bounded classes after norm localization. There is no assertion of a single event serving all pairs simultaneously:*

$$\sum_i q_i\big(\beta_i^{(t)}\big)^2 \xrightarrow{\mathbb{P}} \sum_i q_i\,\rho_i^{(t)}, \qquad u^\top\beta^{(t)} \xrightarrow{\mathbb{P}} \sum_i u_i\,\alpha_i^{(t)}w_i^*, \qquad t = 0, 1, \ldots, k \tag{D.8}$$

In particular the hop-t excess risk converges,

$$\mathcal{E}^{(t)} \xrightarrow{\mathbb{P}} \bar{\mathcal{E}}^{(t)} = \sum_i \lambda_i\big[(1-\alpha_i^{(t)})^2 p_i + \nu_i^{(t)}\big],$$

and hop $t{+}1$'s realized channel inputs (the Γ-quadratic) converge to the recursion's $(\rho^{(t)}, \Gamma^{(t+1)})$.

The recursion tracks aggregate moments, not independent coordinatewise Gaussian errors at every stage. Its proof uses Lemma G.3 at each fresh design. Combining the two quadratic contributions into one weight vector reveals a useful contraction and avoids a branching induction over separately propagated terms.

Lemma J.1 (One hop and its backward weight operator)**.** *Under the bounds of Lemma G.3, let β be independent of the new design, and let $\beta' = S_\eta\beta$ fit its noiseless labels. For $q \in \mathbb{R}^d$, define*

$$(\mathcal{T}q)_j = q_j F_j^2 + \frac{c_j(1-F_j)^2}{n(1-\Xi)}\sum_i \frac{q_i c_i}{(c_i+\tau)^2}. \tag{J.1}$$

On each fixed ball for β, conditionally on β and on any bounded weights and vectors measurable before drawing the new design,

$$\sum_i q_i(\beta_i')^2 - \sum_j (\mathcal{T}q)_j\beta_j^2 \xrightarrow{\mathbb{P}} 0, \tag{J.2}$$
$$z^\top\beta' - (F \odot z)^\top\beta \xrightarrow{\mathbb{P}} 0.$$

The probability bounds are uniform over these bounded choices. Moreover $\|\beta'\| \le \|\beta\|$ pointwise, $\mathcal{T}$ is a positive linear operator, and

$$0 \le (\mathcal{T}\mathbf{1})_j = F_j - \frac{\eta c_j}{(c_j+\tau)^2(1-\Xi)} \le F_j, \qquad \|\mathcal{T}q\|_\infty \le \max_j F_j \, \|q\|_\infty. \tag{J.3}$$

Proof. The eigenvalues of S_η lie in $[0,1]$, proving the pathwise contraction. Apply Lemma G.3 with target β and zero noise. Its quadratic prediction is

$$\sum_i q_i F_i^2 \beta_i^2 + \frac{\sum_j c_j (1-F_j)^2 \beta_j^2}{n(1-\Xi)} \sum_i \frac{q_i c_i}{(c_i+\tau)^2},$$

which is (J.1). Its coefficients are nonnegative when $q \ge 0$. Substituting $q = \mathbf{1}$ and $\tau B_C = n(1-\Xi) - n\eta/\tau$ from (G.1) gives the equality in (J.3). Positivity then implies $|\mathcal{T}q| \le \mathcal{T}|q| \le \|q\|_\infty \mathcal{T}\mathbf{1}$. □

Proof of Theorem D.9. **1. Initialize at the teacher.** Fix the signal, with bounded norm. The teacher reduction and Lemma G.3 give the stage-zero linear and quadratic laws, with profiles $\alpha^{(0)} = F^{\mathrm{w}}$ and $\rho^{(0)} = (F^{\mathrm{w}})^2 \odot p + v^{\mathrm{w}}$. The teacher norm is tight by (G.7).

2. Propagate one aggregate backward. For a fixed terminal weight vector q, apply Lemma J.1 at the last stage. It changes q to the single deterministic vector $\mathcal{T}_t q$ at the preceding stage. Its norm does not increase, by (J.3). Repeating this step reaches the teacher in t applications. For a linear functional, the corresponding operation is $z \mapsto F^{(t)} \odot z$, which also does not increase its norm. The schedule is deterministic; these backward weights are therefore independent of the designs at which they are tested. All realized targets stay in the same teacher norm ball, since each fit is a contraction. At fixed depth a finite union bound, followed by removal of that ball, makes the sum of approximation errors vanish.

3. Identify the forward recursion. The adjoint identity for (J.1) is

$$\sum_j (\mathcal{T}_t q)_j \rho_j^{(t-1)} = \sum_i q_i \left[(F_i^{(t)})^2 \rho_i^{(t-1)} + v_i^{(t)} \right],$$

where $v^{(t)}$ is evaluated at $\rho^{(t-1)}$ with zero label noise. The linear identity gives $\alpha^{(t)} = F^{(t)} \odot \alpha^{(t-1)}$. Writing $\rho^{(t)} = (\alpha^{(t)})^2 \odot p + \nu^{(t)}$ proves exactly (D.7). Taking $q_i = \lambda_i$ and $z_i = \lambda_i w_i^*$ then yields the risk formula by expanding its square. For a random signal, localization and the prior concentration condition (G.9) give the stated readings. □

Theorem J.2 (A risk approximation uniform over all depths)**.** *Suppose the teacher obeys the hypotheses of Theorem 3.2. Consider any deterministic infinite schedule of independent downstream batches, with constants $\delta_0, K, \eta_0 > 0$ such that, for every d, t,*

$$1 + \delta_0 \le d/n_t \le K, \qquad 0 \le \eta_t \le \eta_0.$$

Let $\mathcal{R}_t^{(d)}$ be the prior-averaged finite-dimensional chain formula and $\beta^{(t)}$ the actual stage-t estimator, including the teacher at $t = 0$. Then

$$\sup_{t \ge 0} \left| \mathcal{E}(\beta^{(t)}) - \mathcal{R}_t^{(d)} \right| \xrightarrow{\mathbb{P}} 0. \tag{J.4}$$

The convergence also holds in L^1. In particular it covers any dimension-dependent depth, with no upper bound on its growth. For deterministic signals of bounded norm, the analogous probability statement holds for the formulas evaluated at the realized profile.

Proof. **1. The deterministic profile loses energy geometrically.** By Lemma G.1, $F_i^{(t)} \le r := M^2/(M^2+\delta_0) < 1$ for all stages. The adjoint relation and (J.3) give

$$\sum_i \rho_i^{(t)} = \sum_j (\mathcal{T}_t \mathbf{1})_j \rho_j^{(t-1)} \le r \sum_i \rho_i^{(t-1)} \le r^t \sum_i \rho_i^{(0)}. \tag{J.5}$$

The last initial sum is bounded. Since $(\alpha_i^{(t)})^2\omega_i^2 \le \rho_i^{(t)}$, Cauchy–Schwarz in the formula gives

$$\sup_{t\ge T}|\mathcal{R}_t^{(d)} - r_d^2| \le M\left(Cr^T + 2\sqrt{Cr^T\sum_i \omega_i^2}\right), \qquad r_d^2 = \sum_i \lambda_i\omega_i^2.$$

Thus the entire deterministic tail approaches the zero-prediction risk uniformly in dimension.

2. A finite prefix controls the entire random tail. For each fixed T, the proved aggregate recursion, applied to $q = \mathbf{1}$, gives $\|\beta^{(T)}\|^2 - \sum_i \rho_i^{(T)} \to 0$ in probability. Pathwise contraction then bounds every later norm by $\|\beta^{(T)}\|$. For every $t \ge T$,

$$\left|\mathcal{E}(\beta^{(t)}) - \|w^*\|_\Sigma^2\right| \le M\|\beta^{(T)}\|^2 + 2M\|w^*\|\,\|\beta^{(T)}\|.$$

The signal norm is tight, and $\|w^*\|_\Sigma^2 - r_d^2 \to 0$ by the main theorem's fourth-moment condition. Given an error tolerance, first choose T so that the bounds containing r^T are sufficiently small. Then take $d \to \infty$. The finite union for $0 \le t \le T$ and the bound for the whole remaining tail prove (J.4). This argument does not sum one-step errors over a growing number of stages.

3. Pass to expected errors. All random risks are bounded by $2M(\|\beta_{\mathrm{w}}\|^2 + \|w^*\|^2)$, independently of depth. The deterministic risks are uniformly bounded by Step 1 and the bounded initial profile. Lemma G.5 therefore bounds the second moment of the supremum in (J.4). Uniform integrability proves its L^1 convergence. For deterministic signals, use $\|w^*\|_\Sigma^2$ in place of r_d^2 throughout. □

J.2 An exact finite-dimensional forgetting bound

Proposition J.3 (Repeated fresh projections lose the predictor)**.** *Let $\Sigma \succ 0$, $\kappa_\Sigma = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, and let each independent Gaussian batch have $n_t \le d$ rows. Fit noiseless pseudo-labels by minimum norm or a nonnegative ridge. For any initial predictor independent of these batches with finite second moment,*

$$\mathbb{E}\|\beta^{(T)}\|^2 \le \mathbb{E}\|\beta^{(0)}\|^2 \prod_{t=1}^{T}\left[1 - \frac{1 - n_t/d}{\kappa_\Sigma}\right]. \tag{J.6}$$

For isotropic minimum-norm stages this is an equality. No high-dimensional approximation is used.

Proof. **1. Compare the unobserved component to a Haar subspace.** Write $A = \Sigma^{1/2}$ and let V be the row space of a standard Gaussian $n \times d$ matrix. The row space of the actual design is AV. For deterministic u,

$$\operatorname{dist}(u, AV)^2 = \inf_{v\in V}\|A(A^{-1}u - v)\|^2 \ge \lambda_{\min}(\Sigma)\operatorname{dist}(A^{-1}u, V)^2.$$

Since V is Haar distributed,

$$\mathbb{E}\operatorname{dist}(u, AV)^2 \ge \lambda_{\min}(\Sigma)(1 - n/d)\|A^{-1}u\|^2 \ge \frac{1 - n/d}{\kappa_\Sigma}\|u\|^2.$$

2. Iterate the conditional contraction. For minimum norm, the fitted norm squared is $\|u\|^2 - \operatorname{dist}(u, AV)^2$. The squared ridge fitting operator is bounded above by the projection onto AV, so its norm is no larger. Condition on each preceding predictor and iterate expectations. Under isotropy, the displayed distance comparison is an equality. □

For example, if $n_t/d \le 1 - \epsilon$ and $\kappa_\Sigma \le \kappa_0$, the predictor energy is at most $(1 - \epsilon/\kappa_0)^T\mathbb{E}\|\beta^{(0)}\|^2$. With uniformly bounded signal and teacher second moments and $\|\Sigma\| \le M$, its expected risk consequently approaches zero prediction at a geometric rate, by Cauchy–Schwarz. Thus some early stages can improve prediction while unlimited compression ultimately removes the signal as well. This concerns fresh designs: reusing one minimum-norm design gives $P^2 = P$, so later fits leave the first prediction unchanged.

Remark J.4 (Scope of the chain statements). The finite-depth recursion and its uniform-depth extension are estimator results under bounded spectra. Their coordinate profiles are bookkeeping for aggregate moments, not a claim of a joint Gaussian coupling of all iterates. The extension requires a uniform distance from the interpolation threshold and a deterministic schedule. It does not cover adaptive choices based on earlier designs, nor give power-law rate approximations. At a power-law spectrum, the comparisons below remain statements about the deterministic calculus. The forgetting bound (J.6) is exact, but its condition-number dependence can make it uninformative when the spectrum deteriorates with dimension.

J.3 The population filter objective and the strictness criterion

Idealize the student hops as deterministic filters applied to the teacher output; the connection to a finite budget — the toll — is Section J.6. This is exactly the filter objective (C.8), so Proposition C.4 applies verbatim and the whole section takes place in one Hilbert space:

$$R(F) \;=\; \|F-g\|_\pi^2 + \mathrm{Irr}, \qquad g_i = \frac{c_i}{\rho_i}, \quad \pi_i = \lambda_i \rho_i, \tag{J.7}$$

with $c_i = F_i^{\mathrm{w}}\omega_i^2$ and $\rho_i = (F_i^{\mathrm{w}})^2\omega_i^2 + v_i^{\mathrm{w}}$ as in Definition C.3. Comparing the single-step family $\mathcal{R}_1$ with the two-hop family $\mathcal{R}_2$ (Definition C.5) is therefore a question about *approximating one fixed vector g from two nested sets*: write $R_1(t) := R(\varphi_t)$ and $R_2(t_1,t_2) := R(\varphi_{t_1}\varphi_{t_2})$.

Lemma J.5 (Inclusion). *$\varphi_0 = \mathbf{1}$, so $\varphi_t\varphi_0 = \varphi_t$ and $\mathcal{R}_1 \subset \mathcal{R}_2$; hence $\min_{\mathcal{R}_2} R \le \min_{\mathcal{R}_1} R$ — at the population level a second hop never hurts.*

The content is therefore entirely in *when the inclusion is strict*, which by Proposition C.6 (ii)–(iii) is a statement about one inner product: whether the residual $\varphi_{t^*} - g$ at the single-step optimum is π-orthogonal to the direction in which the second hop moves the filter.

Theorem D.10 (Chains vs. the best single step). *Fix a teacher and a total unlabeled budget N; compare the best single ridge student ($R_1^* = \min_\lambda \bar{\mathcal{E}}_s$) with the best two-hop chain (R_2^*, optimized over the split and both ridges); write $R_1^{\mathrm{pop},*}, R_2^{\mathrm{pop},*}$ for their population ($N\to\infty$) counterparts. Assume the single-step population risk R_1 attains its infimum over $[0,\infty]$ at an interior t^*, so that $R_1(t^*) = R_1^{\mathrm{pop},*}$; put $F_i^* = \lambda_i/(\lambda_i + t^*)$, and define the* filter residuals

$$\Delta_i := F_i^{\mathrm{w}}(1 - F_i^* F_i^{\mathrm{w}})\,\omega_i^2 - F_i^* v_i^{\mathrm{w}}.$$

(i) Population level. *$R_2^{\mathrm{pop},*} \le R_1^{\mathrm{pop},*}$ always, and $R_k^*(N) \ge R_k^{\mathrm{pop},*}$ at every budget; the excess over the population value is at most an explicit* splitting cost, *which vanishes as N grows whenever the per-hop effective dimension $\mathrm{ed}(\tau) := \sum_j \lambda_j^2/(\lambda_j+\tau)^2$ is $o(N)$ at the relevant ridges — true at fixed d, and along the power law (C.2) at fixed $d/N > 1$ exactly when those ridges stand far above the interpolation scale, $\tau N^a \to \infty$;* false *for a spectrum bounded above and below in the proportional regime, where a noiseless minimum-norm hop at $\Sigma = I$, $N/d \to \alpha$ pays the N-independent toll $\alpha(1-\alpha)\|u\|^2$ (Proposition J.20, Remark J.22).*

Vanishing is not the same as negligible: along the power law at the tuned *ridges the splitting cost is $\Theta(N^{-1+1/(a+b)})$ for $b < a+1$, $\Theta(N^{-1+1/(2a+1)})$ for $b > a+1$ (past the Tikhonov saturation point), and $\Theta\big(N^{-1+1/(2a+1)}(\log N)^{1/(2a+1)}\big)$ at the saturation boundary $b = a+1$ — in no case $O(1/N)$ — and it stays a fixed fraction of $R_1^{\mathrm{pop},*}$ at every d, so the population gap is approached in absolute but not in relative terms (Proposition J.25, Remark J.21).*

The boundary factor is literally the same $(\log N)^{1/(2a+1)}$ that Theorem D.14 (iii) carries, and for the same reason: $b = a+1$ makes $a+b = 2a+1$, the logarithmic case of the underlying power-law sums. If the strictness criterion

$$\sum_i F_i^*(1-F_i^*)\,\Delta_i \;<\; 0 \qquad\Longleftrightarrow\qquad \mathrm{Cov}_\pi\big(\Delta,\; 1/\lambda\big) \;<\; 0, \quad \pi_i \propto (F_i^*)^2, \tag{D.10}$$

holds at that t^*, then the two-hop family wins strictly, *$R_2^{\mathrm{pop},*} < R_1^{\mathrm{pop},*}$.*

Globality of t^* is not removable: an instance meeting every standing hypothesis has a strictly negative criterion at a local *interior optimum sitting 42% above $R_1^{\mathrm{pop},*}$, so the argument controls the wrong quantity (Remark J.7).*

(ii) Finite budgets: a chain-wins certificate. *At the fully specified instance $d = 4000$, $m = 1200$, $n_{\mathrm{w}} = 400$, $N = 2000$, $\sigma^2 = 0.25$, $r^2 = 1$, $(a, b) = (1.2, 2.0)$ — i.e. $(\psi, \bar{\kappa_{\mathrm{w}}}, N/d) = (0.3, 3, 0.5)$ — the* annealed *recursion values satisfy $R_2^* \le 0.016302 < 0.017217 \le R_1^*$, a rigorous margin. (This is a statement about the deterministic recursion of Theorem D.9 at that instance — in its annealed reading, ω_i^2 in place of the realized p_i, which is the reading the certificate computes and the one the population analysis uses; the two readings are distinct along the power law, Remark G.4 (b) and Remark J.19 — certified by interval arithmetic, not about a finite-sample random risk.)*

Let t^* be the single-step optimum in the sense of the theorem: a *global* minimizer of R_1 over $[0, \infty]$ that lies in the interior, so $t^* \in (0, \infty)$, $R_1'(t^*) = 0$ and $R_1(t^*) = R_1^{\mathrm{pop},*}$ (numerically $t^* \in (0.002, 0.07)$ in the power-law family below). Write $F_i^* := \lambda_i/(\lambda_i + t^*)$ and $\Delta_i := c_i - F_i^* \rho_i = F_i^{\mathrm{w}}(1 - F_i^* F_i^{\mathrm{w}})\omega_i^2 - F_i^* v_i^{\mathrm{w}}$ (positive where the single ridge under-filters coordinate i, negative where it over-filters).

Remark J.6 ((D.9) is ridgeless, and one-directional). The proof of Remark C.7 (b) establishes (D.9) for an MNI teacher and, together with $\sum_i \lambda_i^2 c_i > 0$ — the teacher passes some signal — places the minimizer of $t \mapsto R_1(t)$ in the interior of $[0, \infty]$. Both hypotheses are needed, neither is implied by an interior optimum, and the identity itself is ridgeless. Throughout, $A := \sum_{i \le m} \lambda_i \omega_i^2/(\lambda_i + \tau_{\mathrm{w}})^2$ and $B := \sum_{i \le m} \lambda_i/(\lambda_i + \tau_{\mathrm{w}})^2$, as in that proof.

(a) A positive ridge changes the identity when the fitted target has nonzero energy. For $\lambda_{\mathrm{w}} > 0$ the ridge branch of (F.3) replaces $\Gamma_{\mathrm{w}} B = (\tau_{\mathrm{w}}^2 A + s_{\mathrm{w}}^2)/\tau_{\mathrm{w}}$ by $\Gamma_{\mathrm{w}} B = B(\tau_{\mathrm{w}}^2 A + s_{\mathrm{w}}^2)/(\tau_{\mathrm{w}} B + n_{\mathrm{w}} \lambda_{\mathrm{w}}/\tau_{\mathrm{w}})$, so (F.2) gives

$$R_1'(0) + 2\,\frac{s_{\mathrm{w}}^2}{\tau_{\mathrm{w}}} \;=\; \frac{2\lambda_{\mathrm{w}} n_{\mathrm{w}} \left(\tau_{\mathrm{w}}^2 A + s_{\mathrm{w}}^2\right)}{\tau_{\mathrm{w}} \left(\tau_{\mathrm{w}}^2 B + \lambda_{\mathrm{w}} n_{\mathrm{w}}\right)} \;>\; 0,$$

strictly positive when $s_{\mathrm{w}}^2 > 0$ (in particular under Assumption C.1), since $A, B \ge 0$.

So a ridge teacher can have $R_1'(0) > 0$, making $t = 0$ a local minimum of R_1 and the standing hypothesis of Theorem D.11 false: at the power-law instance $a = 1.5$, $b = 1$, $d = 600$, $\psi = 0.3$, $\bar{\kappa_{\mathrm{w}}} = 3$, $\sigma^2 = 0.25$ one measures $R_1'(0) = +0.778$ at $\lambda_{\mathrm{w}} = 0.4$ and $+0.591$ at $\lambda_{\mathrm{w}} = 1.5$, against $-2s_{\mathrm{w}}^2/\tau_{\mathrm{w}} = -1.179$ and -0.326.

(b) $s_{\mathrm{w}}^2 > 0$ does not suffice. Put all the signal in the tail, $\omega_i^2 = 0$ for $i \le m$. Then $c \equiv 0$ — the head passes no signal and the tail is not passed at all — so $c_i - \varphi_{t,i}\rho_i = -\varphi_{t,i} v_i^{\mathrm{w}} < 0$ at every head coordinate, where $v_i^{\mathrm{w}} > 0$ because $\Gamma_{\mathrm{w}} \ge s_{\mathrm{w}}^2/(n_{\mathrm{w}}(1 - \Xi_{\mathrm{w}})) > 0$.

Hence $R_1' < 0$ on all of $[0, \infty)$: R_1 is strictly decreasing, with no stationary point and its infimum only at $t = \infty$, although $s_{\mathrm{w}}^2 = \sigma^2 + A_m > 0$ and every clause of Assumption C.1 holds. (Instance: $d = 400$ with λ_i spread over $[0.25, 4]$, $m = 200$, $n_{\mathrm{w}} = 100$, $\sigma^2 = 0.3$ and all signal on $i > m$, so $s_{\mathrm{w}}^2 = 1.3$; $R_1'(t) < 0$ at $t = 10^{-3}, 1, 10, 10^3, 10^6$.)

(c) $s_{\mathrm{w}}^2 > 0$ is not necessary: bias-fed fluctuation already forces an interior optimum. Let the head carry $m/2$ coordinates at $\lambda = 1/\eta$ and $m/2$ at $\lambda = \eta$, $0 < \eta < 1$; take $n_{\mathrm{w}} = m/2$, place all head signal on the small-λ block with $\sum_i \omega_i^2 = w$ there, leave the tail signal-free, and set $\sigma^2 = 0$, so $s_{\mathrm{w}}^2 = 0$.

The MNI fixed point is $\tau_{\mathrm{w}} = 1$ exactly and $\Gamma_{\mathrm{w}} = w/m > 0$: the teacher is noiseless but biased, and (C.6) converts that bias into emitted fluctuation. Then $R_1(0) = w/2$, $R_1(\infty) = \eta w$, $R_1'(0) = 0$ and $R_1''(0) = w(3\eta - 1)/(\eta + 1) < 0$ for $\eta < 1/3$, so $t = 0$ is a strict local *maximum*. At $\eta = 1/9$, $w = 1$ the interior minimum sits at $t^* = 26545.3245$ with $R_1(t^*) = 0.111111064619$ — below $R_1(\infty) = 1/9$ by 4.649×10^{-8} and below $R_1(0) = 1/2$ by 78% (every value independent of m; computed at 60 digits).

So "chains filter noise, not bias" is false at the population level: what a chain filters is the emitted fluctuation v^{w}, and (C.6) feeds that from the stage's own bias as well as from s_{w}^2. Remark J.9 states the surviving form of the moral — it is $v^{\mathrm{w}} \equiv 0$, not $s_{\mathrm{w}}^2 = 0$, that removes the gain.

Proof of part (i): the strictness criterion. **Step 1: identify the residual and the two tangent directions.** Everything is one π-inner product. The residual at the single optimum is $\varphi_{t^*} - g$, and $\pi_i(\varphi_{t^*} - g)_i = \lambda_i\rho_i(F_i^* - c_i/\rho_i) = -\lambda_i\Delta_i$, so Δ is the residual in disguise: $\Delta_i > 0$ *means coordinate i is under-filtered by the single ridge*. The two directions in which the two-hop family moves the filter at $(t^*, 0)$ are $\partial F_i/\partial t_2 = -F_i^*/\lambda_i$ and $\partial F_i/\partial t_1 = -(F_i^*)^2/\lambda_i$.

By Proposition C.6 (ii) the single-step optimality of t^* says exactly that the residual is orthogonal to the *second* of these,

$$\sum_i (F_i^*)^2\Delta_i = 0,$$

and the question is whether it is also orthogonal to the first. Concretely, since $\partial R/\partial F_i = 2\pi_i(F_i - g_i) = -2\lambda_i\Delta_i$ at F^*,

$$\frac{\partial R_2}{\partial t_2}\Big|_{(t^*,0)} = 2\sum_i F_i^*\Delta_i = 2\sum_i F_i^*(1 - F_i^*)\Delta_i = 2t^*\sum_i \frac{(F_i^*)^2\Delta_i}{\lambda_i}, \tag{J.8}$$

where the second and third forms use the *single-optimality identity*

$$\frac{\partial R_2}{\partial t_1}\Big|_{(t^*,0)} = 2\sum_i (F_i^*)^2\Delta_i = R_1'(t^*) = 0 \tag{J.9}$$

(subtract (J.9) from $2\sum F_i^*\Delta_i$, then use $F_i^*(1 - F_i^*) = (F_i^*)^2 t^*/\lambda_i$).

Step 2: use stationarity to obtain the covariance criterion. Because (J.9) makes Δ mean-zero under $\pi_i \propto (F_i^*)^2$,

$$\sum_i (F_i^*)^2\Delta_i/\lambda_i = \Big(\textstyle\sum_i (F_i^*)^2\Big)\,\mathrm{Cov}_\pi(\Delta,\, 1/\lambda), \tag{J.10}$$

reproducing the covariance form (D.10). Since $t^* > 0$ and $\sum(F_i^*)^2 > 0$, the sign of $\partial R_2/\partial t_2|_{(t^*,0)}$ equals that of $\mathrm{Cov}_\pi(\Delta, 1/\lambda)$.

Step 3: improve on the global single-step optimum. Suppose this is < 0. Then $R_2(t^*, \varepsilon) < R_2(t^*, 0)$ for all small $\varepsilon > 0$, and $R_2(t^*, 0) = R(\varphi_{t^*}\varphi_0) = R(\varphi_{t^*}) = R_1(t^*)$ by Lemma J.5. This alone is only a strict improvement *on the value at* t^*. The theorem's conclusion compares the two *optima*, and it is exactly here that the hypothesis is used in the form stated: because t^* is a global minimizer of R_1 over $[0, \infty]$, $R_1(t^*) = R_1^{\mathrm{pop},*}$, and therefore

$$R_2^{\mathrm{pop},*} \;\le\; R_2(t^*, \varepsilon) \;<\; R_1(t^*) \;=\; R_1^{\mathrm{pop},*},$$

a strict win for the two-hop family.

Under a merely *local* interior optimum the middle inequality still holds but the last equality does not, and the argument controls the wrong quantity: Remark J.7 exhibits an instance satisfying every standing hypothesis at which the criterion is strictly negative at a local interior optimum whose value exceeds $R_1^{\mathrm{pop},*}$ by 42%, so nothing about $R_2^{\mathrm{pop},*}$ versus $R_1^{\mathrm{pop},*}$ follows. □

Remark J.7 (Globality of t^* is necessary, not cosmetic). Read with "t^* a local interior optimum" the strictness clause of Theorem D.10 (i) is *false*. A witness, at finite d, with a truncation sketch and an MNI teacher, and with every clause of Assumption C.1 satisfied (spectrum bounded above and below, $\kappa_{\mathrm{w}} = 37/24 = 1.5417 > 1$, $\sigma^2 > 0$):

$$\Sigma = \mathrm{diag}\left(1.116271^{\times 15},\ 0.009543^{\times 20},\ 0.005375^{\times 12}\right), \quad d = 47, \quad m = 37, \quad n_{\mathrm{w}} = 24,$$

$$\omega^2 = \left(1.103214{\cdot}10^{-3\,\times 15},\ 6.510757^{\times 20},\ 1.409593{\cdot}10^{-2\,\times 12}\right), \qquad \sigma^2 = 0.004497.$$

The teacher channel is $\tau_{\mathrm{w}} = 0.0127154883941$, $\Xi_{\mathrm{w}} = 0.7715357$, $\Gamma_{\mathrm{w}} = 0.0749315094447$.

Here R_1 has *three* interior stationary points. At the smallest, $t^\circ = 6.7406{\cdot}10^{-5}$, the point is a strict local minimum and the criterion (D.10) is $-6.064{\cdot}10^{-3} < 0$ — the hypothesis in its local reading holds — yet $R_1(t^\circ) = 1.793832$, while $R_1^{\mathrm{pop},*} = 1.261554$ is attained at the third stationary point $t^* = 53.0647$, where the criterion is $+9.950{\cdot}10^{-3} > 0$.

What follows with no further computation. At t° the perturbation argument of the proof above delivers exactly $R_2(t^\circ, \varepsilon) < R_1(t^\circ) = 1.793832$ and nothing more, whereas the conclusion of

Theorem D.10 (i) asserts $R_2^{\mathrm{pop},*} < 1.261554$. The quantity the proof controls sits 42.2% above the quantity the theorem talks about, so under the local reading the theorem has no proof, and the shortfall is a fixed fraction of the risk rather than a rounding error. That alone retires the local reading.

And the conclusion itself appears to fail — numerically. Searching the two-hop family at this instance finds nothing below the single-ridge slice: a 701×701 logarithmic sweep of (t_1, t_2) over $[10^{-10}, 10^7]^2$ together with the edges $t_j \in \{0, \infty\}$; 225 multistart local refinements in 30-digit arithmetic, all of which return $R_1^{\mathrm{pop},*}$; and a scan of the relaxed quadrant Q from 3477 starting points that locates *no* critical point of ψ in the open quadrant $\{e_1, e_2 > 0\}$ at all, so that by Lemma J.14 (K2) the minimum over Q^+ is attained on its boundary.

The apparent value is $R_2^{\mathrm{pop},*} = R_1^{\mathrm{pop},*} = 1.2615544627$, attained on the single-ridge slice $t_2 = 0$: a tie. We report that as a numerical finding and not as a certificate — unlike Theorem D.10 (ii) and Theorem D.12 it has not been enclosed in interval arithmetic, and a global two-dimensional enclosure is what such a claim would need (Remark J.27 discusses the same obstacle). The case against the local reading does not rest on it.

The instance is not a small-d artifact. Replicating every eigenvalue block κ times together with $m, n_{\mathrm{w}} \mapsto \kappa m, \kappa n_{\mathrm{w}}$ and $\omega_i^2 \mapsto \omega_i^2/\kappa$ — the scaling that holds ψ, κ_{w} and the limiting spectral and signal profiles of Assumption C.1 fixed while $d \to \infty$ — leaves τ_{w}, Ξ_{w}, $\kappa\Gamma_{\mathrm{w}}$, all three stationary points, all three values of R_1 and both criterion values *exactly* invariant; checked at $\kappa = 1, 8, 32$ ($d = 47, 376, 1504$) to 30 digits.

Consistency with what is proved: at the *global* optimum the criterion is positive here, so Theorem D.10 (i) as stated is silent at this instance, and Theorem D.11 is likewise unviolated — a positive criterion does not force a tie in general (Theorem D.12), it merely fails to force a win. What the witness kills is only the local reading. Every consumer of Theorem D.10 (i) in this paper — Theorem D.11 (whose preamble already assumes the infimum is *attained* at an interior t^*), its parts (ii)–(iv), and Theorem D.12 (which uses the criterion's *sign* only) — is on the global reading.

Local necessity. If $\partial R_2/\partial t_2|_{(t^*,0)} > 0$, then with $\partial R_2/\partial t_1|_{(t^*,0)} = 0$ the point $(t^*, 0)$ is a KKT point at which no infinitesimal second hop helps: first-order strictness along t_2. At *equality* the first-order test is silent and second-order terms decide — after re-optimizing t_1, the envelope's curvature is $\det H/(2A_2)$, which can be negative: this is exactly the fold of Remark J.18, where an infinitesimal second hop *does* help despite a vanishing criterion. Local necessity is therefore a strict-inequality statement; the boundary case belongs to the global analysis of Section J.4. □

Whether "$\mathrm{Cov}_\pi \geq 0$" also rules out *distant* interior configurations — the global converse — is resolved in Section J.4: unconditionally yes at the two extremes of the alignment axis, yes under explicit checkable conditions on the band between, and *no in general* — an interval-arithmetic certificate exhibits a criterion-positive instance where a physical two-hop chain strictly beats every single ridge (Theorems D.11 and D.12). In particular no unconditional argument (quasiconvexity of the product-of-resolvents family included) can exist.

Remark J.8 (Interpretation). $1/\lambda_i$ is tail-heavy, and (J.10) re-weights the balanced residual Δ (mean 0 under π) toward the tail. The two-hop chain adds extra shrinkage concentrated on small-λ coordinates, so it helps iff the tail is *under-filtered* ($\Delta < 0$ there): more shrinkage removes noise without discarding signal.

Remark J.9 (Noiseless teacher — chains filter *noise*, verified). As the teacher noise vanishes ($v^{\mathrm{w}} \to 0$) the single optimum $t^* \to 0$ and $F_i^* \to 1$, so the criterion weight $F_i^*(1 - F_i^*) \to 0$ and $\partial R_2/\partial t_2|_{(t^*,0)} \to 0$ (checked: at $(a, b) = (1.2, 2.0)$, scaling $v^{\mathrm{w}} \to \varepsilon v^{\mathrm{w}}$ sends the criterion $-0.50 \to 0$ as $\varepsilon \to 0$). At a strictly noiseless teacher ($v^{\mathrm{w}} = 0$): $t^* = 0$, $F_i^* = 1$, and the boundary derivative $\partial R_2/\partial t_2|_{(0,0)} = 2\sum_i F_i^{\mathrm{w}} \omega_i^2 (1 - F_i^{\mathrm{w}}) \geq 0$ (> 0 for a finite-sample teacher; $= +0.021$ at $(1.2, 2.0)$), so the two-hop chain never strictly wins. Mechanism: no teacher noise, no chain gain.

J.4 Global necessity: the trichotomy and the certified counterexample

Work at the population level of (J.7) in the power-law family, and pass to the elementary-symmetric coordinates $e_1 = t_1 + t_2$, $e_2 = t_1 t_2$: the two-hop filter depends on (t_1, t_2) only through $g_i = e_1\lambda_i + e_2$, via $F_i = \lambda_i^2/(\lambda_i^2 + g_i)$, so the objective $\psi(e_1, e_2) := R(F(e_1, e_2)) = \sum_i h_i(g_i)$ is *separable* across coordinates.

The physical region $\{t_1, t_2 \geq 0\}$ maps onto $D = \{e_2 \geq 0,\ e_1 \geq 2\sqrt{e_2}\}$, which we relax to the closed quadrant $Q = [0,\infty)^2 \supseteq D$; the single-ridge family is the edge $\{e_2 = 0\}$, so $\min_Q \psi = \inf_{e_2=0} \psi$ implies the same over D (the relaxation is used only for exclusions, never to claim ties).

Write $r_i(g) := c_i - F_i\rho_i$ for the per-coordinate residual (the Wiener gap; at the single optimum $r_i = \Delta_i$) and $w_i := F_i^2 r_i$. An interior critical point satisfies the *double orthogonality* $\sum_i w_i = \sum_i w_i/\lambda_i = 0$, and the transversal derivative at the single optimum,

$$D_0 := \partial_{e_2}\psi|_{(t^*,0)} = 2\sum_i (F_i^*)^2 \Delta_i/\lambda_i,$$

has the same sign as criterion (D.10). Two structure lemmas drive everything.

Lemma J.10 (Sign dictionary and root count)**.** *At any $(e_1, e_2) \in Q$ and $\lambda > 0$, with the power-law channel data ($c = F^{\mathrm{w}}\omega^2$, $\rho = (F^{\mathrm{w}})^2\omega^2 + v^{\mathrm{w}}$, $\beta := b/a$),*

$$\operatorname{sign} r(\lambda) = \operatorname{sign} P(\lambda), \qquad P(\lambda) = G\big[(e_1{+}\tau_{\mathrm{w}})\lambda^2 + (e_2{+}\tau_{\mathrm{w}}e_1)\lambda + \tau_{\mathrm{w}}e_2\big] - \Gamma_{\mathrm{w}}\,\lambda^{2-\beta}, \quad \text{(J.11)}$$

a four-term fewnomial. By the generalized Descartes rule (Laguerre; real exponents), P has at most as many positive roots as coefficient sign changes: at most two *for $\beta \in (0,2)$, at most* one *at the extremes $\beta = 0$ and $\beta \geq 2$; and $P \not\equiv 0$ on $Q \setminus \{(0,0)\}$, so on $m \geq 3$ grid points the residual vector is never identically zero.*
End behavior fixes the admissible sign patterns along the spectrum: for $\beta = 0$, subsequences of $(-\,\text{head}, +\,\text{tail})$; on the single-ridge edge with $\beta \geq 1$, subsequences of $(+\,\text{head}, -\,\text{tail})$; for $\beta > 2$ the term $-\Gamma_{\mathrm{w}}\lambda^{2-\beta}$ dominates at $\lambda \to 0^+$, giving a single crossing $(+,-)$ everywhere on the open quadrant.
At $\beta = 2$ exactly, that term merges with the constant: P is a genuine quadratic with constant term $G\tau_{\mathrm{w}}e_2 - \Gamma_{\mathrm{w}}$, whose sign is unrestricted — when it is ≥ 0 (e.g. $e_2 \geq \Gamma_{\mathrm{w}}/(G\tau_{\mathrm{w}})$) all coefficients are positive and there is no *crossing, the all-$(+)$ pattern. Either way the pattern at $\beta = 2$ is a subsequence of $(+,-)$ with at most one sign change — which is all Lemma J.11 consumes, so every downstream exclusion is unaffected.*

Lemma J.11 (Chebyshev exclusions)**.** *$\{1, 1/\lambda\}$ is a Chebyshev system on $(0,\infty)$: if at a solution of the double orthogonality the residual pattern has at most one sign change, then testing against $v(\lambda) = 1/\lambda_0 - 1/\lambda$, with λ_0 at the change, makes $\sum_i w_i v(\lambda_i)$ a sum of one-signed terms equal to zero, forcing $r \equiv 0$ on the grid — impossible by Lemma J.10. Hence no interior critical point exists wherever the pattern count is ≤ 1; a first-order KKT variant excludes local minima on the edge $\{e_1 = 0,\ e_2 > 0\}$ when the pattern there is $(-\,\text{head}, +\,\text{tail})$; and the telescoping identity (J.15) converts a one-change pattern at the single optimum into a strict criterion sign.*

The band $0 < b < a$ requires three transversality/no-saddle conditions, stated here formally. Write $\Sigma w(e_1, e_2) := \sum_i w_i$ and $A_k := \sum_i h_i''(g_i)\,\lambda_i^k$ for $k = 0, 1, 2$, so that the Hessian of ψ in (e_1, e_2) is $H = \left(\begin{smallmatrix} A_2 & A_1 \\ A_1 & A_0 \end{smallmatrix}\right)$ (since $\partial g_i/\partial e_1 = \lambda_i$, $\partial g_i/\partial e_2 = 1$) and $\det H = A_2A_0 - A_1^2$.

Definition J.12 (The transversality conditions (SC1)–(SC3))**.** An instance (a spectrum grid with channel data and prior) satisfies:

- *(SC1)* *e_1-slice transversality* if for every $e_2 \geq 0$, every zero of $e_1 \mapsto \Sigma w(e_1, e_2)$ on $(0,\infty)$ has $A_2 > 0$ there;
- *(SC2)* *A-edge transversality* if every zero of $e_2 \mapsto \sum_i w_i(0, e_2)/\lambda_i$ on $(0,\infty)$ has $A_0 > 0$ there;
- *(SC3)* *no degenerate double-orthogonality points* if every solution of the double-orthogonality system $\Sigma w = \sum_i w_i/\lambda_i = 0$ with $e_1 \geq 0$, $e_2 \geq 0$, $(e_1, e_2) \neq (0,0)$ has $\det H > 0$.

Remark J.13 (Scope and checkability of (SC))*.* The domain in *(SC3)* is deliberately *closed*: $e_2 = 0$ covers the threshold case $D_0 = 0$, and $e_1 = 0$ covers A-edge double points, which the two lemmas above do not exclude for $\beta \in (0,1)$. Each condition is semialgebraic on the instance — rational in (e_1, e_2) for a fixed grid, hence finitely decidable per instance. None is claimed as a general theorem; Theorem D.12 shows that *(SC3)* can genuinely fail.

Nor are they a technical convenience in the band part (iv) covers: the *one-change* hypothesis that carries parts (i)–(iii) through (K3)–(K4) is simply unavailable there. At $(a, b) = (1.2, 0.15)$, $\psi = 0.3$, $\bar{\kappa_{\mathrm{w}}} = 3$, $d/N = 2$, $d = 2 \cdot 10^5$ the residual Δ at the tuned t^* changes sign *twice* — $+$ on $i \leq 773$, $-$ on $774 \leq i \leq 5802$, $+$ again on $i \geq 5803$ — so $(t^*, 0)$ is not one-change and (K4) is silent at it. That

is a numerical observation, and it is the reason part (iv) must carry *(SC1)–(SC3)* rather than reuse the Chebyshev test.

Theorem D.11 (Global necessity: a trichotomy)**.** *Population level, power-law family; for (ii)–(iv) assume the single-ridge risk attains its infimum at an interior t^* — automatic for the MNI teacher (C.3) here, by the two conditions stated with (D.9), and not automatic for a ridge teacher.*

(i) $b = 0$: unconditional tie. $\min_Q \psi = \inf_{e_2=0} \psi$*: no two-hop chain, physical or relaxed, ever strictly beats the best single ridge.*

Moreover at every interior single optimum the pattern is $(-\,\mathrm{head}, +\mathrm{tail})$ *and the criterion is* strictly positive *— the $b = 0$ failure of (D.10) is a theorem, for every overparametrized teacher ($\tau_{\mathrm{w}} > 0$; Assumption C.1).*

(ii) $b \geq a$: chains always win. *At every interior single optimum the pattern is exactly* $(+\,\mathrm{head}, -\mathrm{tail})$ *and the criterion is strictly negative: the two-hop chain strictly wins for* all $b \geq a$.

The criterion-failing region is therefore contained in $\{b < a\}$, and necessity is vacuous on $b \geq a$.

(iii) $b \geq 2a$: equal-ridge rigidity. *ψ has no critical point in the open quadrant; consequently every minimizer of ψ over the physical region D lies on the parabola $t_1 = t_2$: the optimal two-hop chain has exactly equal ridges $t_1^* = t_2^* > 0$.*

(iv) $0 < b < a$: conditional necessity. *Under the transversality/no-saddle conditions* (SC1)–(SC3) *of Definition J.12 — per-instance semialgebraic conditions, sampled at scale with zero violations on the criterion-failing region, not exhaustively verified — criterion ≥ 0 at t^* implies the global tie $\min_Q \psi = R_1^{\mathrm{pop},*}$. Combined with sufficiency (Theorem D.10 (i)): under* (SC)*, strict global win $\iff$ criterion < 0.*

Parts (i)–(iii) share four elementary tools. Throughout $c_i = \rho_i = 0$ for $i > m$, hence $r_i = w_i = 0$ there and *every sum runs over $i \leq m$*, where $F_i > 0$; the grid $\lambda_i = i^{-a}$ strictly decreases, so *head* means small i and *tail* large i. In the Wiener language of Proposition C.4 everything below is one residual: $\pi_i(F_i - g_i) = -\lambda_i r_i$, and the two directions in which the two-hop family moves the filter are $\partial F_i/\partial e_1 = -F_i^2/\lambda_i$ and $\partial F_i/\partial e_2 = -F_i^2/\lambda_i^2$, so an interior critical point is exactly a residual that is π-orthogonal to *both* (Proposition C.6 (ii)) — equivalently, w-orthogonal to the Chebyshev pair $\{1, 1/\lambda\}$ — which (K3) shows a residual with at most one sign change cannot be.

Lemma J.14 (Four tools)**.** *Assume $\tau_{\mathrm{w}} > 0$, $G > 0$, $\Gamma_{\mathrm{w}} \geq 0$, $m \geq 2$. Then:*

(K1) Gradient: *ψ is real-analytic on Q, with $\partial_{e_1}\psi = 2\,\Sigma w$, $\partial_{e_2}\psi = 2\,\Sigma w/\lambda$, and $\psi(e_1, 0) = R_1(e_1)$ on the single-ridge edge.*

(K2) Compactification: *ψ extends continuously to the one-point compactification $Q^+ := Q \cup \{\infty\}$ by $\psi(\infty) := R(\mathbf{0}) = \sum_i \lambda_i \omega_i^2 = \lim_{t\to\infty} \psi(t, 0)$, the zero-filter risk; hence ψ attains its minimum on Q^+, and on $D^+ := D \cup \{\infty\}$.*

(K3) Chebyshev test: *call $p \in Q$* one-change *if for some $\lambda_0 > 0$ off the grid, $r_i(p) \leq 0$ whenever $\lambda_i > \lambda_0$ and $r_i(p) \geq 0$ whenever $\lambda_i < \lambda_0$* (type $(-,+)$)*, or the same with both inequalities reversed* (type $(+,-)$)*; a constant-sign pattern is one-change of either type. If p is one-change and $r(p) \not\equiv 0$ on the grid, then* (a) *p is not a critical point of ψ, and* (b) *if p is of type $(-,+)$ and lies on the A-edge $\{e_1 = 0,\ e_2 > 0\}$, it is not a local minimizer of ψ over Q.*

(K4) Telescoping sign rule: *let $t > 0$ satisfy $R_1'(t) = 0$ and write $\Delta_i := r_i(t, 0)$, $F_i^* := \lambda_i/(\lambda_i + t)$. If $\Delta \not\equiv 0$ then Δ takes both signs; and if $(t, 0)$ is one-change then $\sum_i F_i^*(1 - F_i^*)\Delta_i = \frac{t}{2} D_0(t)$ is > 0 for type $(-,+)$ and < 0 for type $(+,-)$.*

Proof. **(K1).** With $h_i(g) = \lambda_i(\omega_i^2 - 2F(g)c_i + F(g)^2\rho_i)$, $F(g) = \lambda_i^2/(\lambda_i^2 + g)$, $F'(g) = -F^2/\lambda_i^2$, we get $h_i'(g) = 2F^2(c_i - F\rho_i)/\lambda_i = 2F^2 r_i/\lambda_i$; the chain rule with $\partial g_i/\partial e_1 = \lambda_i$, $\partial g_i/\partial e_2 = 1$ gives both derivatives (recorded again as (J.13) below), and $\lambda_i^2 + g_i \geq \lambda_i^2 > 0$ on Q gives analyticity. At $e_2 = 0$, $F_i = \lambda_i/(\lambda_i + e_1)$, i.e. $F = \varphi_{e_1}$, so $\psi(e_1, 0) = R(\varphi_{e_1}) = R_1(e_1)$.

(K2). For $i \leq m$, $|h_i(g_i) - \lambda_i\omega_i^2| \leq \lambda_i F_i(2c_i + \rho_i)$ (using $0 < F_i \leq 1$), while $F_i \leq \lambda_i^2/g_i$ and $g_i = e_1\lambda_i + e_2 \geq \min(\lambda_m, 1)(e_1 + e_2)$; the bound tends to 0 uniformly over the m accessible

coordinates as $e_1 + e_2 \to \infty$, and the $i > m$ terms equal $\lambda_i \omega_i^2$ identically. So $\psi \to R(\mathbf{0})$ along *every* escape to infinity in Q, which is continuity of the extension at the added point; Q^+ is compact and D^+ is a closed subset of it. Finally $\psi(t,0) = R(\varphi_t) \to R(\mathbf{0})$ since $\varphi_t \to \mathbf{0}$ coordinatewise.

(K3). Put $v(\lambda) := 1/\lambda_0 - 1/\lambda$, so $v(\lambda_i) > 0$ for $\lambda_i > \lambda_0$, $v(\lambda_i) < 0$ for $\lambda_i < \lambda_0$, and $v(\lambda_i) \neq 0$ always. As $w_i = F_i^2 r_i$ with $F_i^2 > 0$, type $(-,+)$ makes every product $w_i v(\lambda_i) \leq 0$ and type $(+,-)$ makes every product ≥ 0; in both cases

$$\textstyle\sum_i w_i\, v(\lambda_i) \;=\; \frac{1}{\lambda_0}\,\Sigma w \;-\; \Sigma w/\lambda. \tag{J.12}$$

(a) At a critical point (K1) gives $\Sigma w = \Sigma w/\lambda = 0$, so (J.12) vanishes; a one-signed sum vanishing forces every term to vanish, and $v(\lambda_i) \neq 0$ then forces $r_i = 0$ for all $i \leq m$ — contradicting $r(p) \not\equiv 0$. *(b)* At $p = (0, e_2)$, $e_2 > 0$, the feasible cone of Q is $\{u : u_1 \geq 0\}$, so a local minimizer satisfies the first-order (KKT) conditions $\partial_{e_2}\psi(p) = 0$ (e_2 is unconstrained near p) and $\partial_{e_1}\psi(p) \geq 0$ (the only feasible e_1-variation is positive), i.e. $\Sigma w/\lambda = 0$ and $\Sigma w \geq 0$.

Then (J.12) equals $\Sigma w/\lambda_0 \geq 0$ while type $(-,+)$ makes every summand ≤ 0: again $r \equiv 0$ on the grid, a contradiction. (For type $(+,-)$ the two signs agree and no contradiction arises — which is why the A-edge is excluded in (i) but not for $\beta \geq 1$, where we instead work over D, whose only point with $e_1 = 0$ is the origin.)

(K4). Write $u_i := (F_i^*)^2 \Delta_i$. By (K1), $0 = R_1'(t) = \partial_{e_1}\psi(t,0) = 2\sum_i u_i$, which is (J.9); as $(F_i^*)^2 > 0$, a one-signed $\Delta \not\equiv 0$ would make $\sum_i u_i \neq 0$, so both signs occur.

With $M_k := \sum_{i \geq k} u_i$ we have $M_1 = M_{m+1} = 0$ and $u_i = M_i - M_{i+1}$, so Abel summation gives exactly (J.15):

$$\sum_i u_i/\lambda_i = \sum_{k \geq 2} (1/\lambda_k - 1/\lambda_{k-1}) M_k,$$

with strictly positive weights.

Let $\{1, \dots, j\}$ be the head $\{i : \lambda_i > \lambda_0\}$. For type $(-,+)$, $u_i \leq 0$ on the head and $u_i \geq 0$ on the tail, so $M_k = -\sum_{i<k} u_i \geq 0$ for $k \leq j+1$ and $M_k = \sum_{i \geq k} u_i \geq 0$ for $k \geq j+1$: every $M_k \geq 0$. The negative u_i live in the head, so $j \geq 1$ and $k_0 := \min\{i : u_i > 0\} \geq j+1 \geq 2$, whence $M_{k_0} \geq u_{k_0} > 0$ (all indices $\geq k_0$ are tail indices): the k_0-th telescoped term is > 0, the rest ≥ 0, so $\sum_i u_i/\lambda_i > 0$. Type $(+,-)$ is the same with u replaced by $-u$.

Finally $F_i^*(1 - F_i^*) = (F_i^*)^2 t/\lambda_i$, so

$$\sum_i F_i^*(1 - F_i^*)\Delta_i = t \sum_i u_i/\lambda_i = \tfrac{t}{2} D_0(t)$$

— the computation (J.8), which used only stationarity (J.9) and so is valid at any interior stationary t. □

Proof of part (i). Here $\beta = 0$; the hypotheses used are $\tau_{\mathrm{w}} > 0$, $G > 0$ (Assumption C.1) and $m \geq 2$, the interior-optimum assumption entering only in Step 3.

Step 1 (the pattern, at every point of Q). At $\beta = 0$ the term $-\Gamma_{\mathrm{w}}\lambda^{2-\beta}$ of (J.11) merges with the leading term, leaving the genuine quadratic $P(\lambda) = A\lambda^2 + B\lambda + C$ with $A := G(e_1 + \tau_{\mathrm{w}}) - \Gamma_{\mathrm{w}}$, $B := G(e_2 + \tau_{\mathrm{w}} e_1) \geq 0$, $C := G\tau_{\mathrm{w}} e_2 \geq 0$. Fix $p = (e_1, e_2) \in Q \setminus \{(0,0)\}$; then $B > 0$ because $\tau_{\mathrm{w}}, G > 0$, so $P \not\equiv 0$.

If $A \geq 0$ then $P > 0$ on $(0, \infty)$; if $A < 0$ then P is concave with $P(0) = C \geq 0$, slope $B > 0$ at the origin and $P \to -\infty$, so $P > 0$ on $(0, \lambda_\times)$ and $P < 0$ on $(\lambda_\times, \infty)$ for a unique $\lambda_\times > 0$. Since $\operatorname{sign} r_i = \operatorname{sign} P(\lambda_i)$, in either case p is one-change of type $(-,+)$ — negative head, positive tail — with a cut off the grid, and $r(p) \not\equiv 0$ there as soon as $m \geq 2$, P having at most one positive zero. (The root count is elementary here; no appeal to the generalized Descartes rule of Lemma J.10 is needed.)

Step 2 (exclusions, then compactify). By Step 1 and Lemma J.14 (K3), ψ has no critical point in the open quadrant $\{e_1, e_2 > 0\}$ and no point of the A-edge $\{e_1 = 0,\ e_2 > 0\}$ is a local minimizer

of ψ over Q. By (K2), ψ attains its minimum over the compact Q^+, at some p^*. If $p^* = \infty$ then $\inf_Q \psi = \lim_{t\to\infty} \psi(t,0) \geq \inf_{e_2=0} \psi \geq \inf_Q \psi$, so all three agree.

Otherwise $p^* \in Q$ is a global — hence local — minimizer over Q, so by the two exclusions it lies on the closed single-ridge edge $\{e_2 = 0\}$ (which contains the origin). Either way $\inf_Q \psi = \inf_{e_2=0} \psi = \min_{t\in[0,\infty]} R_1(t) = R_1^{\mathrm{pop},*}$, the last equalities by (K1)–(K2) with $\varphi_\infty = \mathbf{0} \in \mathcal{R}_1$ (Definition C.5); in particular the single-ridge infimum is attained. As $\{e_2 = 0\} \subset D \subset Q$, the same value is the infimum over the physical region: no two-hop chain, physical or relaxed, strictly beats the best single ridge.

Step 3 (the criterion at an interior optimum). Let $t^* \in (0,\infty)$ be a stationary point of R_1. Then $(t^*, 0) \in Q \setminus \{(0,0)\}$, so by Step 1 $\Delta = r(t^*, 0)$ is one-change of type $(-,+)$ with $\Delta \not\equiv 0$; by (K4) both signs occur — the pattern is therefore *exactly* $(-\,\mathrm{head}, +\,\mathrm{tail})$, both parts nonempty — and $\sum_i F_i^*(1-F_i^*)\Delta_i = \frac{t^*}{2} D_0 > 0$: criterion (D.10) fails strictly, at every interior single optimum, for every overparametrized teacher. □

Proof of part (ii). Here $\beta \geq 1$; hypotheses: $\tau_{\mathrm{w}} > 0$, $G > 0$, $\Gamma_{\mathrm{w}} \geq 0$, $m \geq 2$, and the standing assumption of an interior optimum $t^* \in (0,\infty)$ — supplied, for the MNI teacher, by (D.9) together with $\sum_i \lambda_i^2 F_i^{\mathrm{w}} \omega_i^2 > 0$ (both in the proof of Remark C.7 (b)), and both automatic in the power-law family; it is *not* equivalent to $s_{\mathrm{w}}^2 > 0$ (Remark J.6).

Step 1 (edge pattern, by monotonicity). On the edge $e_2 = 0$, $e_1 = t > 0$, (J.11) factors as $P(\lambda) = \lambda\, q(\lambda)$ with $q(\lambda) = G(t+\tau_{\mathrm{w}})\lambda + G\tau_{\mathrm{w}} t - \Gamma_{\mathrm{w}} \lambda^{1-\beta}$. Since $1-\beta \leq 0$, the last term is nondecreasing in λ and the first is strictly increasing, so q is strictly increasing on $(0,\infty)$: it has at most one zero and no sign change other than $-$ to $+$ as λ grows. Hence $(t,0)$ is one-change of type $(+\,\mathrm{head}, -\,\mathrm{tail})$ with $r \not\equiv 0$ on the grid ($m \geq 2$).

Step 2 (both signs, then strictness). $R_1'(t^*) = 0$, so (K4) applies: Δ takes both signs, the pattern is *exactly* $(+\,\mathrm{head}, -\,\mathrm{tail})$ with both parts nonempty, and telescoping gives $\sum_i F_i^*(1-F_i^*)\Delta_i = \frac{t^*}{2} D_0 < 0$: criterion (D.10) holds strictly. (K4) asks only for stationarity, so this conclusion holds at *every* interior stationary point of R_1 — in particular at the global interior minimizer the preamble supplies, which is the one Step 3 uses.

Step 3 (a physical chain strictly wins). Take t^* to be the interior point at which R_1 attains its infimum, as the preamble provides. As $\partial_{e_2}\psi(t^*,0) = D_0 < 0$, $\psi(t^*,\varepsilon) < \psi(t^*,0) = R_1(t^*) = R_1^{\mathrm{pop},*}$ for small $\varepsilon > 0$, the last equality being exactly the *globality* of t^* — the same quantifier as in Theorem D.10 (i), and not removable there or here (Remark J.7); and (t^*,ε) is physical for $0 < \varepsilon < (t^*)^2/4$, its hops being the roots $t_{1,2} = \frac{1}{2}(t^* \pm \sqrt{(t^*)^2 - 4\varepsilon}) > 0$ of $z^2 - t^* z + \varepsilon$.

This is sufficiency (Theorem D.10 (i)) applied to a criterion now *proved* negative. Consequently $\{b : \text{criterion} \geq 0\} \subseteq [0,a)$: the criterion fails nowhere on $[a,\infty)$ and necessity is vacuous on $b \geq a$, so at every d the crossing $b^*(a;d)$ of Remark D.13 satisfies $b^*(a;d) \leq a$. Strictness does *not* follow. $b^*(a;d) < a$ would require the criterion to hold on a whole interval $[b_0, a]$ with $b_0 < a$, and nothing proved here excludes a failure set accumulating at a from below: part (ii) reaches only $b \geq a$, and Remark D.13 deliberately declines to assume that the sign change is unique, so continuity is not available either. Numerically $b^*(a;d)$ lies far below a (Remark D.13). □

Proof of part (iii). Here $\beta \geq 2$; since $a > 1$ this forces $b > a$, so part (ii) and its hypotheses are in force — in particular the preamble hypothesis that R_1 attains its infimum over $[0,\infty]$ at an interior t^*. Nothing below uses $s_{\mathrm{w}}^2 > 0$.

Step 1 (no critical point in the open quadrant). Fix $e_1, e_2 > 0$. In (J.11) the bracket $(e_1+\tau_{\mathrm{w}})\lambda^2 + (e_2+\tau_{\mathrm{w}}e_1)\lambda + \tau_{\mathrm{w}} e_2$ has nonnegative coefficients and a strictly positive leading one, hence is strictly increasing on $(0,\infty)$, while $2-\beta \leq 0$ makes $-\Gamma_{\mathrm{w}}\lambda^{2-\beta}$ nondecreasing. So P is strictly increasing, has at most one positive zero and is $\not\equiv 0$: every point of the open quadrant is one-change of type $(+,-)$ with $r \not\equiv 0$ on the grid.

By Lemma J.14 (K3)(a) there is no critical point of ψ in the open quadrant — the first assertion of (iii). (At $\beta = 2$ exactly, $-\Gamma_{\mathrm{w}}\lambda^0$ merges with $G\tau_{\mathrm{w}} e_2$ and the constant term has unrestricted sign; monotonicity, hence the argument, is unaffected — the $\beta = 2$ split of Lemma J.10.)

Step 2 (a finite minimizer over D exists). By part (ii) there is a physical $p_0 = (t^*, \varepsilon) \in D$ with $\psi(p_0) < R_1^{\mathrm{pop},*} \le R_1(\infty) = \psi(\infty)$. By (K2) ψ attains its minimum over the compact D^+; the value at ∞ is beaten by $\psi(p_0)$, so every minimizer p^* lies in D. The interior of $D = \{e_2 \ge 0,\ e_1 \ge 2\sqrt{e_2}\}$ is $\{e_2 > 0,\ e_1 > 2\sqrt{e_2}\}$, a subset of the open quadrant, where a minimizer would be a critical point — excluded by Step 1. Hence $p^* \in \partial D = \{e_2 = 0,\ e_1 \ge 0\} \cup \{e_1 = 2\sqrt{e_2},\ e_2 \ge 0\}$.

Step 3 (the single-ridge edge is excluded). Step 2 has already produced a point $p_0 \in D$ with $\psi(p_0) < R_1^{\mathrm{pop},*}$. Since $\{(t, 0) : t \ge 0\} \subset D$ and $\psi(t, 0) = R_1(t) \ge R_1^{\mathrm{pop},*}$ for *every* $t \ge 0$ — the latter by definition of the global infimum — every point of the edge $\{e_2 = 0\}$, the origin included, satisfies $\psi > \psi(p_0) \ge \min_{D^+} \psi$. Hence no minimizer lies on that edge.

That is the whole step, and it is worth recording what the one line buys: it uses no derivative of R_1 at 0, no $s_{\mathrm{w}}^2 > 0$, and not (D.9) — hence no minimum-norm restriction on the teacher. One might instead try to exclude $\bar{t} = 0$ by quoting $R_1'(0) = -2s_{\mathrm{w}}^2/\tau_{\mathrm{w}} < 0$; that identity is minimum-norm-only and *fails for every ridge teacher* (Remark J.6 (a)), whereas the theorem admits any teacher meeting its interior-infimum preamble, ridge teachers included. What the step does need is that $R_1^{\mathrm{pop},*}$ is the *global* single-ridge value — the quantifier installed in Theorem D.10 (i) and Remark J.7. Under a merely local reading, part (ii) would deliver only $\psi(p_0) < R_1(t^*)$ at some possibly non-global t^* and this step would not close.

Step 4 (conclusion). What remains of ∂D is the parabola $\{e_1 = 2\sqrt{e_2},\ e_2 > 0\}$, i.e. $t_1 = t_2 = \sqrt{e_2} > 0$: every minimizer over the physical region has exactly equal ridges $t_1^* = t_2^* > 0$. Deep in the aligned regime the optimal two-hop chain is rigid. □

Remark J.15 (The threshold $b \ge 2a$ is sufficient, not necessary). What $b \ge 2a$ (i.e. $\beta \ge 2$) buys is Step 1's monotonicity, and nothing in the *phenomenon* appears to need it. Numerically the equal-ridge conclusion persists far below the threshold. At $d = 4000$, $\psi = 0.3$, $\bar{\kappa_{\mathrm{w}}} = 3$, $\sigma^2 = 1/4$, $r^2 = 1$, minimizing the population two-hop objective over the physical region on a grid in the asymmetry $(t_1 - t_2)/(t_1 + t_2)$ with the scale $t_1 + t_2$ re-optimized at every grid point: at $a = 1.2$ (threshold $2a = 2.4$) the minimizer has exactly $t_1^* = t_2^*$ for every scanned $b \in [0.40,\ 2.6]$, and the symmetric point is a strict local minimum in the asymmetry there (second derivative $+6.7$ at $b = 0.40$, $+22.8$ at $b = 0.50$), while at $b = 0.36$ it is a local maximum and the optimum moves to an asymmetric interior point.

At $a = 2.0$ (threshold 4.0) the same holds for every scanned $b \in [0.5,\ 3.9]$ and fails by $b = 0.2$. In both cases the breakdown sits near the criterion crossing $b^*(a; 4000)$, not near $2a$: as $b \downarrow 0$ the optimum must approach the single-ridge edge $t_2 = 0$, which is the $b = 0$ tie of part (i). This is a scan, not a theorem — extending (iii) below $2a$ would need a replacement for Step 1's sign argument, which we do not have.

Remark J.16 (Hypothesis roll-call). Parts (i)–(iii) use only $\tau_{\mathrm{w}} > 0$ and $G > 0$ (Assumption C.1), $\Gamma_{\mathrm{w}} \ge 0$ (automatic from (C.6)), $m \ge 2$ grid points (and $m \ge 3$ wherever Lemma J.10 is invoked), and — for (ii), for (iii) through its Step 2, and for the second half of (i) only — the preamble hypothesis that R_1 attains its infimum over $[0, \infty]$ at an interior t^*, which for the MNI teacher follows from $s_{\mathrm{w}}^2 > 0$ together with $\sum_i \lambda_i^2 F_i^{\mathrm{w}} \omega_i^2 > 0$ (Remark C.7 (b)) but is equivalent to neither (Remark J.6).

Two clauses of that sentence are load-bearing. First, the hypothesis is used in its *global* form everywhere: under the local reading Step 3 of (ii) controls the wrong quantity and Step 3 of (iii) does not close at all (Remark J.7). Second, no step uses $s_{\mathrm{w}}^2 > 0$ directly — Step 3 of (iii) formerly did, through the minimum-norm-only identity (D.9), and no longer does — so parts (i)–(iii) hold verbatim for a ridge teacher that meets the preamble, which is exactly the scope the preamble announces.

The global tie of (i) is unconditional. No part uses *(SC1)–(SC3)*, and none uses the generalized Descartes rule: in these three regimes the root count of Lemma J.10 follows from monotonicity of P (concavity, at $\beta = 0$), independently of multiplicity conventions.

Proof of part (iv). Write $\psi^* := R_1^{\mathrm{pop},*}$ and abbreviate $\Sigma w := \sum_i w_i$ and $\Sigma w/\lambda := \sum_i w_i/\lambda_i$ at the current point of Q. Two elementary facts are used throughout (direct differentiation of $h_i(g) = \lambda_i(\omega_i^2 - 2F(g)c_i + F(g)^2\rho_i)$ with $F(g) = \lambda_i^2/(\lambda_i^2 + g)$, $F'(g) = -F^2/\lambda_i^2$, $g_i = e_1\lambda_i + e_2$):

$$h_i'(g) = \tfrac{2}{\lambda_i} F^2 r_i \implies \partial_{e_1}\psi = 2\,\Sigma w, \quad \partial_{e_2}\psi = 2\,\Sigma w/\lambda;$$

$$\frac{\partial r_i}{\partial e_1} = \frac{\rho_i F^2}{\lambda_i} > 0, \quad \frac{\partial r_i}{\partial e_2} = \frac{\rho_i F^2}{\lambda_i^2} > 0 : \tag{J.13}$$

along either coordinate direction each residual r_i crosses zero at most once, upward. Assume criterion ≥ 0 at t^*, i.e. $D_0 \geq 0$.

Step 0 (slice ends). By (J.13), $r_i(\cdot, e_2)$ crosses 0 in e_1 at most once, upward, at the root of $P(\lambda_i) = 0$ [from the dictionary (J.11)], which is *affine* in the parameters:

$$x_i(e_2) = \frac{\Gamma_{\mathrm{w}} \lambda_i^{2-\beta} - G\big(\tau_{\mathrm{w}} \lambda_i^2 + e_2 \lambda_i + \tau_{\mathrm{w}} e_2\big)}{G\, \lambda_i (\lambda_i + \tau_{\mathrm{w}})},$$

decreasing in e_2. For $e_1 > X^* := \max_i \max(x_i(0), 0)$ — finite, independent of e_2 — every $r_i > 0$, hence $\Sigma w > 0$: every e_1-slice ends positive, uniformly in e_2. Since ψ is real-analytic on Q and tends to the zero-filter risk $\sum_i \lambda_i \omega_i^2$ as $e_1 + e_2 \to \infty$, each slice minimum over $e_1 \geq 0$ is attained in $[0, X^*]$.

Step 1 (slice classification under (SC1)). Fix $e_2 \geq 0$. Under *(SC1)* every zero of $e_1 \mapsto \Sigma w(e_1, e_2)$ on $(0, \infty)$ is a transversal up-crossing ($\partial_{e_1}(2\,\Sigma w) = 2A_2 > 0$ there), so there is at most one: two consecutive up-crossings would need a down-crossing between them.

Two cases. *(B)*: one zero $\xi(e_2) \in (0, X^*]$; then $\Sigma w < 0$ before it and > 0 after, so $e_1 \mapsto \psi$ strictly decreases then increases: unique argmin $\xi(e_2)$. *(Z)*: no zero; then $\Sigma w \geq 0$ on $(0, \infty)$ (the slice ends positive and cannot cross), so ψ is nondecreasing in e_1: argmin $e_1 = 0$.

Define the envelope $M(e_2) := \min_{e_1 \geq 0} \psi(e_1, e_2)$, equal to $\psi(\xi(e_2), e_2)$ on B and to $\psi(0, e_2)$ on Z; M is continuous (an envelope of continuous functions with the locally uniform coercivity of Step 0; the argmin is unique on B). By construction $\min_Q \psi = \min_{e_2 \geq 0} M(e_2)$, and $e_2 = 0$ is of type B with $\xi(0) = t^*$ and $M(0) = \psi^*$: on the edge, $\Sigma w = 0$ is exactly the single-optimality identity (J.9) at the interior optimum t^* (which exists by the standing assumption), unique by *(SC1)*.

Step 2 (branch regularity; the Schur form). On a B-component, $A_2 > 0$ at $(\xi(e_2), e_2)$ and the implicit function theorem make ξ a C^1 function with $\xi'(e_2) = -A_1/A_2$; along the branch,

$$M'(e_2) = 2\, D(e_2), \quad D(e_2) := \Sigma w/\lambda \big|_{(\xi(e_2),\, e_2)}, \qquad D'(e_2) = \tfrac{1}{2}\big(A_0 - A_1^2/A_2\big) = \frac{\det H}{2A_2} \tag{J.14}$$

— the envelope theorem for the first identity ($\partial_{e_1}\psi = 0$ at the argmin), and for the second, $D' = \frac{1}{2}\,(\partial_{e_2} + \xi'\,\partial_{e_1})\,\partial_{e_2}\psi = \frac{1}{2}(A_0 - A_1^2/A_2)$: the Schur complement of the slice entry A_2 in the Hessian H.

On a Z-interval, $M'(e_2) = \partial_{e_2}\psi(0, e_2) = 2\,\Sigma w(0, e_2)/\lambda =: \psi'_A(e_2)$. (The identity (J.14) was verified numerically along a branch: $dD/de_2 = -4.1127 \times 10^5$ against $\det H/(2A_2) = -4.1151 \times 10^5$, relative error 6×10^{-4} at a finite-difference step.)

Step 3 (sign propagation on B). Claim: on the B-component containing $e_2 = 0$, $D \geq 0$ throughout, given $D(0) = D_0/2 \geq 0$. Suppose not, and let $s_0 := \inf\{e_2 : D(e_2) < 0\}$ within the component. By continuity $D(s_0) = 0$; then $(\xi(s_0), s_0)$ solves the double orthogonality [$\Sigma w = 0$ along the branch, $\Sigma w/\lambda = D(s_0) = 0$; the case $s_0 = 0$ is the threshold case $D_0 = 0$, covered because the domain of *(SC3)* is deliberately closed], so *(SC3)* with (J.14) gives $D'(s_0) > 0$.

Then $D > 0$ on $(s_0, s_0 + \delta)$, contradicting that points with $D < 0$ accumulate at s_0 from the right; and $D < 0$ just before s_0 would contradict the minimality of s_0. No such s_0 exists.

Step 4 (regime switches; the double-corner closure). A B-component can end only by $\xi \downarrow 0$: the branch is bounded by X^* (Step 0); an interior tangency is excluded by *(SC1)*; and ξ *converges* at the end of its component — two distinct cluster values would, by connectedness of the cluster set of a continuous function, force a whole interval of zeros of $\Sigma w(\cdot, \bar{e}_2)$, whose interior points have $A_2 = 0$, violating *(SC1)* — while a positive limit would let the branch continue past the end by the implicit function theorem.

At such an end $\bar{e}_2$: $\Sigma w(0, \bar{e}_2) = 0$ and $M'(\bar{e}_2^+) = \psi'_A(\bar{e}_2) = 2 \lim_{e_2 \uparrow \bar{e}_2} D \geq 0$ by Step 3. *Can the corner be double*, i.e. also $\psi'_A(\bar{e}_2) = 0$, so that $\Sigma w = \Sigma w/\lambda = 0$ at $(0, \bar{e}_2)$? No: *(SC3)* —

whose closed domain includes the edge $e_1 = 0$ precisely for this case — gives $\det H > 0$ there, which forces $A_2 > 0$ for free ($A_2 \geq 0$ as the limit of the branch values, and $A_2 = 0$ would give $\det H = -A_1^2 \leq 0$).

Both signs then transfer to a left-neighbourhood, and that step is not skippable: the mean-value point θ_y below is an *interior* branch point, where $\Sigma w = 0$ but in general $\Sigma w/\lambda \neq 0$, so it is *not* a double-orthogonality point and *(SC3)* is silent there — *(SC1)* alone gives only $A_2 > 0$. But $\det H$ and A_2 are continuous in (e_1, e_2) and the branch point $(\xi(y), y)$ tends to $(0, \bar{e}_2)$ as $y \uparrow \bar{e}_2$, so there is $\delta > 0$ with $\det H > 0$ *and* $A_2 > 0$ at every branch point with $y \in (\bar{e}_2 - \delta, \bar{e}_2)$.

Now D extends continuously to $\bar{e}_2$ with value $\psi'_A(\bar{e}_2)/2 = 0$ and is differentiable inside, so the mean value theorem applied to (J.14) gives, for $y \in (\bar{e}_2 - \delta, \bar{e}_2)$ and some $\theta_y \in (y, \bar{e}_2)$,

$$D(y) \;=\; D(\bar{e}_2) - (\bar{e}_2 - y)\, D'(\theta_y) \;=\; -(\bar{e}_2 - y)\left[\frac{\det H}{2A_2}\right](\theta_y) \;<\; 0,$$

contradicting Step 3's propagated $D \geq 0$.

So B-components never end at double corners. At every end $\bar{e}_2$ we then have $\psi'_A(\bar{e}_2) = 2\lim_{e_2 \uparrow \bar{e}_2} D \geq 0$ by Step 3 and $\psi'_A(\bar{e}_2) \neq 0$ by the above, hence $\psi'_A(\bar{e}_2) > 0$ strictly — and this covers both continuations. If a Z-interval of positive length follows, ψ'_A *enters it strictly positive*. If instead a new B-component starts immediately at $\bar{e}_2$ (a B→B touch, with no intervening Z), it starts with $D(\bar{e}_2^+) = \psi'_A(\bar{e}_2)/2 > 0$ and Step 3 applies afresh, exactly as at the Z→B re-emergence below.

Across the Z-interval, *(SC2)* makes every zero of ψ'_A an up-crossing, so the Step-3 argument with A_0 in place of $\det H/A_2$ propagates $\psi'_A \geq 0$; moreover a Z-interval cannot *end* at a ψ'_A-zero under *(SC2)* — an up-crossing forces $\psi'_A < 0$ just before it, contradicting the propagated sign.

Hence at a Z→B re-emergence $\hat{e}_2$ (a slice zero re-emerging from $e_1 = 0$; from $e_1 = \infty$ it cannot, by Step 0), $\psi'_A(\hat{e}_2) > 0$ strictly, and the new branch starts with $D(\hat{e}_2^+) = \psi'_A(\hat{e}_2)/2 > 0$: Step 3 applies afresh to the new component.

The induction over components is legitimate because the switch points are zeros of the real-analytic function $e_2 \mapsto \Sigma w(0, e_2) \not\equiv 0$ (as $e_2 \to \infty$ it behaves like $\left[\sum_i \lambda_i^4 c_i\right]/e_2^2 > 0$), hence isolated: the regime map is locally finite. (Empirically, on every instance tested the regime map is a single B→Z transition with no re-emergence and no double corner; extending the *(SC3)* scan to the closed domain costs nothing extra.)

Step 5 (conclusion). M is continuous with one-sided derivatives ≥ 0 everywhere: M is non-decreasing, so $\min_Q \psi = \min_{e_2 \geq 0} M(e_2) = M(0) = \psi^*$; and since the physical region satisfies $E \subset D \subset Q$, the same tie holds over D: no physical two-hop chain with $t_2 > 0$ strictly beats the best single ridge. Combined with the sufficiency direction (Theorem D.10 (i)): under *(SC)*, strict global win $\iff$ criterion < 0. □

The counterexample below sits precisely where *(SC3)* fails near the threshold (numerically: a saddle with $\det H < 0$ coexists with the dip, Remark J.18), so the conditions are necessary, not decorative; their status — semialgebraic and finitely decidable per instance, sampled at scale with zero violations on the criterion-failing region, but *not* proved as general inequalities — is exactly as recorded immediately after Definition J.12 and in the statement of part (iv).

For the next theorem, fix the standard teacher configuration $(m, n_w, \sigma^2) = (1200, 400, 0.25)$, $d = 4000$, $(a, b) = (2.0, 0.0675)$, and the signal normalization $\sum_i \lambda_i \omega_i^2 = r^2 = 1$. All certified objective values below refer to the *centered* objective: the filter-independent constant $\sum_i \lambda_i \omega_i^2 = r^2 = 1$ — the *full* sum, never-learned tail included — is dropped, hence the negative values. Both sides of the comparison shift by the same constant, so the margin is centering-invariant.

Theorem D.12 (No unconditional converse)**.** *At the power-law instance $d = 4000$, $m = 1200$, $n_w = 400$, $\sigma^2 = 0.25$, $r^2 = 1$, $(a, b) = (2.0, 0.0675)$ — all values below referring to the* centered *population objective ψ of Section J.4, which drops a filter-independent constant, so the margin is centering-invariant — the transversal derivative at the single optimum $t^* = 2.9057934773 \times 10^{-3}$ is $D_0 := \partial_{e_2}\psi|_{(t^*,0)} = +0.3667620 > 0$, so criterion (D.10) is positive.*

Nevertheless, the interior point $(e_1, e_2) = (2.9043477340 \times 10^{-3},\ 1.8160576582 \times 10^{-9})$ — the physical two-hop chain $(t_1, t_2) = (2.90372 \times 10^{-3},\ 6.25424 \times 10^{-7})$, both hops strictly positive —

satisfies

$$\psi(e_1, e_2) - R_1^{\mathrm{pop},*} \;=\; -2.20937164288882 \times 10^{-10} \;<\; 0, \tag{D.11}$$

Thus criterion ≥ 0 *does* not *imply global optimality of the single ridge, and Theorem D.10 (i) cannot be upgraded to an unconditional iff: any unconditional necessity theorem on* $0 < b < a$ *must carry an (SC3)-type hypothesis.*

Every quantity above is enclosed by interval arithmetic with directed rounding, so the sign of (D.11) *is a rigorous computer-assisted verification, not a floating-point evaluation.*

Remark J.17 (Certification details). The derivative D_0 has the same *sign* as criterion (D.10), which equals $(t^*/2)\, D_0 \approx 5.33 \times 10^{-4}$; only the sign is consumed. At the displayed interior point, the stationarity residuals are $\leq 3.5 \times 10^{-49}$ and $\det H = +7.96 \times 10^{11} > 0$, so it is a genuine local minimum, all in 50-digit arithmetic.

The global step over the single-ridge edge is certified in directed-rounding *interval* arithmetic (`certificates/l05e_iv_cert.py`; `mpmath.iv`, 60 digits). With τ_{w}, Γ_{w}, and all channel data rigorously enclosed, $\min_{t\in[0,\infty]} R_1^{\mathrm{pop}}(t) \geq -0.958054197449903009935$ holds by certified interval bounds on an exact cell decomposition of $[0, 2]$, a second-order bound on the cell containing the minimizer ($R_1'' > 0$ certified there), and a termwise-monotone tail bound for $t \geq 2$ — the ridgeless endpoint $t = 0$ and the limit $t \to \infty$ included.

Since the certificate point is an exact rational, with hops $t_1 \geq t_2 > 0$ certified, and satisfies $\psi(e_1, e_2) \leq -0.958054197670840174223$, the certified margin is $\min_t R_1^{\mathrm{pop}} - \psi(e_1, e_2) \geq 2.20937164288 \times 10^{-10} > 0$. Thus the two-hop point beats *every* single ridge, not merely a grid.

The same run certifies that the single optimum t^* is the *unique* global minimizer of R_1^{pop} (enclosed to width 2.5×10^{-27}), with criterion value $D_0(t^*) = 0.3667619816676915\ldots > 0$. The 50-digit computation was independently re-implemented from scratch with every printed digit matching.

Remark J.18 (The fold: numerical phenomenology, not theorem). What governs the counterexample is the *order* of the phase transition at $b^*(a; d)$; here we report numerics, explicitly not at certificate strength, and every value in this remark is computed at $d = 4000$ (the crossing itself moves with d; Remark D.13). Along the envelope branch, $D'(0) = \det H(t^*, 0)/(2A_2)$; at the threshold $b = b^*(a; d)$ the sign of $\det H(t^*, 0)$ flips from positive ($a \in \{0.8, 1.2, 1.5\}$: $+6.3\times10^{7}$, $+7.5\times10^{9}$, $+1.2 \times 10^{10}$) to negative ($a = 2.0$: -2.0×10^{12}).

In the *supercritical* (second-order) regime the interior minimum detaches from $(t^*, 0)$ exactly at b^*, and the local criterion *is* the global phase boundary — verified at $a = 1.2$, where the interior dip appears within $\pm1\%$ of b^*, onset exactly at the criterion sign change.

In the *subcritical* regime a saddle–minimum pair is born by a fold *before* the criterion crosses (a first-order transition), and the true population boundary $b^{**}(a)$ sits strictly below $b^*(a; d)$: bisection places the tricritical point at $a^{**} \approx 1.54$ (bracket $(1.539, 1.543)$; two independent implementations agree), and at $a = 2.0$, $b^{**}(2.0) \approx 0.06746$ against $b^*(2.0; 4000) = 0.067606$ — a sliver of relative width $(b^* - b^{**})/b^* \approx 0.22\%$, inside which criterion transversal derivatives D_0 reach $+0.49$ (criterion $= (t^*/2)D_0$, same sign) while the chain's win is 10^{-11}–5×10^{-10} deep.

These numbers are configuration-specific (the fold-vs-transcritical *mechanism* was robust in every probed configuration); we present the fold structure at numeric grade, and only the counterexample of Theorem D.12 at certificate grade.

J.5 The power-law threshold

Write $M_k = \sum_{i\geq k} (F_i^*)^2 \Delta_i$ ($M_1 = 0$ by (J.9)). Telescoping with $1/\lambda_i$ increasing in i,

$$\sum_i (F_i^*)^2 \Delta_i/\lambda_i = \sum_{k\geq 2} \big(1/\lambda_k - 1/\lambda_{k-1}\big) M_k, \tag{J.15}$$

so a single sign change $+ \to -$ in Δ (positive head, negative tail) makes every $M_k < 0$, hence (J.15) < 0: the chain wins.

At $b = 0$ (isotropic prior) the tail carries as much per-coordinate signal as the head, so the single ridge already *over*-filters the tail; the residual has the reversed $- \to +$ structure and the criterion *fails*

— a theorem for every overparametrized teacher (Theorem D.11 (i)); numerically $\partial R_2/\partial t_2|_{(t^*,0)} = +0.07\ldots+4.8 > 0$, the two-hop optimizer returning $t_2 = 0$. The criterion's sign is continuous in b, provably > 0 at $b = 0$ and provably < 0 from $b = a$ on; numerically the crossing is unique for each a, and *at* $d = 4000$ (with $\psi = 0.3$, $\bar{\kappa_w} = 3$, $\sigma^2 = 1/4$, $r^2 = 1$) it satisfies $b^*(a; 4000) \in (0, \frac{1}{2})$:

a	0.8	1.2	1.5	2.0	2.5	3.0
$b^*(a; 4000)$	0.253	0.286	0.202	0.068	0.013	0.002

At this d the crossing is *non-monotone in* a — $b^*(0.8; 4000) = 0.253 < b^*(1.2; 4000) = 0.286$, rising to a peak ≈ 0.30 near $a \approx 1.1$ before decaying to 0.002 by $a = 3$.

The whole row moves with d, and downward: see the d-table of Remark D.13, where the same four capacity exponents lose between a factor 1.4 and a factor 87 in b^* across $d = 10^3$ to $3.2 \cdot 10^4$. (The $a = 0.8$ column sits outside the standing assumption $a > 1$ and is shown only to exhibit the non-monotonicity of the crossing; every theorem in this paper requires $a > 1$.) The status ledger: both ends of the axis are theorems ($b = 0$: no strict win and criterion > 0, Theorem D.11 (i); $b \geq a$: strict win, part (ii)); the crossing values in the table are computed numerically and mark the *local*-criterion crossing, which is the exact global phase boundary in the second-order regime $a \lesssim 1.54$ and an over-estimate beyond, by the hairline sliver of Remark J.18.

What is at certificate strength there is the sliver's *non-emptiness* — Theorem D.12 exhibits a point inside it, at $(a, b) = (2, 0.0675)$, where the local criterion is strictly positive yet a two-hop chain provably beats every single ridge — and not its width, which Remark J.18 reports at numeric grade.

J.6 Finite budgets: the phase transition

At total budget N, compare $R_1^*(N) = \min_\lambda R_1^{\text{fin}}(N, \lambda)$ (single hop, all N) with $R_2^*(N) = \min_{f,\lambda_1,\lambda_2} R_2^{\text{fin}}$ (split $n_1 = fN$, $n_2 = (1-f)N$), both from the recursion of Theorem D.9.

Division of labor (what the certificate certifies). Both quantities above are *defined* as values of that deterministic recursion, and the certificate below is an interval-arithmetic statement about those values. This distinction matters here because the instance is deliberately outside the transcribed convergence engines: its spectrum is power law with $a = 1.2$, so Assumption C.1's bounded-condition-number clause fails, and both hops carry strictly positive ridges. At that instance the recursion is read at limit-formula level throughout (Remarks J.4 and K.16). So Theorem D.10 (ii) is a rigorous statement about the limit formulas at a fully specified instance — which is what the phase-transition claim needs, since the phase boundary is a property of those formulas — and not a finite-sample probabilistic statement; the paired simulation quoted with it is corroboration, not part of the certificate.

Remark J.19 (Which recursion the certificate computes: the annealed one). Theorem D.9 is stated in the *quenched* convention: the per-hop fluctuation is computed at the realized profile $\rho_i^{(t)} = (\alpha_i^{(t)})^2 p_i + \nu_i^{(t)}$ with $p_i = (w_i^*)^2$, and replacing p_i by ω_i^2 throughout gives the *annealed* recursion. The certificate `certificates/chainA_iv_cert.py` propagates the annealed one: every occurrence of the signal in it is the deterministic $\omega_i^2 = Gi^{-b}$, so the enclosed quantities — $\rho_i = (F_i^{\text{w}})^2\omega_i^2 + v_i^{\text{w}}$, the Wiener data $c_i, W_i, \Gamma_{\text{w}}$ and every risk built from them — are annealed.

So is the population objective (J.7) and every population statement of Sections J.3 and J.4, which are written in ω^2 from the start: the chain analysis of this paper lives in the annealed calculus, and the certificate is consistent with it.

This has to be said out loud, because the sufficient condition for the two readings to agree is unavailable here. They agree under the prior moment condition (G.9), and Remark G.4 (b) shows that condition *fails* along the power law — for the Gaussian prior itself, $V_d \to 2G^2\zeta(2a + 2b) > 0$ by (G.10). Instance A is a power-law instance ($a = 1.2$, $b = 2$), so Theorem D.10 (ii) is to be read as an interval-arithmetic certificate for the *annealed* limit formulas at a fully specified instance: not a claim about the quenched recursion there, and not a finite-sample probabilistic claim. Nothing else in the chain section turns on the distinction, since nothing else in it is quenched.

(i) The splitting cost vanishes. The finite risk of a hop equals its population risk plus the self-noise floor

$$\sum_i \lambda_i v_i^{\mathrm{self}} = \Gamma \sum_i \lambda_i^2/(\lambda_i+\tau)^2 = \Gamma\, n\Xi = \mathrm{num}\cdot\frac{\Xi}{1-\Xi}, \quad \mathrm{num} = \tau^2 \sum_j \lambda_j \rho_j/(\lambda_j+\tau)^2 \geq 0. \tag{J.16}$$

Since $v_i^{\mathrm{self}} \geq 0$, the population value is a lower bound: $R_1^{\mathrm{fin}}(N,\lambda) \geq R_1^{\mathrm{pop}}(\tau(\lambda)) \geq R_1^{\mathrm{pop},*}$, and likewise for R_2.

Proposition J.20 (The splitting cost: an explicit envelope, and when it vanishes)**.** *For a stage with spectrum (λ_j), budget n and induced ridge τ from (C.5), write*

$$\mathrm{ed}(\tau) \;:=\; \sum_j \frac{\lambda_j^2}{(\lambda_j+\tau)^2}, \qquad \textit{so that} \qquad \Xi = \mathrm{ed}(\tau)/n$$

by (C.6), and set $\bar{E} := r^2 + \sum_j \lambda_j v_j^{\mathrm{w}}$, a teacher-determined constant. Fix $\varepsilon > 0$ and ridges t^ε and $(t_1^\varepsilon, t_2^\varepsilon)$ that are ε-optimal for the two population problems and admissible *at the budgets they are used at, i.e. $t^\varepsilon \geq \tau^{\mathrm{MNI}}(N)$ and $t_1^\varepsilon, t_2^\varepsilon \geq \tau^{\mathrm{MNI}}(N/2)$.*
Write

$$\mathfrak{T}^{(1)} := \bar{E}\,\frac{\Xi^{(1)}}{1-\Xi^{(1)}}, \qquad \mathfrak{t}_1 := \bar{E}\,\frac{\Xi_1}{1-\Xi_1}, \qquad \mathfrak{t}_2 := (\bar{E}+\mathfrak{t}_1)\frac{\Xi_2}{1-\Xi_2}, \qquad \mathfrak{T}^{\mathrm{ch}} := \mathfrak{t}_1+\mathfrak{t}_2,$$

with $\Xi^{(1)} := \mathrm{ed}(t^\varepsilon)/N$ and $\Xi_j := \mathrm{ed}(t_j^\varepsilon)/(N/2)$, all assumed < 1. Then at every budget N

$$0 \leq R_1^*(N) - R_1^{\mathrm{pop},*} \leq \varepsilon + \mathfrak{T}^{(1)}, \qquad 0 \leq R_2^*(N) - R_2^{\mathrm{pop},*} \leq \varepsilon + \mathfrak{T}^{\mathrm{ch}}. \tag{J.17}$$

Consequently, along any family on which

$$\mathrm{ed}(\tau) \;=\; o(N) \qquad \textit{at those ridges} \tag{J.18}$$

the envelope vanishes and $R_1^*(N) \to R_1^{\mathrm{pop},*}$, $R_2^*(N) \to R_2^{\mathrm{pop},*}$, so $R_2^*(N) - R_1^*(N) \to R_2^{\mathrm{pop},*} - R_1^{\mathrm{pop},*}$ — the population gap, which is ≤ 0 always (Lemma J.5) and < 0 when the criterion holds. (In a family whose population problem itself moves with N that last sentence is true but empty; Remark J.21 says what survives and what does not.) Two clean realizations of (J.18):

- (a) Rigorous, classical. *Underparametrized ridge student, d fixed, $N \to \infty$, $\lambda > 0$ fixed. Here $\kappa_{\mathrm{s}} = d/N < 1$ is* outside *the proportional overparameterized regime of Assumption C.1, so this leg rests not on Theorem D.1 but on classical ridge consistency: $\widetilde{X}^\top \widetilde{X}/N \to \Sigma$ a.s., hence*

  $$\widehat{w}_{\mathrm{s}} = (\widetilde{X}^\top\widetilde{X}/N + \lambda I)^{-1}(\widetilde{X}^\top \widehat{y}/N) \to (\Sigma+\lambda I)^{-1}\Sigma\beta_{\mathrm{w}} = F(\lambda)\beta_{\mathrm{w}}$$

  a.s.; and $\mathrm{ed}(\tau) \leq d = O(1) = o(N)$.
- (b) Power-law, at ridges above the interpolation scale. $\kappa_{\mathrm{s}} = d/N > 1$ *fixed, $d \to \infty$, spectrum $\lambda_j = j^{-a}$ with $a > 1$.*
  Comparison with $\int_0^\infty (1+\tau x^a)^{-2} dx$ gives

  $$\mathrm{ed}(\tau) = \sum_{j\leq d}(1+\tau j^a)^{-2} \;\leq\; 1 + J_a\,\tau^{-1/a}, \qquad J_a := \int_0^\infty \frac{du}{(1+u^a)^2} < \infty, \tag{J.19}$$

  so (J.18) holds as soon as

  $$\tau N^a \;\longrightarrow\; \infty \qquad \Longleftrightarrow \qquad \tau/\tau^{\mathrm{MNI}}(N) \;\longrightarrow\; \infty, \tag{J.20}$$

  the equivalence because $\tau^{\mathrm{MNI}}(N) \asymp N^{-a}$ (Lemma J.23 below, which is elementary and independent of this proposition); and this same condition is what makes τ admissible with room to spare. This — not "ridges bounded away from 0" — is the correct hypothesis, and the distinction is not pedantic: *the ridges this paper actually tunes to satisfy $t^* \to 0$, so a bounded-away-from-0 hypothesis is never available at them, whereas (J.20) holds there by a polynomial margin (Proposition J.25 (ii)–(iii), every $b \geq 0$).*

Hypothesis (J.18) is not decorative: Remark J.22 exhibits a stage where the splitting cost is bounded below *by a positive constant, uniformly in N, and identifies the axis — the ridge scale — that actually decides it.*

Proof. A two-sided sandwich; no uniform convergence over ridge families is used.

Lower. By (J.16) every hop's self-noise is nonnegative termwise, so at any budget and any ridges $R_1^{\mathrm{fin}} \geq R_1^{\mathrm{pop}}(\tau) \geq R_1^{\mathrm{pop},*}$ and $R_2^{\mathrm{fin}} \geq R_2^{\mathrm{pop}}(t_1,t_2) \geq R_2^{\mathrm{pop},*}$; minimizing, $R_k^*(N) \geq R_k^{\mathrm{pop},*}$ for $k = 1, 2$.

Upper. The near-minimizers satisfy $R_1^{\mathrm{pop}}(t^\varepsilon) \leq R_1^{\mathrm{pop},*} + \varepsilon$ and $R_2^{\mathrm{pop}}(t_1^\varepsilon, t_2^\varepsilon) \leq R_2^{\mathrm{pop},*} + \varepsilon$ by hypothesis; it remains to check that they are *reachable*. Differentiating (C.5) in τ,

$$\frac{d\lambda}{d\tau} \;=\; 1 - \frac{1}{n}\sum_i \frac{c_i}{c_i+\tau} + \frac{\tau}{n}\sum_i \frac{c_i}{(c_i+\tau)^2} \;=\; 1 - \frac{1}{n}\sum_i \frac{c_i^2}{(c_i+\tau)^2} \;=\; 1 - \Xi(\tau) \;>\; 0$$

— note this is the derivative of $\tau \mapsto \lambda(\tau)$, so the inverse map $\lambda \mapsto \tau(\lambda)$ has derivative $(1-\Xi)^{-1}$. Hence $\tau \mapsto \lambda(\tau)$ is a strictly increasing continuous bijection from $[\tau^{\mathrm{MNI}}(n), \infty)$ onto $[0, \infty)$, and a target τ is realized by an admissible ridge $\lambda \geq 0$ *if and only if* $\tau \geq \tau^{\mathrm{MNI}}(n)$ — which is exactly the admissibility hypothesis placed on t^ε and $(t_1^\varepsilon, t_2^\varepsilon)$. (Along a family with $\tau^{\mathrm{MNI}}(n) \to 0$ any fixed positive target is eventually admissible, which covers realization (a); in the power-law family the tuned ridges are admissible by (J.20), and this is the only place a lower bound on the ridges is used.) Evaluate the chain at those ridges with the even split $n_1 = n_2 = N/2$ (the population risk does not depend on the split).

Writing $H^{(t)}$ for the composed filter after t hops, $v^{(t)}$ for the accumulated fluctuation profile, $E_t := \sum_j \lambda_j \rho_j^{(t)}$ for the energy entering hop $t{+}1$ and $\mathrm{toll}_t := \sum_i \lambda_i v_i^{\mathrm{self},t}$, the recursion of Theorem D.9 gives exactly

$$R_1^{\mathrm{fin}} = R_1^{\mathrm{pop}} + \mathrm{toll}_1, \qquad R_2^{\mathrm{fin}} = R_2^{\mathrm{pop}} + \sum_i \lambda_i (F_i^{(2)})^2 v_i^{\mathrm{self},1} + \mathrm{toll}_2.$$

By (J.16), $\mathrm{toll}_t = \mathrm{num}_t\, \Xi_t/(1-\Xi_t)$ with $\mathrm{num}_t = \tau_t^2 \sum_j \lambda_j \rho_j^{(t-1)}/(\lambda_j+\tau_t)^2 \leq E_{t-1}$, and $\Xi_t = \mathrm{ed}(\tau_t)/(N/2)$.

The energies stay bounded: $E_0 \leq r^2 + \sum_j \lambda_j v_j^{\mathrm{w}} = \bar{E}$ (using $H_j^{(0)} \leq 1$), and $E_1 \leq r^2 + \sum_j \lambda_j v_j^{(1)} \leq \bar{E} + \mathrm{toll}_1$ since $(F_j^{(1)})^2 \leq 1$. Since $u \mapsto u/(1-u)$ is increasing on $[0,1)$, this gives $\mathrm{toll}_1 \leq \bar{E}\,\Xi_1/(1-\Xi_1) = \mathfrak{t}_1$, then $\mathrm{toll}_2 \leq (\bar{E}+\mathfrak{t}_1)\Xi_2/(1-\Xi_2) = \mathfrak{t}_2$, and $\sum_i \lambda_i (F_i^{(2)})^2 v_i^{\mathrm{self},1} \leq \mathrm{toll}_1 \leq \mathfrak{t}_1$; the same computation for a single hop on the full budget N gives $\mathrm{toll} \leq \mathfrak{T}^{(1)}$.

Hence $R_1^*(N) \leq R_1^{\mathrm{fin}}(N, t^\varepsilon) \leq R_1^{\mathrm{pop},*} + \varepsilon + \mathfrak{T}^{(1)}$ and $R_2^*(N) \leq R_2^{\mathrm{pop},*} + \varepsilon + \mathfrak{t}_1 + \mathfrak{t}_2$, which is (J.17): no limit has been taken and no uniformity over ridge families is used. Under (J.18) every Ξ above tends to 0, so $\mathfrak{T}^{(1)}, \mathfrak{T}^{\mathrm{ch}} \to 0$; letting $\varepsilon \downarrow 0$ after $N \to \infty$ gives $\limsup_N R_k^*(N) \leq R_k^{\mathrm{pop},*}$. Combining with the lower half and subtracting gives the claim. □

Remark J.21 (Absolute versus relative: what (J.17) gives in a family whose population problem moves). (J.17) holds at each fixed budget, so it applies verbatim to a family in which the teacher, the spectrum and hence the population optima all move with N — the power-law family, where $d = \bar{\kappa}_{\mathrm{s}} N$. There its two readings part company.

Absolute: correct, and empty. Along the power law at the tuned ridges (J.18) does hold, so $R_2^*(N) - R_1^*(N) \to R_2^{\mathrm{pop},*} - R_1^{\mathrm{pop},*}$. But both sides tend to 0 — $R_1^{\mathrm{pop},*} \to 0$ by Proposition J.25 (ii), and $0 \geq R_2^{\mathrm{pop},*} - R_1^{\mathrm{pop},*} \geq -R_1^{\mathrm{pop},*}$ — so the limit "$0 \to 0$" controls no sign. In a moving family the absolute conclusion of Proposition J.20 is true and carries no information.

The order of quantifiers, first. Proposition J.20 fixes $\varepsilon > 0$ *before* the budget and bounds the excess by $\varepsilon + \mathfrak{T}^{(1)}$. Along a moving family that alone is fatal to any relative reading: with $P := R_1^{\mathrm{pop},*} \to 0$, a fixed $\varepsilon > 0$ already gives $E \geq \varepsilon$ and hence $E/P \to \infty$, so (J.21) below would say nothing whatever. The envelope is available at *each* budget and for *every* $\varepsilon > 0$, so the legal repair — and the reading intended throughout this remark — is to choose $\varepsilon = \varepsilon_N \downarrow 0$ *after* N.

For the one-hop term one may take $\varepsilon_N = 0$ outright, since Proposition J.25 (ii) shows the one-hop population infimum is attained at an admissible ridge t^*; for the two-hop term no such attainment is established here, so a vanishing sequence is the honest device. The arithmetic below is at $\varepsilon_N = 0$ in the one-hop term and $\varepsilon_N \downarrow 0$ in the two-hop term, and none of it survives at a fixed ε.

Relative: the envelope pins it, but only to within the toll fraction. Write $P := R_1^{\mathrm{pop},*}$, $G := R_2^{\mathrm{pop},*} - P \in [-P, 0]$, $A := R_2^*(N) - R_1^*(N)$ and $E := \varepsilon_N + \mathfrak{T}^{(1)} + \mathfrak{T}^{\mathrm{ch}}$. Then (J.17) gives $|A - G| \le E$ and $R_1^*(N) \in [P, P + E]$, whence

$$\left| \frac{A}{R_1^*(N)} - \frac{G}{P} \right| = \frac{|AP - GR_1^*(N)|}{R_1^*(N)\,P} \le \frac{P|A - G| + |G|\,|P - R_1^*(N)|}{P^2} \le \frac{2E}{P}, \qquad \text{(J.21)}$$

using $|G| \le P$. Two consequences, and a limit.

(1) The sign certificate, and why it cannot fire in this family. If $|G| > E$ then $A < 0$: the chain still wins at the finite budget. In a moving power-law family that comparison is unavailable, and structurally so rather than by slack in the estimates. On one side $G \in [-P, 0]$ forces $|G|/P \le 1$. On the other, $E \ge \mathfrak{T}^{(1)}$ and $\mathfrak{T}^{(1)} \asymp P$: by Proposition J.25 (i) the constant $\bar{E} = r^2 + \Gamma_{\mathrm{w}} n_{\mathrm{w}} \Xi_{\mathrm{w}}$ is $\asymp 1$, and by (iii)–(v) $\Xi^{(1)} = \mathrm{ed}(t^*)/N \asymp \mathrm{toll}(t^*) \asymp P$; so E/P is bounded below by a positive constant while $|G|/P$ cannot exceed 1.

Measured at the three instances below, $E/P = 4.140,\ 3.492,\ 3.197$ against $|G|/P = 0.091,\ 0.120,\ 0.145$, i.e. $|G|/E = 0.022,\ 0.034,\ 0.045$ — the certificate needs $|G|/E > 1$, and the ratio is nowhere near a trajectory to 1. The honest reading is that the sign certificate is usable in the *fixed*-population realization (a) of Proposition J.20 and not in a moving family; what settles the moving case at a finite budget is the interval certificate of Theorem D.10 (ii), which does not pass through (J.17) at all. (Two of the three orders here — $\mathfrak{T}^{\mathrm{ch}} \asymp P$ and $|G| \asymp P$ — are numerical observations and not theorems: Proposition J.25 analyses the *one-hop* toll at the *one-hop* tuned ridge, whereas $\mathfrak{T}^{\mathrm{ch}}$ lives at two ridges on the split budget $N/2$, and G does not occur in that proposition at all.)

(2) What is *not* proved. The relative convergence

$$\frac{R_2^*(N) - R_1^*(N)}{R_1^*(N)} \longrightarrow \frac{R_2^{\mathrm{pop},*} - R_1^{\mathrm{pop},*}}{R_1^{\mathrm{pop},*}}$$

does *not* follow from (J.21), and we do not prove it. The obstruction is structural, not a slack in the estimates: by Proposition J.25 (i) and (v) the toll is a fixed fraction of the risk in this family, so $E/P \not\to 0$ and the right-hand side of (J.21) stays bounded away from 0. Establishing the relative limit would require showing that the single-hop and the two-hop tolls agree to leading order — a cancellation between two pipelines with different splits and different ridges — which the sandwich cannot see, since it bounds each toll separately.

Numerically the cancellation does appear to occur. At $(a, b) = (1.2, 2)$, $\psi = 0.3$, $\bar{\kappa_{\mathrm{w}}} = 3$, $d/N = 2$, the finite-budget and the population relative gaps — $(R_2^*(N) - R_1^*(N))/R_1^*(N)$ and G/P — are

$$(-6.91\%,\ -9.09\%) \text{ at } d = 10^3, \qquad (-10.39\%,\ -12.04\%) \text{ at } d = 4 \cdot 10^3,$$
$$(-13.45\%,\ -14.50\%) \text{ at } d = 3.2 \cdot 10^4.$$

Note that *both* move with d; what shrinks is their discrepancy, $2.18 \to 1.65 \to 1.05$ percentage points.

What (J.21) charges for is the quantity $2E/P$, and that does not shrink to anything usable: it reads $8.28,\ 6.98,\ 6.39$ at the same three instances — sixty to a hundred and thirty times the discrepancy it would have to control, decreasing far too slowly to be on its way to 0. (These are envelope quantities, not realized tolls, and the gap is where one would expect it: $\mathfrak{T}^{(1)}$ carries $\bar{E} = 1.228,\ 1.229,\ 1.229$ where the realized toll numerator carries $0.155,\ 0.188,\ 0.214$, and $\mathfrak{T}^{\mathrm{ch}}$ is a two-hop bound on the split budget $N/2$.

The *realized* toll fraction $\mathrm{toll}(t^*)/R_1^{\mathrm{pop},*}$ of Remark J.26 is the much smaller $0.054 \to 0.057 \to 0.061$; that is the fraction *(v)* speaks about, and it is not the quantity (J.21) charges.) So the two tolls are cancelling to an accuracy the envelope cannot certify. We record that as a numerical observation and not as a theorem. (The two-hop recursion used for these figures reproduces the interval certificate of Theorem D.10 (ii) at $d = 4000$ to eight digits, $R_1^{\mathrm{pop},*} = 0.01721745$ against the certified ≥ 0.0172174469 and $R_2^*(N) = 0.01630177$ against the certified ≤ 0.0163017768.) The load-bearing finite-budget claim of this paper, Theorem D.10 (ii), is a certificate at a fixed instance and does not depend on any of this.

Remark J.22 ((J.18) is necessary, and what decides it is the *ridge scale*). Hypothesis (J.18) is a joint statement about the spectrum and the scale of the ridge, not a convenience of the proof, and it fails for the flattest spectrum of all. Take $\Sigma = I_d$, one noiseless minimum-norm hop on $n = N$ fresh inputs with input profile u, and $\alpha := N/d \in (0, 1)$ held fixed as $d \to \infty$.

Then (C.5) reads $d/(1+\tau) = N$, so $\tau = (1-\alpha)/\alpha$, $F_i \equiv \alpha$, $\Xi = \alpha$ and $\mathrm{ed}(\tau) = d\alpha^2 = \alpha N$: of order N, not $o(N)$. By (J.16) the toll is then

$$\sum_i \lambda_i v_i^{\mathrm{self}} \;=\; \frac{\Xi}{1-\Xi}\,\tau^2 \sum_j \frac{\lambda_j u_j^2}{(\lambda_j+\tau)^2} \;=\; \frac{\alpha}{1-\alpha}\,(1-\alpha)^2\|u\|^2 \;=\; \alpha(1-\alpha)\,\|u\|^2,$$

independent of N and strictly positive: the splitting cost does *not* vanish, at any budget. (Consistency: the hop's population risk is $(1-\alpha)^2\|u\|^2$, and population plus toll is $(1-\alpha)\|u\|^2$ — exactly $\mathbb{E}\|(I-\widetilde{P})u\|^2$ for a projection onto a uniformly random N-dimensional subspace, which is what a noiseless minimum-norm hop computes. The channel formula is right; it is the vanishing of the toll that is false here.)

The axis is the ridge scale, not the shape of the spectrum. It is tempting to read the contrast above as "power law good, bounded spectrum bad". That is not what (J.18) measures. For *every* spectrum $\mathrm{ed}(\tau)/n$ is exactly Ξ, and at an interpolating ridge Cauchy–Schwarz on $\sum_i F_i = n$ gives $\Xi \geq 1/\kappa$ (Lemma G.1), so $\mathrm{ed}(\tau^{\mathrm{MNI}}(n)) \geq n/\kappa$ is of order n and never $o(n)$ — power law included. The power law is no exception: at $a = 1.2$, $d/N = 2$, the student stage's own MNI ridge gives $\mathrm{ed}(\tau^{\mathrm{MNI}}(N))/N = 0.58729,\ 0.58776,\ 0.58777$ at $d = 500,\ 16000,\ 256000$ — flat in d, and nowhere near 0.

What differs between the two spectra is whether *any useful* ridge escapes that scale. Along the power law there is a whole window: by (J.19) every τ with $\tau N^a \to \infty$ has $\mathrm{ed}(\tau) = o(N)$, while $\tau \to 0$ keeps the filter nondegenerate, and the tuned ridge sits strictly inside the window with a polynomial margin (Proposition J.25 (ii)–(iii), every $b \geq 0$). For a spectrum bounded above and below, $\lambda_j \in [1/M, M]$ with $d \asymp N$, the window is *empty*: since $\lambda \mapsto \lambda^2/(\lambda+\tau)^2$ is increasing,

$$\mathrm{ed}(\tau) \;=\; \sum_{j \leq d} \frac{\lambda_j^2}{(\lambda_j+\tau)^2} \;\geq\; \frac{d}{(1+M\tau)^2},$$

so $\mathrm{ed}(\tau) = o(N)$ forces $\tau \to \infty$, and then $F_j \leq M/(M+\tau) \to 0$ uniformly: the only ridges that avoid the toll are those that shrink the estimator to $\mathbf{0}$. The $\Sigma = I$ computation above is the extreme instance, read at the interpolating ridge.

So the correct scope clause is about the ridge scale: the splitting cost vanishes on ridges with $\tau/\tau^{\mathrm{MNI}}(N) \to \infty$ and does not on ridges of order $\tau^{\mathrm{MNI}}(N)$; what a power-law spectrum buys is that the first set contains *useful* ridges. Every chain statement in this paper that is made *at a budget* — Theorem D.10 (i)'s splitting-cost clause and Theorem D.10 (ii)'s certificate at $a = 1.2$ — is made at a power-law spectrum and at ridges in that set, licensed by Proposition J.25 (ii)–(iii). (Theorem D.11 needs no such clause: it is a population statement at $N = \infty$, where no toll is paid, and its part (ii) covers all $b \geq a$, the saturation boundary included.) So this is a scope clause rather than a gap — but one that has to be carried explicitly, and Theorem D.10 (i) carries it. What it does *not* buy is a splitting cost that is negligible *relative* to the risk: see Proposition J.25 (i) and (iv)–(v), and Remark J.21.

The splitting cost along the power law: exact order

The realizations above say *when* the splitting cost vanishes. Along the power law it does vanish — and at a rate that is neither $O(1/N)$ nor fast enough to be negligible against the risk it is charged against. Two elementary estimates, then the statement.

Lemma J.23 (Power-law fixed point). *Let $\lambda_j = j^{-a}$ with $a > 1$, let $1 \leq n < p$ be integers with $n \geq \zeta(a)$, and let $\tau > 0$ solve $\sum_{j \leq p} \lambda_j/(\lambda_j+\tau) = n$. Then $\tau \leq 1$ and*

$$\frac{(a+1)(p-n)}{(p+1)^{a+1}} \;\leq\; \tau \;\leq\; \left(1 + \tfrac{2^{a-1}}{a-1}\right)^a n^{-a}. \tag{J.22}$$

In particular, if $p/n \to \bar{\kappa} \in (1, \infty)$ then $\tau \asymp p^{-a} \asymp n^{-a}$, with constants depending only on $(a, \bar{\kappa})$.

Proof. If $\tau > 1$ then $n = \sum_{j\le p} \lambda_j/(\lambda_j+\tau) < \tau^{-1}\sum_{j\ge 1} j^{-a} < \zeta(a)$, contrary to hypothesis; so $\tau \le 1$ and $J := \lfloor \tau^{-1/a} \rfloor \ge 1$.

Upper bound. Bound $\lambda_j/(\lambda_j+\tau)$ by 1 for $j \le J$ and by λ_j/τ for $j > J$. Since $x \mapsto x^{-a}$ decreases, $\sum_{j>J} j^{-a} \le \int_J^\infty x^{-a}\,dx = J^{1-a}/(a-1)$, and $J \ge \tau^{-1/a}/2$ (if $\tau^{-1/a} \ge 2$ then $J > \tau^{-1/a} - 1 \ge \tau^{-1/a}/2$; otherwise $J = 1 \ge \tau^{-1/a}/2$), so $J^{1-a} \le 2^{a-1}\tau^{(a-1)/a}$. Hence

$$n \;\le\; J + \frac{1}{\tau}\cdot\frac{2^{a-1}\tau^{(a-1)/a}}{a-1} \;\le\; \tau^{-1/a}\Big(1+\frac{2^{a-1}}{a-1}\Big),$$

which rearranges to the stated upper bound.

Lower bound.

$$p - n = \sum_{j\le p} \tau/(\lambda_j+\tau) \le \tau\sum_{j\le p}\lambda_j^{-1} = \tau\sum_{j\le p} j^a \le \tau\int_1^{p+1} x^a dx \le \tau(p+1)^{a+1}/(a+1).$$

□

Lemma J.24 (Two-sided power-law sums)**.** *Let $a > 1$, $t \in (0,1]$, $p \ge 1$ an integer, and $j_t := t^{-1/a} \ge 1$.*

(a) $\frac{1}{8}\min(p, j_t) \;\le\; \sum_{j\le p} \frac{1}{(1+tj^a)^2} \;\le\; \min\big(p,\; 1 + J_a\, t^{-1/a}\big)$, *where* $J_a := \int_0^\infty (1+u^a)^{-2}du < \infty$.

(b) *Let $s > 1$ and suppose $4j_t \le p$. Then, with constants depending only on (a, s),*

$$\Sigma_s(t) \;:=\; \sum_{j\le p} j^{-s}\Big(\frac{tj^a}{1+tj^a}\Big)^2 \;\asymp\; \begin{cases} t^{(s-1)/a}, & s < 2a+1,\\ t^2\log(2/t), & s = 2a+1,\\ t^2, & s > 2a+1, \end{cases}$$

i.e. $\Sigma_s(t) \asymp t^{\min\{2,\,(s-1)/a\}}$ away from the boundary case $s = 2a+1$, where the same holds up to the logarithm. (Above the boundary the exponent saturates at 2 and cannot be $(s-1)/a$: the $j = 1$ term alone is already $\asymp t^2$.) Moreover $\sum_{j>p} j^{-s} \asymp p^{1-s}$.

Proof. Throughout we use the elementary comparison, valid for $\alpha \ne 1$ and $J \ge J_0(\alpha)$,

$$\sum_{j\le J} j^{-\alpha} \;\asymp_\alpha\; \max\{1,\, J^{1-\alpha}\}, \qquad \sum_{j\le J} j^{-1} \;\asymp\; \log(2J), \qquad \sum_{j>J} j^{-\alpha} \;\asymp_\alpha\; J^{1-\alpha}\ (\alpha > 1), \tag{J.23}$$

each obtained by comparing with $\int x^{-\alpha}dx$ over the same range: for $\alpha \le 0$ the summand increases and one compares on $(J/2, J]$, for $\alpha > 0$ it decreases and one compares on $[1, J+1]$.

(a) Each summand is ≤ 1, giving the bound p; and since $x \mapsto (1+tx^a)^{-2}$ decreases, $\sum_{j\ge1}(1+tj^a)^{-2} \le 1 + \int_0^\infty (1+tx^a)^{-2}dx = 1 + J_a t^{-1/a}$ by $u = t^{1/a}x$, with $J_a < \infty$ because $2a > 1$. For the lower bound, every $j \le \min(p, j_t)$ has $tj^a \le 1$, hence $(1+tj^a)^{-2} \ge 1/4$; there are $\lfloor \min(p, j_t)\rfloor \ge \frac{1}{2}\min(p, j_t)$ such j, as $\min(p, j_t) \ge 1$.

(b) Split at j_t. For $j > j_t$ we have $tj^a > 1$, so the bracket lies in $(\frac{1}{2}, 1)$ and that part of Σ_s is $\asymp \sum_{j_t<j\le p} j^{-s}$; by (J.23) the upper bound is $\le C j_t^{1-s}$, and for the lower bound $\sum_{j_t<j\le 2j_t} j^{-s} \ge \lfloor j_t\rfloor (2j_t)^{-s} \ge 2^{-s-1} j_t^{1-s}$, legitimate because $2j_t \le p$. So this part is $\asymp j_t^{1-s} = t^{(s-1)/a}$.

For $j \le j_t$ we have $tj^a \le 1$, so the bracket lies in $[\frac{1}{4}(tj^a)^2, (tj^a)^2]$ and that part is $\asymp t^2\sum_{j\le j_t} j^{2a-s}$. Put $\alpha := s - 2a$. By (J.23): if $\alpha < 1$, equivalently $(s-1)/a < 2$, this is $\asymp t^2 j_t^{1-\alpha} = t^{(s-1)/a}$; if $\alpha > 1$ it is $\asymp t^2$; if $\alpha = 1$ it is $\asymp t^2\log(2j_t) \asymp t^2\log(2/t)$. Adding the two parts and using $t \le 1$ gives the claim; the final assertion is the third estimate in (J.23). □

Proposition J.25 (The power-law splitting cost: exact order, and a fixed fraction of the risk)**.** *Take the power-law family (C.2) with $a > 1$, $b \geq 0$, and fixed $\psi \in (0,1)$, $\bar{\kappa_{\mathrm{w}}} > 1$, $\bar{\kappa_{\mathrm{s}}} > 1$, $\sigma^2 > 0$, $r^2 > 0$: head $m = \lceil \psi d \rceil$, teacher budget $n_{\mathrm{w}} = \lceil m/\bar{\kappa_{\mathrm{w}}} \rceil$, MNI teacher, and one student hop on the total budget $N = \lceil d/\bar{\kappa_{\mathrm{s}}} \rceil$.*
Put

$$\theta := \min\Big\{2,\ \frac{a+b-1}{a}\Big\}, \qquad \varrho := \frac{a}{a\theta+1} \quad \Big(= \frac{a}{a+b} \text{ when } b \leq a+1\Big),$$

so that $\theta\varrho + \varrho/a = 1$. write Φ_{ab} be the bias profile of Lemma K.8, and write

$$\Lambda_d := \begin{cases} 1, & b \neq a+1, \\ \log d, & b = a+1, \end{cases}$$

the boundary factor. The saturation boundary $b = a+1$ is included*: it is the logarithmic case of Lemma J.24 (b) and of Φ_{ab}, and Λ_d is exactly the correction it produces.*
All $\asymp$ below have constants depending only on $(a, b, \psi, \bar{\kappa_{\mathrm{w}}}, \bar{\kappa_{\mathrm{s}}}, \sigma^2, r^2)$ and hold for $d \geq d_0$ of the same dependence.

(i) $\tau_{\mathrm{w}} \asymp d^{-a}$, $F_i^{\mathrm{w}} \asymp 1$ *uniformly over $i \leq m$*, $1 - \Xi_{\mathrm{w}} \asymp 1$, $\Gamma_{\mathrm{w}} \asymp d^{-1}$ *and* $\Gamma_{\mathrm{w}} m \asymp 1$*; consequently the constant $\bar{E} = r^2 + \Gamma_{\mathrm{w}} n_{\mathrm{w}} \Xi_{\mathrm{w}}$ of Proposition J.20 satisfies $\bar{E} \asymp 1$.*

(ii) $R_1^{\mathrm{pop}}(t) \asymp \Phi_{ab}(t) + \Gamma_{\mathrm{w}} t^{-1/a}$ *uniformly on $t \in [Km^{-a}, 1]$ for a constant $K = K(a, \bar{\kappa_{\mathrm{w}}})$, while $R_1^{\mathrm{pop}} \geq c > 0$ off that interval.*
Hence every minimizer obeys

$$t^* \asymp (d\,\Lambda_d)^{-\varrho}, \qquad R_1^{\mathrm{pop},*} \asymp d^{-\theta\varrho}\,\Lambda_d^{\varrho/a} \;=\; d^{-(1-\varrho/a)}\Lambda_d^{\varrho/a},$$

and $t^*/\tau_{\mathrm{w}} \asymp t^* N^a \asymp d^{\,a-\varrho}\Lambda_d^{-\varrho} \to \infty$.

(iii) $\mathrm{ed}(t^*) \asymp (d\,\Lambda_d)^{\varrho/a} \to \infty$*, yet* $\mathrm{ed}(t^*) = o(N)$*: (J.18) holds, at ridges tending to* 0 *— which is why realization (b) must be read through (J.20) and not through a bounded-away-from-*0 *clause.*

(iv) $\mathrm{toll}(t^*) \asymp \mathrm{ed}(t^*)/N \asymp N^{-(1-\varrho/a)}\Lambda_N^{\varrho/a}$*, so*

$$N \cdot \mathrm{toll}(t^*) \;\asymp\; (N\Lambda_N)^{\varrho/a} \;\longrightarrow\; \infty :$$

the splitting cost vanishes but is not $O(1/N)$.
For $(a,b) = (1.2, 2)$ the exponent is $1 - \varrho/a = 1 - 1/(a+b) = 11/16$ with $\Lambda_N = 1$; at the saturation boundary $b = a+1$ the same formula reads $N^{-2a/(2a+1)}(\log N)^{1/(2a+1)}$ — the boundary logarithm that Theorem D.14 (iii) carries, and for the same reason.

(v) $\mathrm{toll}(t^*) \asymp R_1^{\mathrm{pop},*}$*: the splitting cost is a fixed fraction of the risk, at every d. The proposition proves only the $\asymp$; the fraction* measured *across the instances of Remark J.26 lies in* [0.054, 0.094].

Proof. Write $\lambda_j = j^{-a}$, $\omega_j^2 = Gj^{-b}$ with $G = r^2/\sum_{j \leq d} j^{-(a+b)}$; since $a + b > 1$ the normalizing sum lies in $[1, \zeta(a+b)]$, so $G \asymp 1$. Constants c, C may change from line to line; $c_1, c_2, \ldots$ are fixed.

(i). Lemma J.23 with $(p, n) = (m, n_{\mathrm{w}})$ gives $c_1 m^{-a} \leq \tau_{\mathrm{w}} \leq c_2 m^{-a}$, and $m \asymp d$. For $i \leq m$, $F_i^{\mathrm{w}} = (1 + \tau_{\mathrm{w}} i^a)^{-1}$ is nonincreasing in i with $F_m^{\mathrm{w}} = (1 + \tau_{\mathrm{w}} m^a)^{-1} \in [(1+c_2)^{-1}, 1]$, so $F_i^{\mathrm{w}} \asymp 1$ uniformly. By (G.1), $1 - \Xi_{\mathrm{w}} = n_{\mathrm{w}}^{-1} \sum_{i \leq m} F_i^{\mathrm{w}}(1 - F_i^{\mathrm{w}}) \leq 1$; and for $m/2 < i \leq m$ one has $\lambda_i \in [m^{-a}, 2^a m^{-a}]$, whence

$$F_i^{\mathrm{w}}(1 - F_i^{\mathrm{w}}) = \frac{\lambda_i \tau_{\mathrm{w}}}{(\lambda_i + \tau_{\mathrm{w}})^2} \;\geq\; \frac{m^{-a} \cdot c_1 m^{-a}}{\big((2^a + c_2) m^{-a}\big)^2} \;=\; \frac{c_1}{(2^a + c_2)^2} \;=:\; c_3,$$

and summing over the at least $m/3$ such i (for $m \geq 6$) against $n_{\mathrm{w}} \leq m$ gives $1 - \Xi_{\mathrm{w}} \geq c_3/3$.

Finally $\Gamma_{\mathrm{w}} = (\tau_{\mathrm{w}}^2 A + s_{\mathrm{w}}^2)/(n_{\mathrm{w}}(1 - \Xi_{\mathrm{w}}))$ with $A = \sum_{i \leq m} \lambda_i \omega_i^2/(\lambda_i + \tau_{\mathrm{w}})^2$: the numerator lies in $[\sigma^2, \sigma^2 + 2r^2]$, because $\tau_{\mathrm{w}}^2/(\lambda_i + \tau_{\mathrm{w}})^2 \leq 1$ makes $\tau_{\mathrm{w}}^2 A \leq \sum_i \lambda_i \omega_i^2 = r^2$, and $s_{\mathrm{w}}^2 = \sigma^2 + \sum_{i > m} \lambda_i \omega_i^2 \in [\sigma^2, \sigma^2 + r^2]$. With $n_{\mathrm{w}} \asymp d$ and $1 - \Xi_{\mathrm{w}} \asymp 1$ this gives $\Gamma_{\mathrm{w}} \asymp d^{-1}$, hence $\Gamma_{\mathrm{w}} m \asymp 1$. Finally $\bar{E} = r^2 + \sum_j \lambda_j v_j^{\mathrm{w}} = r^2 + \Gamma_{\mathrm{w}} n_{\mathrm{w}} \Xi_{\mathrm{w}}$ (by $\lambda_j v_j^{\mathrm{w}} = \Gamma_{\mathrm{w}}(F_j^{\mathrm{w}})^2$ and $\sum_{j \leq m}(F_j^{\mathrm{w}})^2 = n_{\mathrm{w}} \Xi_{\mathrm{w}}$), which is $\geq r^2 > 0$ and $\leq r^2 + \Gamma_{\mathrm{w}} n_{\mathrm{w}} \asymp 1$: so $\bar{E} \asymp 1$.

(ii). Split $R_1^{\mathrm{pop}}(t) = \mathcal{B}(t) + \mathcal{V}(t)$,

$$\mathcal{B}(t) := \sum_{i\le m} \lambda_i \omega_i^2 (1 - F_i F_i^{\mathrm{w}})^2 + T_m, \quad T_m := \sum_{i>m} \lambda_i \omega_i^2, \quad \mathcal{V}(t) := \sum_{i\le m} \lambda_i F_i^2 v_i^{\mathrm{w}},$$

with $F_i = \lambda_i/(\lambda_i + t)$.

Variance. Since $\lambda_i v_i^{\mathrm{w}} = \Gamma_{\mathrm{w}}(F_i^{\mathrm{w}})^2$, we get

$$\mathcal{V}(t) = \Gamma_{\mathrm{w}} \sum_{i\le m} F_i^2 (F_i^{\mathrm{w}})^2 \asymp \Gamma_{\mathrm{w}} \sum_{i\le m} (1 + t i^a)^{-2} \asymp \Gamma_{\mathrm{w}} \min(m, t^{-1/a}),$$

by (i) and Lemma J.24 (a), for *every* $t \in (0, 1]$.

Bias. For $t \ge \tau_{\mathrm{w}}$ and every i, cross-multiplying $\tau_{\mathrm{w}}\lambda_i \le t\lambda_i$ gives $1 - F_i^{\mathrm{w}} = \tau_{\mathrm{w}}/(\lambda_i + \tau_{\mathrm{w}}) \le t/(\lambda_i + t) = 1 - F_i$; since $1 - F_i F_i^{\mathrm{w}} = (1 - F_i) + F_i(1 - F_i^{\mathrm{w}})$, this sandwiches $1 - F_i \le 1 - F_i F_i^{\mathrm{w}} \le 2(1 - F_i)$ — an estimate free of b. Hence the head sum is $\asymp \sum_{i\le m} \lambda_i \omega_i^2 (1 - F_i)^2 = G\,\Sigma_{a+b}(t) \asymp \Phi_{ab}(t)$ by Lemma J.24 (b) with $s = a + b$, $p = m$, valid once $4t^{-1/a} \le m$: for $b \ne a + 1$ that lemma gives $t^\theta = \Phi_{ab}(t)$ directly, and at $b = a + 1$ (i.e. $s = 2a + 1$) its logarithmic branch gives $t^2 \log(2/t)$, which is $\asymp t^2 \log(e/t) = \Phi_{ab}(t)$ on $(0, 1]$ because $\log 2 \le \log(2/t)/\log(e/t) \le 1$ there.

And $T_m \asymp m^{-(a+b-1)} \le m^{-a\theta}$ because $a\theta \le a + b - 1$. Set $K := \max\{4^a, c_2, 1\}$. For $t \in [Km^{-a}, 1]$ both provisos hold, $T_m \le CK^{-\theta} t^\theta \le CK^{-\theta}\Phi_{ab}(t)$ (using $\Phi_{ab}(t) \ge t^\theta$ on $(0, 1]$), and $\min(m, t^{-1/a}) = t^{-1/a}$; so $R_1^{\mathrm{pop}}(t) \asymp \Phi_{ab}(t) + \Gamma_{\mathrm{w}} t^{-1/a}$ there.

Off the interval. If $t < Km^{-a}$ then $\mathcal{V}(t) \ge c\,\Gamma_{\mathrm{w}} \min(m, K^{-1/a} m) \ge c_4 \Gamma_{\mathrm{w}} m \asymp 1$. If $t > 1$ then the head sum of $\mathcal{B}$ is nondecreasing in t, so $\mathcal{B}(t) \ge \mathcal{B}(1) \asymp 1$. Either way $R_1^{\mathrm{pop}}(t) \ge c > 0$, whereas the next paragraph exhibits points of $[Km^{-a}, 1]$ at which $R_1^{\mathrm{pop}} \to 0$; so for $d \ge d_0$ every minimizer lies in $[Km^{-a}, 1]$.

The minimizer. Put $\gamma := \Gamma_{\mathrm{w}} \asymp d^{-1}$ and $h(t) := \Phi_{ab}(t) + \gamma t^{-1/a}$, so (renaming constants) $c_1 h \le R_1^{\mathrm{pop}} \le c_2 h$ on $[Km^{-a}, 1]$. This h is exactly the objective of Lemma K.10 at $A = 1$, $B = \gamma$, whose parameters are $\mu = a\theta + 1 = a/\varrho$, $Q(1, \gamma) = \gamma^{1-1/\mu} = \gamma^{\theta\varrho} \asymp d^{-\theta\varrho}$ and $\ell = 1 + \log(1/\gamma) \asymp \log d$.

Case $b \ne a + 1$, where $\Phi_{ab}(t) = t^\theta$ and $\Lambda_d = 1$. Then h is minimized at $t_\gamma = (\gamma/(a\theta))^{a/(a\theta+1)} = (\gamma/(a\theta))^{\varrho} \asymp d^{-\varrho}$, with $\gamma t_\gamma^{-1/a} = a\theta\, t_\gamma^\theta$ and $h(t_\gamma) = (1 + a\theta) t_\gamma^\theta$; since $\varrho < a$, $t_\gamma \in [Km^{-a}, 1]$ for $d \ge d_0$. For a minimizer t^*, $c_1 h(t^*) \le R_1^{\mathrm{pop}}(t^*) \le R_1^{\mathrm{pop}}(t_\gamma) \le c_2 h(t_\gamma)$, so $h(t^*) \le C\, t_\gamma^\theta$; combined with $h(t) \ge t^\theta$ and $h(t) \ge \gamma t^{-1/a}$ this gives $t^{*\theta} \le C t_\gamma^\theta$ and $\gamma t^{*\,-1/a} \le C\gamma t_\gamma^{-1/a}$, i.e. $t^* \asymp t_\gamma \asymp d^{-\varrho}$. Consequently $R_1^{\mathrm{pop},*} \asymp h(t_\gamma) \asymp t_\gamma^\theta \asymp d^{-\theta\varrho}$.

Case $b = a + 1$, where $\theta = 2$, $\mu = 2a + 1$, $\varrho = a/(2a + 1)$, $\Phi_{ab}(t) = t^2 \log(e/t)$ and $\Lambda_d = \log d$. Here Lemma K.10 (c) at $A = 1 \ge \gamma = B$ is two-sided:

$$\tfrac{a}{2(2a+1)}\, Q\ell^{1/(2a+1)} \;\le\; \inf_{0<t\le 1} h(t) \;\le\; h(t^\dagger) \;\le\; 3\, Q\ell^{1/(2a+1)}.$$

Here

$$Q\ell^{1/(2a+1)} \asymp d^{-\frac{2a}{2a+1}} (\log d)^{\frac{1}{2a+1}},$$

at $t^\dagger = \gamma^{a/(2a+1)} \ell^{-a/(2a+1)} \asymp (d \log d)^{-a/(2a+1)}$, which lies in $[Km^{-a}, 1]$ for $d \ge d_0$ because $a/(2a + 1) < a$. So $R_1^{\mathrm{pop},*} \le c_2 h(t^\dagger) \le C\, d^{-\theta\varrho} \Lambda_d^{\varrho/a}$ and $R_1^{\mathrm{pop},*} \ge c_1 \inf_{(0,1]} h \ge c\, d^{-\theta\varrho} \Lambda_d^{\varrho/a}$, using $\theta\varrho = 2a/(2a + 1)$ and $\varrho/a = 1/(2a + 1)$: this is the stated order of $R_1^{\mathrm{pop},*}$.

It remains to localize t^*, and one bootstrap does it. Write $h(t^*) \le c_1^{-1} R_1^{\mathrm{pop},*} \le C\, Q\ell^{1/(2a+1)}$.

Lower. $h(t) \ge \gamma t^{-1/a}$ gives $t^{*\,-1/a} \le C\, Q\ell^{1/(2a+1)}/\gamma$, and $Q/\gamma = \gamma^{2a/(2a+1)-1} = \gamma^{-1/(2a+1)} \asymp d^{1/(2a+1)}$, so $t^{*\,-1/a} \le C(d \log d)^{1/(2a+1)}$, i.e. $t^* \ge c(d \log d)^{-a/(2a+1)}$.

Upper. Since $\log(e/t) \ge 1$ on $(0, 1]$, $h(t) \ge t^2$, so $t^{*2} \le C\, Q\ell^{1/(2a+1)}$ and hence $t^* \le C\, d^{-a/(2a+1)} (\log d)^{1/(2(2a+1))}$; in particular $\log(e/t^*) \ge \frac{a}{2a+1} \log d - C' \log\log d \ge c \log d$ for

$d \geq d_0$. Feeding that back into $h(t) \geq t^2 \log(e/t)$,

$$t^{*2} \;\leq\; \frac{C\,Q\ell^{1/(2a+1)}}{c \log d} \;\asymp\; d^{-\frac{2a}{2a+1}} (\log d)^{\frac{1}{2a+1}-1} \;=\; (d \log d)^{-\frac{2a}{2a+1}},$$

i.e. $t^* \leq C(d \log d)^{-a/(2a+1)}$. Together, $t^* \asymp (d \log d)^{-\varrho} = (d\Lambda_d)^{-\varrho}$, as claimed.

In both cases $t^*/\tau_{\mathrm{w}} \asymp (d\Lambda_d)^{-\varrho}/d^{-a}$ and $t^* N^a \asymp (d\Lambda_d)^{-\varrho} d^a$, both $\asymp d^{\,a-\varrho}\Lambda_d^{-\varrho} \to \infty$ since $\varrho < a$ and Λ_d is at most logarithmic; and $\theta\varrho = 1 - \varrho/a$ by $\varrho(\theta + 1/a) = 1$.

(iii). By Lemma J.24 (a) with $p = d$, $\mathrm{ed}(t^*) \asymp \min(d, t^{*\,-1/a}) = t^{*\,-1/a} \asymp (d\Lambda_d)^{\varrho/a}$, the minimum being the second term because $\varrho/a = 1/(a\theta + 1) < 1$ and Λ_d is at most logarithmic. Hence $\mathrm{ed}(t^*)/N \asymp d^{\varrho/a - 1}\Lambda_d^{\varrho/a} \to 0$.

(iv). By (J.16), $\mathrm{toll}(t^*) = \mathrm{num} \cdot \Xi/(1-\Xi)$ with $\Xi = \mathrm{ed}(t^*)/N \to 0$ by (iii), so $\Xi/(1-\Xi) \asymp \mathrm{ed}(t^*)/N$; it remains to show $\mathrm{num} = \sum_j \lambda_j (1 - F_j^*)^2 \rho_j \asymp 1$, where $\rho_j = (F_j^{\mathrm{w}})^2 \omega_j^2 + v_j^{\mathrm{w}}$. Above,

$$\mathrm{num} \;\leq\; \sum_j \lambda_j \rho_j \;\leq\; \sum_j \lambda_j \omega_j^2 + \Gamma_{\mathrm{w}} \sum_{j \leq m} (F_j^{\mathrm{w}})^2 \;=\; r^2 + \Gamma_{\mathrm{w}} n_{\mathrm{w}} \Xi_{\mathrm{w}} \;\leq\; r^2 + \Gamma_{\mathrm{w}} n_{\mathrm{w}} \;\asymp\; 1.$$

Below, restrict to $m/2 < j \leq m$: there $\lambda_j v_j^{\mathrm{w}} = \Gamma_{\mathrm{w}} (F_j^{\mathrm{w}})^2 \geq c\Gamma_{\mathrm{w}}$ by (i), while $F_j^* \leq \lambda_{\lceil m/2 \rceil}/t^* \leq Cd^{-a}(d\Lambda_d)^{\varrho} \to 0$ (again $\varrho < a$, Λ_d at most logarithmic), so $(1 - F_j^*)^2 \geq 1/4$ for $d \geq d_0$; summing over the at least $m/3$ such j gives $\mathrm{num} \geq c\Gamma_{\mathrm{w}} m \asymp 1$. Therefore

$$\mathrm{toll}(t^*) \asymp \mathrm{ed}(t^*)/N \asymp d^{\varrho/a - 1}\Lambda_d^{\varrho/a} \asymp N^{-(1-\varrho/a)}\Lambda_N^{\varrho/a},$$

and $N \cdot \mathrm{toll}(t^*) \asymp (N\Lambda_N)^{\varrho/a} \to \infty$.

(v). By (iv), (i) and $d/N \asymp 1$,

$$\mathrm{toll}(t^*) \;\asymp\; \frac{\mathrm{ed}(t^*)}{N} \;\asymp\; \frac{d}{N}\,\Gamma_{\mathrm{w}}\,\mathrm{ed}(t^*) \;\asymp\; \mathcal{V}(t^*),$$

and by the proof of (ii),

$$\mathcal{V}(t^*) \asymp \gamma t^{*\,-1/a} \asymp d^{-1}(d\Lambda_d)^{\varrho/a} = d^{-\theta\varrho}\Lambda_d^{\varrho/a} \asymp R_1^{\mathrm{pop},*}.$$

□

Remark J.26 (Why the fraction is constant; and the measured values). The mechanism behind (v) is one line and is not special to the power law. By (J.16) the toll is $\mathrm{num} \cdot \Xi/(1-\Xi)$ with $\Xi = \mathrm{ed}(\tau)/N$, and $\mathrm{num} \asymp \Gamma_{\mathrm{w}} m$ at any ridge far above the scale of the teacher channel, while the variance part of the population risk is $\asymp \Gamma_{\mathrm{w}} \mathrm{ed}(\tau)$. Hence

$$\mathrm{toll}(\tau) \;\asymp\; \frac{m}{N} \cdot \big[\text{variance part of } R_1^{\mathrm{pop}}(\tau)\big], \qquad \frac{m}{N} = \psi\bar{\kappa}_{\mathrm{s}} \asymp 1.$$

A ridge *tuned* to minimize R_1^{pop} balances bias against variance, so at t^* the variance is a fixed fraction of the whole risk — and so, therefore, is the toll.

Driving the splitting cost below the risk would require an untuned ridge pushed deep into the bias-dominated regime, which costs more risk than it saves toll. This is the honest replacement for the $O(1/N)$ reading: the cost of splitting the budget is not a lower-order correction to a chain risk, it is a constant share of it.

Measured at the population optimum for $(a, b) = (1.2, 2)$, $\psi = 0.3$, $\bar{\kappa}_{\mathrm{w}} = 3$, $N/d = 1/2$, $\sigma^2 = 1/4$, $r^2 = 1$ (so $\theta = 11/6$, $\varrho = 3/8$ and $\varrho/a = 1/(a+b) = 5/16$):

d	t^*	$R_1^{\mathrm{pop},*}$	$\mathrm{toll}(t^*)$	$N \cdot \mathrm{toll}$	$\mathrm{toll}/R_1^{\mathrm{pop},*}$
$1.0 \cdot 10^3$	0.07764	$3.975 \cdot 10^{-2}$	$2.152 \cdot 10^{-3}$	1.076	0.05413
$4.0 \cdot 10^3$	0.04811	$1.722 \cdot 10^{-2}$	$9.853 \cdot 10^{-4}$	1.971	0.05723
$3.2 \cdot 10^4$	0.02157	$4.591 \cdot 10^{-3}$	$2.794 \cdot 10^{-4}$	4.470	0.06085
$5.12 \cdot 10^5$	0.00718	$7.365 \cdot 10^{-4}$	$4.671 \cdot 10^{-5}$	11.958	0.06342
$4.096 \cdot 10^6$	0.00316	$1.826 \cdot 10^{-4}$	$1.173 \cdot 10^{-5}$	24.033	0.06425

$N \cdot \mathrm{toll}$ diverges, and the measured local exponent of toll in d moves -0.564, -0.606, -0.645, -0.664 down the table against the predicted $-(1-\varrho/a) = -0.6875$.

The convergence is slow for a specific and checkable reason. Splitting at j_t in the proof of Lemma J.24 (b) leaves, beside the leading t^θ, a term of order t^2 that is present two-sidedly (the $j = O(1)$ end of the inner sum), so the relative correction is at least of order $t^{2-\theta}$, i.e. $d^{-\varrho(2-\theta)}$ at the tuned ridge — at $(a,b) = (1.2, 2)$ that is $d^{-1/16}$, a sixteenth root, which is what the table has to work against.

The diagnosis is falsifiable, and the branches in which the correction is fast bear it out: continuing the same computation to $d = 10^9$ gives measured toll exponents -0.6661 at $(a,b) = (2,1)$ (predicted $-2/3$), -0.6866 at $(3, 0.2)$ (predicted -0.6875), and -0.7482, with $\mathrm{ed}(t^*)$-exponent $+0.2518$, at $(1.5, 4)$ in the saturation branch (predicted -0.75 and $+0.25$). The ratio $\mathrm{toll}/R_1^{\mathrm{pop},*}$ is still creeping upward past the foot of the table — 0.0648, 0.0650, 0.0652 at $d = 10^8$, 10^9, 10^{10} — which is what *(v)* asserts, namely that the ratio stays bounded away from 0; it does not assert that the ratio has converged by $d = 10^7$.

Across six exponent pairs $(a,b) \in \{(1.2,2), (2,1), (3,0.2), (1.5,4), (1.5,2.5), (1.2,2.2)\}$ and eight budget doublings $d = 10^3 \cdot 2^k$, $k \le 7$ — same ψ, $\bar{\kappa_{\mathrm{w}}}$, d/N, σ^2, r^2 — the ratio $\mathrm{toll}/R_1^{\mathrm{pop},*}$ stays in $[0.0541,\ 0.0936]$, rising monotonically in d within every pair. That band is a measurement; what *(v)* proves is the $\asymp$, and the constants its proof carries are far looser than the band.

The saturation boundary $b = a + 1$ is included in the proposition, and the numbers show the predicted logarithm. At $(a,b) = (1.5, 2.5)$ with the same ψ, $\bar{\kappa_{\mathrm{w}}}$, d/N, σ^2, r^2 (so $\theta = 2$, $\varrho = 3/8$, $\varrho/a = 1/4$), the *log-free* normalization $\mathrm{toll}(t^*)\, N^{1-\varrho/a} = \mathrm{toll}(t^*)\, N^{3/4}$ reads 0.2498, 0.2877, 0.3111, 0.3282, 0.3421, 0.3540 at $d = 10^4, \dots, 10^9$: a 42% rise over five decades, not settling.

Dividing by the $\Lambda_N^{\varrho/a} = (\log N)^{1/4}$ that *(iv)* predicts does settle it — 0.1462, 0.1586, 0.1635, 0.1656, 0.1667, 0.1673, the increments halving each decade, apparently toward ≈ 0.168. The ratio $\mathrm{toll}/R_1^{\mathrm{pop},*}$ over the first three of those reads 0.0704, 0.0745, 0.0760: bounded away from 0, which is *(v)* at the boundary. At $(a,b) = (1.2, 2.2)$ the log-free column likewise fails to settle, running 0.2217 to 0.3349.

These are numbers, not a proof; what is proved is the proposition, boundary included — and the mechanism is the one Lemma J.24 (b) names, $s = a + b = 2a + 1$ being exactly its logarithmic case.

(ii) A certified finite-budget chain improvement. Setup $d = 4000$, $a = 1.2$, teacher $(m, n_{\mathrm{w}}, \sigma^2) = (1200, 400, 0.25)$, budget $N = 2000$. Fixed points are enclosed by interval bisection and propagated through directed-rounding risk bounds.

A — chain strictly wins ($b = 2$), fully rigorous:

$$\text{upper: } R_2^*(N) \le R_2^{\mathrm{fin}}(f = 0.773, \lambda_1 = 0.0219, \lambda_2 = 0.0173) \le 0.01630178,$$
$$\text{lower: } R_1^*(N) \ge R_1^{\mathrm{pop},*} = \min_{t \in [0,\infty]} R_1^{\mathrm{pop}}(t) \ge 0.01721744.$$

The lower bound is *global*: directed-rounding interval arithmetic (60 digits) encloses the teacher channel, certifies $R_1^{\mathrm{pop}\prime\prime} > 0$ on the cell containing the minimizer (second-order Taylor bound), covers the rest of $[0, 1/4]$ by adaptive interval cells, and disposes of all $t \ge 1/4$ — including the $t = \infty$ zero-estimator limit — by a termwise-monotone tail bound, giving $\min_{t\in[0,\infty]} R_1^{\mathrm{pop}}(t) \ge 0.0172174469$ (`certificates/chainA_iv_cert.py`; this replaces an earlier bound, which was strongly convex only on $[\hat{t}/2, 2\hat{t}]$ and hence local).

Since *any* single ridge, at any N, has risk $\ge R_1^{\mathrm{pop}}(\tau_{\mathrm{s}}(\lambda)) \ge \min_t R_1^{\mathrm{pop}}$ (self-noise ≥ 0 termwise), we get $R_2^*(N) \le 0.0163018 < 0.0172174 \le R_1^*(N)$: a certified margin ≥ 0.00091567. The fresh paired simulation in Section N gives $+0.001984 \pm 0.000447$ at $d = 4000$ (two standard errors).

Remark J.27 (A finite-budget converse needs a separate argument). At $b = 0$ the population filter gap is zero by Theorem D.11 (i). This fact alone does not compare the optimized finite-budget objectives: both the effective-ridge constraints and fluctuation costs change when data are split. To certify that every two-stage choice loses at a specified budget, one would need a global lower bound over (n_1, τ_1, τ_2). A numerical optimizer's failure to find a winning chain does not provide that bound. No such finite-budget converse is claimed here.

J.7 Mechanism: composition raises the filter order

Proved (algebra). A single ridge hop is the resolvent shrinkage $F = \Sigma(\Sigma + tI)^{-1}$ (Tikhonov), rolloff $F_i \approx \lambda_i/t = O(\lambda_i)$; the two-hop chain is the product of two resolvents $F = \Sigma^2(\Sigma + t_1 I)^{-1}(\Sigma + t_2 I)^{-1}$, rolloff $O(\lambda_i^2)$ — a higher-order low-pass with a sharper cut of the ill-conditioned tail. For the power law $\omega_i^2 \propto i^{-b} = \lambda_i^{b/a}$, i.e. the source condition

$$w^* \in \mathrm{Range}(\Sigma^r), \qquad r = b/(2a), \tag{J.24}$$

so larger b gives a smoother source and a signal-poor tail — exactly the $\Delta < 0$-on-the-tail mechanism of the criterion.

Remark J.28 (Why this is *not* classical qualification). The tempting analogy is the qualification phenomenon (Engl et al., 1996, Chs. 4–5): single Tikhonov has qualification 1 and saturates, iterated Tikhonov (residual $(t/(\lambda + t))^2$) reaches qualification 2. The analogy is *false* for our composed filter, and the distinction is instructive.

Our two-hop filter is the product of resolvents $\lambda^2/((\lambda + t_1)(\lambda + t_2))$; its *head* residual is $1 - F_1 F_2 \approx (t_1 + t_2)/\lambda$ — order one in t, exactly like a single ridge — so its qualification in the classical (bias-side, source-smoothness) sense is still 1, for any number of hops. What jumps is the *tail* rolloff: a single ridge passes tail noise as λ/t, the composed filter as $\lambda^2/(t_1 t_2)$ — an order jump on the *variance* side.

That is the mechanism behind the criterion: on aligned signals the single knob must over-shrink the head to control the tail, and composition relieves precisely this, which is why a noiseless teacher (nothing in the tail to kill) yields no chain gain (Remark J.9).

Only the filter forms and the source condition (J.24) carry theorems; the classical saturation-rate ladder is not invoked anywhere. Unlike self-distillation, whose rounds refit the same training inputs and add neither fresh data nor fresh noise (Mobahi et al., 2020), each hop here also injects fluctuation and must be paid for with a data slice — the tension the criterion resolves.

J.8 From a finite formula certificate to estimator separation

A finite spectrum has a positive smallest eigenvalue, even when its entries were generated from a power-law formula. Holding these levels fixed while replicating their multiplicities gives a bounded-spectrum proportional sequence. This transfers a strict certified formula gap to an estimator theorem. It does not establish an approximation for a power-law spectrum whose smallest eigenvalue keeps decreasing.

Lemma J.29 (Uniformity over the single student's ridge)**.** *Under Theorem 3.2, write $\widehat{w}_{\mathrm{s}}(\eta)$ for the student fit on one common fresh design, and $\mathcal{R}_s^{(d)}(\eta)$ for its finite-dimensional formula. Then*

$$\sup_{\eta \in [0,\infty)} \left| \mathcal{E}(\widehat{w}_{\mathrm{s}}(\eta)) - \mathcal{R}_s^{(d)}(\eta) \right| \xrightarrow{\mathbb{P}} 0. \tag{J.25}$$

The same convergence holds in L^1.

Proof. **1. A compact interval of ridges.** The Gaussian singular-value bounds and the uniform aspect-ratio gap imply that, with probability tending to one, all nonzero eigenvalues of $\widehat{\Sigma} = \widetilde{X}^\top \widetilde{X}/n_{\mathrm{s}}$ lie in a fixed interval $[a_0, a_1] \subset (0, \infty)$ (Vershynin, 2012, Corollary 5.35). On this event the fitting operator $S_\eta = (\widehat{\Sigma} + \eta I)^\dagger \widehat{\Sigma}$ satisfies

$$\|S_\eta - S_{\eta'}\|_{\mathrm{op}} \le a_0^{-1} |\eta - \eta'|, \qquad \eta, \eta' \ge 0,$$

including the continuous minimum-norm endpoint at zero. Since the teacher and signal norms are tight, their realized risk curve has a tight Lipschitz constant on every compact interval.

The deterministic curves have uniformly bounded Lipschitz constants there as well. Indeed, differentiating the fixed-point equation gives

$$\frac{d\tau}{d\eta} = \frac{1}{1 - \Xi}, \qquad \frac{dF_i}{d\eta} = -\frac{c_i}{(c_i + \tau)^2 (1 - \Xi)}.$$

The gap and the lower bound on τ in Lemma G.1, together with $\tau \le \eta + \sum_i c_i/n \le \eta + KM$, bound these derivatives, those of Ξ and Γ, and hence the derivative of the risk formula. Its teacher profile has bounded total mass. Apply Theorem 3.2 on a fixed finite grid, and then let the grid spacing tend to zero. The two equicontinuity bounds yield uniform convergence on each fixed compact interval. This is a simultaneous statement over the fitted ridge path, not only a parameter-uniform bound at a separately fixed ridge.

2. The unbounded parameter tail. On the same spectral event, $\sup_{\eta \ge A} \|S_\eta \beta_{\mathrm{w}}\| \le a_1 \|\beta_{\mathrm{w}}\|/A$. Consequently the entire realized tail approaches $\|w^*\|_\Sigma^2$ as $A \to \infty$, uniformly in large dimension and in probability. In the formula, $\tau \ge \eta$, $F_i \le M/\eta$, and its student second-moment profile has total mass $O(\eta^{-2})$: use the bounded teacher mass and $\sum_i c_i/(c_i+\tau)^2 = O(d/\tau^2)$ in the formula for Γ. Thus its entire tail approaches $r_d^2 = \sum_i \lambda_i \omega_i^2$ uniformly as well. Finally $\|w^*\|_\Sigma^2 - r_d^2 \to 0$ by the prior moment condition. First choose A large, then apply the compact-interval result.

3. Expected uniform error. All fitted student norms are at most $\|\beta_{\mathrm{w}}\|$, and the deterministic curves are uniformly bounded. Lemma G.5 gives uniform integrability of the supremum in (J.25). □

Theorem J.30 (Replicated-spectrum transfer)**.** *Fix a finite template with d_0 positive eigenvalues, a head of size m_0, true-label budget $n_{\mathrm{w},0} < m_0$, and a two-stage allocation $n_{1,0} + n_{2,0} = N_0 < d_0$, with $0 < n_{j,0} < d_0$. Fix its signal variances, noise, and teacher and two-stage ridges. Suppose its deterministic formulas satisfy*

$$\mathcal{R}_{2,0} \le a < b \le \inf_{\eta \ge 0} \mathcal{R}_{1,0}(\eta).$$

For each integer $q \ge 1$, replace coordinate i by q independent coordinates of eigenvalue λ_i and signal variance ω_i^2/q, and multiply every dimension and sample budget by q. Take centered independent signals with a uniformly bounded fourth-moment ratio, and fresh Gaussian designs. For every $\epsilon > 0$,

$$\Pr\left\{ \inf_{\eta \ge 0} \mathcal{E}(\widehat{w}_{\mathrm{s},q}(\eta)) - \mathcal{E}(\beta_q^{(2)}) \ge b - a - \epsilon \right\} \longrightarrow 1. \tag{J.26}$$

The single-stage baseline uses qN_0 inputs; the chain uses the same total budget. The baseline may use the concatenation of the chain's two batches. Its ridge may even be chosen after seeing those data and the target, because the comparison holds simultaneously for all ridges.

Proof. **1. Replication leaves the aggregate formulas unchanged.** Every spectral count and sample size is multiplied by q, so the fixed points τ and the scalars Ξ are unchanged at each ridge. The signal variance of each replica is divided by q; the summed signal energy and omitted-signal noise are unchanged. Hence Γ and each coordinate fluctuation variance are divided by q. Inductively every coordinate second moment is divided by q, and every aggregate risk is exactly the template value. This holds for every single-stage ridge.

2. Verify the estimator theorem's assumptions. The largest and smallest eigenvalues are the fixed template extrema; all aspect ratios remain fixed away from one. Also $(qd_0)(\omega_i^2/q) = d_0\omega_i^2$ is uniformly bounded, and the index-resolved empirical profile converges. Thus the two-stage risk converges to $\mathcal{R}_{2,0}$ by Theorem D.9. By Lemma J.29, the infimum of the realized single-stage risk differs from the infimum of its formula by $o_{\mathbb{P}}(1)$. A union bound proves (J.26). Shared inputs between the two compared procedures cause no problem: convergence of each procedure separately suffices. □

Apply the theorem to the certificate in Theorem D.10 (ii):

$$(d_0, m_0, n_{\mathrm{w},0}, N_0) = (4000, 1200, 400, 2000),$$
$$(n_{1,0}, n_{2,0}) = (1546, 454), \qquad (\eta_1, \eta_2) = (0.0219, 0.0173).$$

The template has $\lambda_i = i^{-1.2}$, $\omega_i^2 = Gi^{-2}$ normalized to signal power one, a minimum-norm teacher, and $\sigma^2 = 0.25$. The directed-rounding certificate gives $b - a > 9.1567 \times 10^{-4}$. Therefore its replicated sequence has a strict estimator-level advantage over every single-stage ridge under the same total unlabeled budget. This is an asymptotic existence result; it gives no finite-dimensional failure probability for the unreplicated $d_0 = 4000$ experiment, and is not a power-law estimator theorem.

K PROOFS OF THE SCALING THEOREMS (THEOREM D.14)

Theorem D.14 (Three scaling regimes)**.** *Power-law family (C.2), $a > 1$; the statements are exact asymptotics of the limiting risk formulas (cf. the division of labor after Assumption C.1).*

(i) Imitation rate. *With a noiseless exact teacher, and with the student's overparametrization ratio bounded away from the interpolation boundary — $n_s < d$ and $\liminf_d d/n_s > 1$, the clause Assumption C.1 already imposes, here also allowing $d/n_s \to \infty$ (sparse data) — the MNI student obeys*

$$\bar{\mathcal{E}}_s(n_s) \asymp n_s^{-\min(a+b-1,\, 2a)}$$

(times $\log n_s$ at the boundary $b = a+1$), with constants depending only on (a, b, r^2) and that bound: the source rate up to the interpolation ceiling $2a$.

The hypothesis is not technical — at $d = n_s+1$ the exponent is strictly larger, $\bar{\mathcal{E}}_s \asymp n_s^{-(a+b)}$ for $b < a+1$ (Remark K.12).

(ii) Noise plateau. *With $\sigma^2 > 0$ and fixed ratios $(\psi, \bar{\kappa_w}, \bar{\kappa_s})$, scaling all of (m, n_w, n_s, d) proportionally drives $\bar{\mathcal{E}}_s$ to an explicit strictly positive limit*

$$\sigma^2\, C_{\mathrm{floor}}(a, \psi, \bar{\kappa_w}, \bar{\kappa_s}) :$$

interpolating pipelines do not obey a decaying scaling law at fixed shape.

This is a limit, not a finite-d lower bound, and the approach is not one-sided: in the power-law families we evaluate $\bar{\mathcal{E}}_s$ falls to $\sigma^2 C_{\mathrm{floor}}$ from above *at small source exponents and rises to it from below at large ones (Remark K.13).*

(iii) No scaling penalty for weak supervision. *With $\sigma^2 > 0$, optimally tuned ridge at both stages, generous capacity $m = \Theta(n_w)$ and an imitation budget that dominates the label budget, $d/n_s \to \bar{\kappa_s} \in (1, \infty)$ with $n_s/n_w \to \infty$ (the regime in which the student is a near-clone and its self-noise is lower order; the proof uses the bound away from 1, not merely $n_s = \Theta(d)$, and Remark K.15 records where),*

$$\bar{\mathcal{E}}_s^*(n_w) \asymp n_w^{-(a+b-1)/(a+b)}$$

for $b < a+1$ (saturating at $n_w^{-2a/(2a+1)}$ for $b > a+1$, with an extra $(\log n_w)^{1/(2a+1)}$ factor at the boundary $b = a+1$) — the same exponent as optimally-ridged supervised regression on the n_w true labels.

Division of labor (what these statements are about). Every result here is a rigorous statement about the *deterministic limit formula* $\bar{\mathcal{E}}_s$ of (D.2) — the composed channel calculus, which is well defined for any summable spectrum and in particular along the power law (C.2). Theorem D.1 proves $\mathcal{E}_s \xrightarrow{\mathbb{P}} \bar{\mathcal{E}}_s$ — at the realized profile; at the ω^2-profile only under Remark G.4 (G.9), which Remark G.4 (b) shows *fails* along the power law — for spectra of bounded condition number, which the power-law family is not.

The rates, floor and optimal exponent below are therefore theorems about $\bar{\mathcal{E}}_s$ (Remark K.16); this is exactly the division of labor announced after Assumption C.1. Two sequence-model limits appear: the *sparse-data* limit ($n_s \to \infty$, $n_s/d \to 0$; the active band is a vanishing fraction) drives the rate (i), and the *proportional* limit ($d \to \infty$ at fixed ratios; truncation binds) drives the floor (ii) and the optimal-ridge exponent (iii).

The model is the sequence model $\lambda_i = i^{-a}$ ($a > 1$), $\omega_i^2 = G i^{-b}$ ($b \geq 0$), normalized by $\sum_i \lambda_i \omega_i^2 = r^2$, so $G = r^2/\zeta(a+b)$ with ζ the Riemann zeta function.

K.1 SPECTRAL TOOLKIT

Lemma K.1 (Sum–integral comparison)**.** *If $g : (0,\infty) \to (0,\infty)$ is non-increasing with $\int_0^\infty g = \mathrm{J} < \infty$, then*

$$\mathrm{J} - \int_0^1 g \;\le\; \int_1^\infty g \;\le\; \sum_{i\ge 1} g(i) \;\le\; \int_0^\infty g \;=\; \mathrm{J}.$$

Peeling off the first term sharpens the upper bound to $\sum_{i\ge1} g(i) \le g(1) + \int_1^\infty g$, and the two bounds together bracket the deficit *from both sides:*

$$\int_0^1 g \;-\; g(1) \;\le\; \mathrm{J} - \sum_{i\ge1} g(i) \;\le\; \int_0^1 g.$$

Consequently, for a family (g_τ) with $g_\tau(1) = o(\mathrm{J}(\tau))$ as $\tau \to 0$ — no growth condition on J is needed — one has $\sum_{i\ge1} g_\tau(i) = \mathrm{J}(\tau)(1+o(1))$ if and only if $\int_0^1 g_\tau = o(\mathrm{J}(\tau))$.
Neither half is free. The "if" half needs no hypothesis beyond the bracket, but its premise is not removable, and divergence of $\mathrm{J}(\tau)$ is no substitute for it: $g_\tau(x) = \min\{\tau^{-1}, x^{-1}\}$ on $(0,1]$ and e^{1-x} on $(1,\infty)$ is positive, non-increasing and integrable with $\mathrm{J}(\tau) = 2 + \log(1/\tau) \to \infty$ and $g_\tau(1) = 1 = o(\mathrm{J}(\tau))$, yet $\sum_{i\ge1} g_\tau(i) = 1 + (e-1)^{-1}$ does not depend on τ, so the ratio tends to 0 — as it must, since $\int_0^1 g_\tau/\mathrm{J}(\tau) \to 1$.
The "only if" half is where the hypothesis $g_\tau(1) = o(\mathrm{J}(\tau))$ is spent, and it is false without it: $g_\tau(x) = \tau^{-1}$ on $(0,2]$ and τe^{2-x} on $(2,\infty)$ has $\mathrm{J}(\tau) = 2\tau^{-1} + \tau$ and $\sum_{i\ge1} g_\tau(i) = 2\tau^{-1} + \tau(e-1)^{-1}$, so the conclusion holds*, while $\int_0^1 g_\tau/\mathrm{J}(\tau) \to \frac12$; there g_τ is constant on $(0,1]$, so $\int_0^1 g_\tau = g_\tau(1)$ and the deficit bracket degenerates to $0 \le \mathrm{J} - \sum_{i\ge1} g_\tau(i) \le \tau^{-1}$.*
Each use in this paper is either of the two-sided bracket alone (Lemma K.9, Proposition K.14) or of the "if" half at a kernel bounded by 1 (Lemma K.3), where $\int_0^1 g_\tau \le 1 = o(\mathrm{J}(\tau))$ is immediate; the "only if" half is never invoked. Unimodal variant *(used where the integrand rises then falls, e.g. the K_{ab} recombination at $b < a$): if g is unimodal, the same bracketing on each monotone piece gives $|\sum_{i\ge1} g(i) - \int_0^\infty g| \le 2\sup g$, and $\sup g/\mathrm{J}(\tau) = O(\tau^{1/a})$ for the kernels in this section — which is the corresponding hypothesis, and it is what delivers the $(1+o(1))$ conclusion there.*

Proof. Monotone bracketing $\int_i^{i+1} g \le g(i) \le \int_{i-1}^i g$, summed over $i \ge 1$, gives the displayed chain; summing the right-hand inequality over $i \ge 2$ only and adding $g(1)$ gives $\sum_{i\ge1} g(i) \le g(1) + \int_1^\infty g$, and both deficit bounds follow using $\int_1^\infty g = \mathrm{J} - \int_0^1 g$.

For the equivalence, write $\mathrm{D}(\tau) := \mathrm{J}(\tau) - \sum_{i\ge1} g_\tau(i) \ge 0$: if $\int_0^1 g_\tau = o(\mathrm{J})$ then $\mathrm{D} \le \int_0^1 g_\tau = o(\mathrm{J})$; conversely, if $\mathrm{D} = o(\mathrm{J})$ then $\int_0^1 g_\tau \le \mathrm{D} + g_\tau(1) = o(\mathrm{J})$, the last term by hypothesis — the only place that hypothesis is used, and the second witness shows it cannot be dropped. For the unimodal case, split at the mode and bracket each side, losing at most one term ($\le \sup g$) per side.

For the first witness, $\int_0^1 \min\{\tau^{-1}, x^{-1}\}dx = 1 + \log(1/\tau)$ and $\int_1^\infty e^{1-x}dx = 1$, while $\sum_{i\ge1} g_\tau(i) = 1 + \sum_{i\ge2} e^{1-i} = 1 + (e-1)^{-1}$; for the second, $\int_0^2 \tau^{-1}dx = 2\tau^{-1}$, $\int_2^\infty \tau e^{2-x}dx = \tau$ and $\sum_{i\ge1} g_\tau(i) = 2\tau^{-1} + \tau\sum_{k\ge1} e^{-k} = 2\tau^{-1} + \tau(e-1)^{-1}$. □

Definition K.2 (Three spectral constants)**.** For $a > 1$, $b \ge 0$, $a + b > 1$,

$$I_a := \int_0^\infty \frac{du}{1+u^a} = \frac{\pi/a}{\sin(\pi/a)}, \quad J_a := \int_0^\infty \frac{du}{(1+u^a)^2}, \quad K_{ab} := \int_0^\infty \frac{u^{a-b}}{(1+u^a)^2}\,du,$$

the last finite iff $b < a + 1$. The Beta integral $\int_0^\infty u^{s-1}(1+u^a)^{-n}\,du = \frac1a \mathrm{B}(s/a,\, n - s/a)$ gives $J_a = \frac1a \mathrm{B}(1/a, 2 - 1/a) = (1 - 1/a)I_a$ and $K_{ab} = \frac1a \mathrm{B}(\frac{a-b+1}{a}, \frac{a+b-1}{a})$.

Lemma K.3 (Sparse-data fixed point; $\Xi_\infty = 1 - 1/a$)**.** *For the full spectrum, $n \to \infty$ with $n/d \to 0$, the ridgeless fixed point solving $\sum_i \lambda_i/(\lambda_i + \tau) = n$ satisfies $\tau = c_a n^{-a}(1 + o(1))$ with $c_a := I_a^a$, and $\Xi(n) \to \Xi_\infty(a) := J_a/I_a = 1 - 1/a \in (0, 1)$; in particular $1 - \Xi_\infty = 1/a$.*

Proof. By Lemma K.1 with $g(x) = 1/(1 + \tau x^a)$ — for which the "if" half of the asymptotic clause applies because $g \le 1$, hence $g(1) \le \int_0^1 g \le 1 = o(\tau^{-1/a} I_a)$ — $\sum_i 1/(1 + \tau i^a) = \tau^{-1/a} I_a(1 + o(1))$; setting this $= n$ gives $\tau = c_a n^{-a}(1+o(1))$. Likewise $\sum_i 1/(1+\tau i^a)^2 = \tau^{-1/a} J_a(1+o(1))$, so $\Xi = \tau^{-1/a} J_a/n \to J_a/I_a$, which equals $1 - 1/a$ by the Beta identity of Definition K.2. □

Lemma K.4 (Proportional fixed point)**.** *Fix $\kappa > 1$; spectrum $\lambda_i = i^{-a}$, $i = 1, \dots, d$, $n = d/\kappa$, $d \to \infty$. Then $\tau = \theta d^{-a}$ and $\Xi \to \Xi_{\mathrm{prop}}(a, \kappa)$, where $\theta = \theta(a, \kappa)$ solves $\int_0^1 dx/(1 + \theta x^a) = 1/\kappa$ and $\Xi_{\mathrm{prop}}(a, \kappa) := \kappa \int_0^1 dx/(1 + \theta x^a)^2 \in (0, 1)$. As $\kappa \to \infty$, $\Xi_{\mathrm{prop}} \to 1 - 1/a$.*

Proof. Scale $x = i/d$: $\frac{1}{d}\sum_{i \le d}(1 + \theta(i/d)^a)^{-1} \to \int_0^1 (1 + \theta x^a)^{-1} dx$; the fixed point $n = d/\kappa$ gives $\int_0^1 (1 + \theta x^a)^{-1} dx = 1/\kappa$ (strictly decreasing in θ, unique root). Then $\Xi = (\kappa/d)\sum(1 + \theta(i/d)^a)^{-2} \to \kappa \int_0^1 (1 + \theta x^a)^{-2} dx$; as $\kappa \to \infty$, $\theta \to \infty$ and $u = \theta^{1/a} x$ sends the integrals to $\theta^{-1/a} I_a$, $\theta^{-1/a} J_a$ on $(0, \infty)$, so $\Xi \to J_a/I_a = 1 - 1/a$. □

Remark K.5 (Ξ is not spectrum-only in the proportional regime). One might guess $\Xi = 1 - 1/a$ for all κ; this is false in the proportional regime, where finite overparametrization makes $\Xi_{\mathrm{prop}}(a, \kappa)$ depend on both. The value $1 - 1/a$ is only the $\kappa \to \infty$ edge (e.g. $a = 2$: $\Xi_{\mathrm{prop}} = 0.742, 0.589, 0.528$ at $\kappa = 1.5, 3, 8$, approaching 0.5 only slowly).

Lemma K.6 (The bias-sum asymptotic)**.** *Let $B_0(\tau) := \tau^2 \sum_i \lambda_i \omega_i^2/(\lambda_i + \tau)^2 = G\tau^2 \sum_i i^{a-b}/(1 + \tau i^a)^2$. As $\tau \to 0$,*

$$B_0(\tau) = \begin{cases} G\, K_{ab}\, \tau^{(a+b-1)/a}(1 + o(1)), & b < a + 1, \\ G\, \tau^2\, \frac{1}{a} \log(1/\tau)(1 + o(1)), & b = a + 1, \\ G\, \zeta(b - a)\, \tau^2 (1 + o(1)), & b > a + 1. \end{cases}$$

Proof. Split at $i^* = \tau^{-1/a}$. On the head ($\lambda_i \ge \tau$) the denominator is $1 + O(1)$, so $\sum_{i \le i^*} i^{a-b}$ behaves as $i^{*\,a-b+1}/(a - b + 1)$ ($b < a + 1$), $\log i^*$ ($b = a + 1$), or $\zeta(b - a)$ ($b > a + 1$). On the tail ($\lambda_i < \tau$), $i^{a-b}/(1+\tau i^a)^2 = \tau^{-2} i^{-a-b}(1+o(1))$ sums to $\tau^{(a+b-1)/a-2}/(a+b-1)$. Multiplying by τ^2: for $b < a + 1$ head and tail recombine to the full-integral constant K_{ab} (substitution $u = \tau^{1/a} x$); for $b = a + 1$ the $\tau^2 \log(1/\tau)$ head dominates; for $b > a + 1$ the $\tau^2 \zeta(b - a)$ head dominates. □

Lemma K.7 (Elementary power sums)**.** *For an integer $M \ge 1$: (a) if $q > -1$ there are $0 < \alpha_q \le \beta_q < \infty$, depending only on q, with $\alpha_q M^{q+1} \le \sum_{i \le M} i^q \le \beta_q M^{q+1}$; (b) if $q < -1$ then $1 \le \sum_{i \le M} i^q \le \zeta(-q)$; (c) $\frac{1}{2}\log(eM) \le \sum_{i \le M} i^{-1} \le 2\log(eM)$; (d) if $s > 1$ then $\sum_{i > M} i^{-s} \le M^{1-s}/(s - 1)$.*

Proof. (a) $M^{-(q+1)} \sum_{i \le M} i^q$ is a strictly positive sequence converging to $1/(q + 1) > 0$, hence bounded above and below by positive constants. (b) Monotone in M, with the stated endpoints. (c) $\log(1+M) \le \sum_{i \le M} i^{-1} \le 1 + \log M$; the left side is $\ge \frac{1}{2}\log(eM)$ because $(1+M)^2/M \ge 4 > e$ for $M \ge 1$, and the right side is $\le 2 + 2\log M = 2\log(eM)$. (d) Compare with $\int_M^\infty x^{-s} dx$. □

Lemma K.8 (Truncated bias sum, two-sided)**.** *Let $a > 1$, $b \ge 0$, $a + b > 1$, and for an integer $p \ge 1$ and $\tau > 0$ put*

$$B_0^{(p)}(\tau) \;:=\; \sum_{i \le p} \lambda_i \omega_i^2 \Big(\frac{\tau}{\lambda_i + \tau}\Big)^2 \;=\; G\,\tau^2 \sum_{i \le p} \frac{i^{a-b}}{(1 + \tau i^a)^2},$$

and

$$\Phi_{ab}(\tau) \;:=\; \begin{cases} \tau^{(a+b-1)/a}, & b < a+1, \\ \tau^2 \log(e/\tau), & b = a+1, \\ \tau^2, & b > a+1. \end{cases}$$

Fix $c_1 \in (0,1]$.
There are $0 < c_- \le c_+ < \infty$, depending only on (a,b,c_1), such that

$$c_-\, G\, \Phi_{ab}(\tau) \;\le\; B_0^{(p)}(\tau) \;\le\; c_+\, G\, \Phi_{ab}(\tau) \qquad \textit{whenever } \tau \in (0,1] \textit{ and } \tau \ge c_1 p^{-a}.$$

Proof. Write $i_\tau := \tau^{-1/a} \ge 1$ and $M := \min(\lfloor i_\tau \rfloor, p) \ge 1$. Every $i \le M$ has $\tau i^a \le 1$, so $\frac{1}{4} \le (1+\tau i^a)^{-2} \le 1$; every i with $M < i \le p$ has $i \ge \lfloor i_\tau \rfloor + 1 > i_\tau$, i.e. $\tau i^a > 1$, so $\frac{1}{4}\tau^{-2} i^{-2a} \le (1+\tau i^a)^{-2} \le \tau^{-2} i^{-2a}$.

Splitting the sum at M gives, with $H := \sum_{i \le M} i^{a-b}$ and $T := \sum_{M < i \le p} i^{-(a+b)} \ge 0$,

$$\tfrac{1}{4} G\big(\tau^2 H + T\big) \;\le\; B_0^{(p)}(\tau) \;\le\; G\big(\tau^2 H + T\big). \tag{K.1}$$

The hypothesis $\tau \ge c_1 p^{-a}$ says $i_\tau \le c_1^{-1/a} p$; with $\lfloor x \rfloor \ge x/2$ for $x \ge 1$ this pins M to the scale of i_τ:

$$\tfrac{1}{2}\, c_1^{1/a}\, i_\tau \;\le\; M \;\le\; i_\tau. \tag{K.2}$$

Indeed $M \le \lfloor i_\tau \rfloor \le i_\tau$; if $\lfloor i_\tau \rfloor \le p$ then $M = \lfloor i_\tau \rfloor \ge i_\tau/2$, and otherwise $M = p \ge c_1^{1/a} i_\tau$ (recall $c_1^{1/a} \le 1$).

By Lemma K.7 (d) and (K.2),

$$T \;\le\; \frac{M^{1-a-b}}{a+b-1} \;\le\; \frac{(\frac{1}{2} c_1^{1/a})^{1-a-b}}{a+b-1}\; i_\tau^{1-a-b} \;=\; C(a,b,c_1)\, \tau^{(a+b-1)/a}. \tag{K.3}$$

It remains to evaluate $\tau^2 H$ in the three cases; throughout, $\tau^2 i_\tau^{a-b+1} = \tau^{2-(a-b+1)/a} = \tau^{(a+b-1)/a}$.

Case $b < a+1$. Here $q := a - b > -1$, so Lemma K.7 (a) with (K.2) gives $\tau^2 H \asymp \tau^2 M^{a-b+1} \asymp \tau^2 i_\tau^{a-b+1} = \tau^{(a+b-1)/a}$ (the exponent $a-b+1$ is positive, so (K.2) may be raised to it). With (K.3), both sides of (K.1) are $\asymp G\tau^{(a+b-1)/a}$.

Case $b > a+1$. Here $q = a-b < -1$, so $1 \le H \le \zeta(b-a)$ and $\tau^2 H \asymp \tau^2$. Since $(a+b-1)/a > 2$ and $\tau \le 1$, (K.3) gives $T \le C\tau^{(a+b-1)/a} \le C\tau^2$. Both sides of (K.1) are $\asymp G\tau^2$.

Case $b = a+1$. Here $q = -1$ and Lemma K.7 (c) gives $\frac{1}{2}\log(eM) \le H \le 2\log(eM)$. Now $\log(ei_\tau) = 1 + \frac{1}{a}\log(1/\tau)$ lies between $\frac{1}{a}\log(e/\tau)$ and $\log(e/\tau)$ (as $a > 1$), while $\log(eM) \le \log(ei_\tau)$ and, using $\log(eM) \ge 1$ and (K.2), $\log(ei_\tau) \le \log(eM) + \log(2c_1^{-1/a}) \le \big(1 + \log(2c_1^{-1/a})\big)\log(eM)$. Hence $H \asymp \log(e/\tau)$ and $\tau^2 H \asymp \tau^2 \log(e/\tau)$. Finally $(a+b-1)/a = 2$ here, so (K.3) gives $T \le C\tau^2 \le C\tau^2 \log(e/\tau)$. Both sides of (K.1) are $\asymp G\tau^2 \log(e/\tau)$. □

Lemma K.9 (Student fixed point away from the interpolation boundary)**.** *Let $\lambda_i = i^{-a}$, $i \le d$, $a > 1$, let $n_{\mathrm{s}} < d$ with $\kappa := d/n_{\mathrm{s}} \ge \kappa_0$ for a fixed $\kappa_0 > 1$, and let $(\tau_{\mathrm{s}}, \Xi_{\mathrm{s}})$ be the ridgeless ($\lambda_{\mathrm{s}} = 0$) student channel parameters (C.5)–(C.6). Put $c_1 := \min\{1, (\kappa_0 - 1)/2\}$. Then, for every $n_{\mathrm{s}} \ge I_a$,*

(a) *$c_1\, d^{-a} \;\le\; \tau_{\mathrm{s}} \;\le\; I_a^a\, n_{\mathrm{s}}^{-a} \;\le\; 1$, and also $\tau_{\mathrm{s}} \ge c_1\, (3n_{\mathrm{s}})^{-a}$;*

(b) *$1 - \Xi_{\mathrm{s}} \;\ge\; c_2 \;:=\; \min\big\{2^{-(a+3)}(a+1)^{-1} I_a^{-1},\;\; c_1\kappa_0/(4(a+1))\big\} \;>\; 0$.*

Both constants depend only on (a, κ_0); in particular they are uniform over $\kappa \in [\kappa_0, \infty]$, so the sparse-data regime $n_{\mathrm{s}}/d \to 0$ and any fixed shape $n_{\mathrm{s}}/d \to 1/\bar{\kappa}_{\mathrm{s}} < 1$ are covered at once.

Proof. Write $N(\tau) := \sum_{i \le d} \lambda_i/(\lambda_i + \tau) = \sum_{i \le d} (1 + \tau i^a)^{-1}$, strictly decreasing, with $N(\tau_{\mathrm{s}}) = n_{\mathrm{s}}$ by (C.5), and $i_{\tau_{\mathrm{s}}} := \tau_{\mathrm{s}}^{-1/a}$.

(a). *Upper.* By Lemma K.1, $n_{\mathrm{s}} = N(\tau_{\mathrm{s}}) \le \sum_{i\ge 1}(1+\tau_{\mathrm{s}} i^a)^{-1} \le \int_0^\infty (1+\tau_{\mathrm{s}} x^a)^{-1} dx = I_a \tau_{\mathrm{s}}^{-1/a}$, so $\tau_{\mathrm{s}} \le (I_a/n_{\mathrm{s}})^a \le 1$ for $n_{\mathrm{s}} \ge I_a$.

Lower. Let $\varepsilon > 0$. If $\tau_{\mathrm{s}} \le \varepsilon d^{-a}$ then $\tau_{\mathrm{s}} i^a \le \varepsilon$ for every $i \le d$, so $N(\tau_{\mathrm{s}}) \ge d/(1+\varepsilon)$; since $N(\tau_{\mathrm{s}}) = n_{\mathrm{s}} \le d/\kappa_0$, this forces $1+\varepsilon \ge \kappa_0$. Taking the contrapositive at $\varepsilon = (\kappa_0 - 1)/2$ gives $\tau_{\mathrm{s}} > \frac{\kappa_0 - 1}{2} d^{-a} \ge c_1 d^{-a}$. *Lower in* n_{s}. Each $i \le \min(d, \lfloor i_{\tau_{\mathrm{s}}} \rfloor)$ contributes at least $\frac{1}{2}$ to $N(\tau_{\mathrm{s}})$, so $\min(d, \lfloor i_{\tau_{\mathrm{s}}} \rfloor) \le 2n_{\mathrm{s}}$. If $\lfloor i_{\tau_{\mathrm{s}}} \rfloor \le d$ this gives $i_{\tau_{\mathrm{s}}} < 2n_{\mathrm{s}} + 1 \le 3n_{\mathrm{s}}$, i.e. $\tau_{\mathrm{s}} > (3n_{\mathrm{s}})^{-a}$; otherwise $d \le 2n_{\mathrm{s}}$ and $\tau_{\mathrm{s}} > c_1 d^{-a} \ge c_1 (3n_{\mathrm{s}})^{-a}$.

(b). Write $F_i := \lambda_i/(\lambda_i + \tau_{\mathrm{s}}) = (1+\tau_{\mathrm{s}} i^a)^{-1}$. The fixed point reads $\sum_{i\le d} F_i = n_{\mathrm{s}}$, so

$$1 - \Xi_{\mathrm{s}} \;=\; \frac{1}{n_{\mathrm{s}}}\Big(\sum_{i\le d} F_i - \sum_{i\le d} F_i^2\Big) \;=\; \frac{1}{n_{\mathrm{s}}} \sum_{i\le d} F_i\,(1-F_i). \tag{K.4}$$

Let $M := \min(\lfloor i_{\tau_{\mathrm{s}}} \rfloor, d) \ge 1$ (legitimate since $i_{\tau_{\mathrm{s}}} \ge 1$ by (a)). For $i \le M$ one has $\tau_{\mathrm{s}} i^a \le 1$, hence $F_i \ge \frac{1}{2}$ and $1 - F_i = \tau_{\mathrm{s}} i^a/(1+\tau_{\mathrm{s}} i^a) \ge \frac{1}{2}\tau_{\mathrm{s}} i^a$. Therefore, using $\sum_{i\le M} i^a \ge M^{a+1}/(a+1)$,

$$\sum_{i\le d} F_i(1-F_i) \;\ge\; \tfrac{1}{4}\,\tau_{\mathrm{s}} \sum_{i\le M} i^a \;\ge\; \frac{\tau_{\mathrm{s}}\, M^{a+1}}{4(a+1)}.$$

If $\lfloor i_{\tau_{\mathrm{s}}} \rfloor \le d$ then $M = \lfloor i_{\tau_{\mathrm{s}}} \rfloor \ge i_{\tau_{\mathrm{s}}}/2$, so $\tau_{\mathrm{s}} M^{a+1} \ge 2^{-(a+1)} \tau_{\mathrm{s}} i_{\tau_{\mathrm{s}}}^{a+1} = 2^{-(a+1)} i_{\tau_{\mathrm{s}}} \ge 2^{-(a+1)} n_{\mathrm{s}}/I_a$ by the upper bound in (a), and (K.4) gives $1 - \Xi_{\mathrm{s}} \ge 2^{-(a+3)}(a+1)^{-1} I_a^{-1}$. If instead $\lfloor i_{\tau_{\mathrm{s}}} \rfloor > d$ then $M = d$ and $\tau_{\mathrm{s}} d^{a+1} \ge c_1 d$ by (a), whence $1 - \Xi_{\mathrm{s}} \ge c_1 d/(4(a+1)n_{\mathrm{s}}) = c_1 \kappa/(4(a+1)) \ge c_1 \kappa_0/(4(a+1))$. □

Lemma K.10 (Ridge balance). *Let $a > 1$, $b \ge 0$, $a + b > 1$, let Φ_{ab} be as in Lemma K.8, put*

$$\gamma := \min\{(a+b-1)/a,\ 2\}, \qquad \mu := a\gamma + 1 \quad (\textit{so } \mu = a+b \textit{ if } b < a+1,\ \mu = 2a+1 \textit{ if } b \ge a+1),$$

and $Q(A,B) := A^{1/\mu} B^{1-1/\mu}$ for $A, B > 0$. Then

(a) *for any $\gamma, \delta > 0$,*

$$\inf_{\tau>0} \big[A\tau^{\gamma} + B\tau^{-\delta}\big] = \kappa\, A^{\delta/(\gamma+\delta)} B^{\gamma/(\gamma+\delta)}$$

with $\kappa = (\delta/\gamma)^{\gamma/(\gamma+\delta)} + (\gamma/\delta)^{\delta/(\gamma+\delta)}$;

(b) *if $b \ne a+1$ then $\inf_{0<\tau\le 1} \big[A\Phi_{ab}(\tau) + B\tau^{-1/a}\big] \;\ge\; \kappa_{a,b}\, Q(A,B)$, $\kappa_{a,b} > 0$;*

(c) *if $b = a+1$ and $A \ge B$ then, with $\ell := 1 + \log(A/B) \ge 1$ and $\tau^\dagger := (B/A)^{a/(2a+1)} \ell^{-a/(2a+1)} \in (0,1]$,*

$$\frac{a}{2(2a+1)}\, Q(A,B)\, \ell^{1/(2a+1)} \;\le\; \inf_{0<\tau\le 1} \big[A\Phi_{ab}(\tau) + B\tau^{-1/a}\big] \;\le\; 3\, Q(A,B)\, \ell^{1/(2a+1)},$$

the upper bound being attained at $\tau = \tau^\dagger$.

Proof. (a) The function $g(\tau) = A\tau^{\gamma} + B\tau^{-\delta}$ is strictly convex in $\log\tau$ with $g'(\tau) = 0$ at $\tau^* = (\delta B/(\gamma A))^{1/(\gamma+\delta)}$; substituting gives the displayed value. (b) Apply (a) with $\delta = 1/a$ and γ as defined, noting $\Phi_{ab}(\tau) = \tau^\gamma$ for $b \ne a+1$ and that the infimum over $(0,1]$ is at least the infimum over $(0,\infty)$; the exponents are $\delta/(\gamma+\delta) = 1/(a\gamma+1) = 1/\mu$ and $\gamma/(\gamma+\delta) = 1 - 1/\mu$. (c) Here $\gamma = 2$, $\mu = 2a+1$.

Put $\tau_3 := (B/A)^{a/(2a+1)} \le 1$ (using $A \ge B$), so that

$$A\tau_3^2 \;=\; B\tau_3^{-1/a} \;=\; Q := A^{1/(2a+1)} B^{2a/(2a+1)}, \tag{K.5}$$

and note $\log(e/\tau_3) = 1 + \frac{a}{2a+1}\log(A/B) = 1 + \frac{a}{2a+1}(\ell - 1) \ge \frac{a}{2a+1}\ell$. Set $\tau^{**} := \tau_3 \ell^{-a/(2a+1)} \le \tau_3$ and $\tau_4 := \tau_3 \ell^{1/(2(2a+1))}$. Four cases exhaust $(0,1]$; in each, one of the two terms alone already exceeds $\frac{a}{2(2a+1)} Q \ell^{1/(2a+1)}$.

(i) $\tau \le \tau^{**}$**:** $B\tau^{-1/a} \ge B(\tau^{**})^{-1/a} = B\tau_3^{-1/a}\ell^{1/(2a+1)} = Q\ell^{1/(2a+1)}$ by (K.5).

(ii) $\tau^{**} < \tau \le \tau_3$**:** $\log(e/\tau) \ge \log(e/\tau_3) \ge \frac{a}{2a+1}\ell$ and $A\tau^2 \ge A(\tau^{**})^2 = Q\ell^{-2a/(2a+1)}$, so $A\tau^2\log(e/\tau) \ge \frac{a}{2a+1}Q\ell^{1/(2a+1)}$.

(iii) $\tau_3 < \tau \le \min(\tau_4, 1)$**:**

$$\log(e/\tau) \ge \log(e/\tau_4) = \log(e/\tau_3) - \tfrac{\log \ell}{2(2a+1)} \ge \tfrac{a}{2a+1}\ell - \tfrac{\log \ell}{2(2a+1)} \ge \tfrac{a}{2(2a+1)}\ell,$$

the last step because $\log \ell \le \ell - 1 < a\ell$ for $\ell \ge 1$, $a > 1$; and $A\tau^2 \ge A\tau_3^2 = Q$. Hence $A\tau^2\log(e/\tau) \ge \frac{a}{2(2a+1)}Q\ell \ge \frac{a}{2(2a+1)}Q\ell^{1/(2a+1)}$.

(iv) $\tau_4 < \tau \le 1$**:** $\log(e/\tau) \ge 1$ and $A\tau^2 \ge A\tau_4^2 = Q\ell^{1/(2a+1)}$. Since $\frac{a}{2(2a+1)} < 1$, the four bounds combine to the lower half.

Upper half. Evaluate at $\tau^\dagger = \tau_3\ell^{-a/(2a+1)}$, which lies in $(0, \tau_3] \subseteq (0, 1]$ since $\ell \ge 1$. By (K.5), $B(\tau^\dagger)^{-1/a} = B\tau_3^{-1/a}\ell^{1/(2a+1)} = Q\ell^{1/(2a+1)}$ and $A(\tau^\dagger)^2 = A\tau_3^2\ell^{-2a/(2a+1)} = Q\ell^{-2a/(2a+1)}$. For the logarithm, $\log(e/\tau^\dagger) = \log(e/\tau_3) + \frac{a}{2a+1}\log\ell$, where $\log(e/\tau_3) = 1 + \frac{a}{2a+1}(\ell - 1) \le 1 + \frac{1}{2}(\ell - 1) \le \ell$ (using $\frac{a}{2a+1} < \frac{1}{2}$ and $\ell \ge 1$) and $\frac{a}{2a+1}\log\ell \le \log\ell \le \ell - 1$; hence $\log(e/\tau^\dagger) \le 2\ell$ and

$$A\Phi_{ab}(\tau^\dagger) \;=\; A(\tau^\dagger)^2\log(e/\tau^\dagger) \;\le\; 2\,Q\,\ell^{\,1-2a/(2a+1)} \;=\; 2\,Q\,\ell^{1/(2a+1)}.$$

Adding the two terms gives $A\Phi_{ab}(\tau^\dagger) + B(\tau^\dagger)^{-1/a} \le 3Q\ell^{1/(2a+1)}$, which is the upper half. □

K.2 (I) The imitation rate

The pure-imitation limit is the student stage with an exact teacher $\beta_{\mathrm{w}} = w^*$ (equivalently $\psi \to 1$, $n_{\mathrm{w}} \to \infty$, so $F_i^{\mathrm{w}} \to 1$, $v_i^{\mathrm{w}} \to 0$, $A_m \to 0$): a single noiseless MNI/ridgeless stage with target ω^2 and zero noise. Since $1 - F_i^{\mathrm{s}} = \tau_{\mathrm{s}}/(\lambda_i + \tau_{\mathrm{s}})$, the bias sum is exactly $B_0(\tau_{\mathrm{s}})$, and with $\rho_i = \omega_i^2$, $s^2 = 0$ the self-noise is $\sum_i \lambda_i v_i^{\mathrm{s}} = B_0\,\Xi_{\mathrm{s}}/(1 - \Xi_{\mathrm{s}})$. Hence the elegant collapse

$$\bar{\mathcal{E}}_s = B_0(\tau_{\mathrm{s}}) + B_0(\tau_{\mathrm{s}})\frac{\Xi_{\mathrm{s}}}{1 - \Xi_{\mathrm{s}}} = \frac{B_0(\tau_{\mathrm{s}})}{1 - \Xi_{\mathrm{s}}}\,: \tag{K.6}$$

with zero noise the entire risk is the bias sum inflated by a single scalar. The self-noise is exactly proportional to B_0, so it multiplies the rate by a constant and never changes the exponent.

Proof of part (i). Standing hypothesis — the only one beyond the power law, and the one the statement of Theorem D.14 (i) carries: the student's overparametrization ratio stays away from the interpolation boundary,

$$n_{\mathrm{s}} \;<\; d, \qquad \kappa \;:=\; d/n_{\mathrm{s}} \;\ge\; \kappa_0 \;>\; 1, \tag{K.7}$$

with κ_0 fixed and *no* upper bound on κ: the sparse-data regime $n_{\mathrm{s}}/d \to 0$ and any fixed shape $n_{\mathrm{s}}/d \to 1/\bar{\kappa_{\mathrm{s}}} < 1$ are both instances. This is exactly Assumption C.1's clause "$\kappa_{\mathrm{s}} \to \bar{\kappa_{\mathrm{s}}} > 1$, bounded away from 1", extended to allow $\bar{\kappa_{\mathrm{s}}} = \infty$. It is not technical: Remark K.12 shows the exponent genuinely changes without it. All constants below depend only on (a, b, r^2, κ_0).

The two-sided rate. By (K.6), $\bar{\mathcal{E}}_s = B_0^{(d)}(\tau_{\mathrm{s}})/(1 - \Xi_{\mathrm{s}})$, with $B_0^{(d)}$ the *truncated* bias sum of Lemma K.8 at $p = d$: the student's spectrum has exactly d coordinates, and it is precisely the interplay between d and the effective dimension $\tau_{\mathrm{s}}^{-1/a}$ that (K.7) controls. By Lemma K.9 (b), $c_2 \le 1 - \Xi_{\mathrm{s}} \le 1$, so

$$B_0^{(d)}(\tau_{\mathrm{s}}) \;\le\; \bar{\mathcal{E}}_s \;\le\; c_2^{-1}\,B_0^{(d)}(\tau_{\mathrm{s}}),$$

and it suffices to locate $B_0^{(d)}(\tau_{\mathrm{s}})$. Lemma K.9 (a) supplies exactly what Lemma K.8 requires at $p = d$, namely $\tau_{\mathrm{s}} \ge c_1 d^{-a}$ and $\tau_{\mathrm{s}} \le 1$; hence $B_0^{(d)}(\tau_{\mathrm{s}}) \asymp G\,\Phi_{ab}(\tau_{\mathrm{s}})$.

The same lemma gives $c_1(3n_{\mathrm{s}})^{-a} \le \tau_{\mathrm{s}} \le I_a^a n_{\mathrm{s}}^{-a}$, i.e. $\tau_{\mathrm{s}} = \vartheta\, n_{\mathrm{s}}^{-a}$ with ϑ confined to a fixed compact subset of $(0, \infty)$. Since $\Phi_{ab}(\vartheta\tau) \asymp \Phi_{ab}(\tau)$ uniformly over such ϑ — immediate for the two pure

powers, and at $b = a + 1$ because $\log(e/(\vartheta\tau)) = \log(e/\tau) - \log\vartheta$ is within a factor 2 of $\log(e/\tau)$ as soon as $\log(e/\tau) \geq 2|\log\vartheta|$, i.e. for $n_s \geq n_0(a, \kappa_0)$ — we may replace τ_s by n_s^{-a}:

$$\bar{\mathcal{E}}_s(n_s) \asymp G\,\Phi_{ab}(n_s^{-a}) \asymp \begin{cases} n_s^{-(a+b-1)}, & b < a+1, \\ n_s^{-2a}\log n_s, & b = a+1, \\ n_s^{-2a}, & b > a+1, \end{cases}$$

i.e. $\bar{\mathcal{E}}_s(n_s) \asymp n_s^{-\min(a+b-1,\,2a)}$ with the stated boundary logarithm, since $a+b-1 < 2a \iff b < a+1$. The constants are uniform over $\kappa \in [\kappa_0, \infty]$.

Exact constants in the sparse-data limit. When moreover $n_s/d \to 0$ the constants can be evaluated. First, truncation is then immaterial: $0 \leq B_0(\tau_s) - B_0^{(d)}(\tau_s) \leq \sum_{i>d}\lambda_i\omega_i^2 \leq \frac{G}{a+b-1}d^{1-a-b}$, which is $o(n_s^{-\min(a+b-1,2a)})$ because $d \gg n_s$ and $a+b-1 \geq \min(a+b-1, 2a)$; so $B_0^{(d)}(\tau_s) = B_0(\tau_s)(1+o(1))$ with B_0 the full sum of Lemma K.6. By Lemma K.3, $\tau_s = c_a n_s^{-a}(1+o(1))$ and $1-\Xi_s \to 1/a$, so $1/(1-\Xi_s) \to a$; and by Lemma K.6 at $\tau = \tau_s \to 0$,

$$B_0 = \begin{cases} G\,K_{ab}\,c_a^{(a+b-1)/a}\,n_s^{-(a+b-1)}(1+o(1)), & b < a+1, \\ G\,c_a^2\,n_s^{-2a}\log n_s\,(1+o(1)), & b = a+1, \\ G\,\zeta(b-a)\,c_a^2\,n_s^{-2a}(1+o(1)), & b > a+1, \end{cases}$$

(using $\log(1/\tau_s) = a\log n_s - \log c_a \sim a\log n_s$). Multiplying by $1/(1-\Xi_s) \to a$ gives $\bar{\mathcal{E}}_s = (1+o(1))\,a\,G\,C(a,b)\,n_s^{-\min(a+b-1,2a)}\,L(n_s)$ with $(C, L) = (K_{ab}c_a^{(a+b-1)/a}, 1)$, $(c_a^2, \log n_s)$, $(\zeta(b-a)c_a^2, 1)$ in the three cases. □

Remark K.11 (Boundary honesty at $b = a+1$). At $b = a+1$ the head sum $\sum_{i\leq i^*} i^{-1}$ diverges logarithmically, so the risk is $n_s^{-2a}\log n_s$, strictly larger than n_s^{-2a} by a logarithm — stated, not swept. (Numerically, fitted log-log slopes shallower than $-2a$ by exactly the amount an $n^{-2a}\log n$ law predicts.)

Remark K.12 (Hypothesis (K.7) is necessary: at $n_s/d \to 1$ the exponent is strictly larger). Without a bound $\kappa \geq \kappa_0 > 1$ the rate of part (i) is false, and by a full power of n_s. Take one null direction, $d = n_s + 1$, and $b < a+1$; constants below depend only on (a, b, r^2). Then for $n_s > 2(a+1)$

$$\bar{\mathcal{E}}_s(n_s) \asymp n_s^{-(a+b)}, \tag{K.8}$$

strictly faster than the $n_s^{-(a+b-1)}$ of part (i). *Proof.* Write $F_i = (1+\tau_s i^a)^{-1}$; the fixed point $\sum_{i\leq d}F_i = n_s = d-1$ reads $\sum_{i\leq d}(1-F_i) = 1$.

(1) $\tau_s d^a < 1$. Otherwise $1 - F_i \geq \frac{1}{2}(i/d)^a$ for all $i \leq d$, so $1 = \sum_i(1-F_i) \geq \frac{1}{2d^a}\sum_{i\leq d} i^a \geq \frac{d}{2(a+1)} > 1$ for $d > 2(a+1)$ — a contradiction.

In particular $F_i > \frac{1}{2}$ for every $i \leq d$.

(2) $\tau_s \asymp d^{-(a+1)}$. From $1 - F_i \leq \tau_s i^a$ and $\sum_{i\leq d} i^a \leq 2^{a+1}d^{a+1}/(a+1)$ we get $\tau_s \geq (a+1)2^{-(a+1)}d^{-(a+1)}$; from (1), $1 - F_i \geq \frac{1}{2}\tau_s i^a$ and $\sum_{i\leq d} i^a \geq d^{a+1}/(a+1)$ give $\tau_s \leq 2(a+1)d^{-(a+1)}$.

(3) $1 - \Xi_s \asymp 1/n_s$. By (K.4), $1 - \Xi_s = n_s^{-1}\sum_i F_i(1-F_i)$, and $\frac{1}{2} < F_i \leq 1$ with $\sum_i(1-F_i) = 1$ puts this in $[\frac{1}{2n_s}, \frac{1}{n_s}]$.

(4) $B_0^{(d)}(\tau_s) \asymp G\tau_s^2 d^{a-b+1}$. By (1), $\tau_s i^a < 1$ on $i \leq d$, so $(1+\tau_s i^a)^{-2} \in [\frac{1}{4}, 1]$ and $B_0^{(d)}(\tau_s) \asymp G\tau_s^2\sum_{i\leq d} i^{a-b} \asymp G\tau_s^2 d^{a-b+1}$ by Lemma K.7 (a), $a - b > -1$. Now (K.6) with (2), (3), (4) and $d = n_s + 1 \asymp n_s$ gives $\bar{\mathcal{E}}_s \asymp n_s \cdot Gd^{-2(a+1)}d^{a-b+1} \asymp G\,n_s^{-(a+b)}$, which is (K.8). □

The mechanism is visible in (2)–(3): as $\kappa \downarrow 1$ the student's effective dimension $\tau_s^{-1/a}$ overshoots d, so the truncation of the spectrum at d — not the τ-rolloff — decides the bias sum, while $1 - \Xi_s$ collapses at rate $\kappa - 1$. Both effects are absent as soon as κ is bounded away from 1, which is what Lemma K.9 quantifies. (Numerically at $(a, b) = (2, 1)$, $r^2 = 1$: $\bar{\mathcal{E}}_s\,n_s^3 = 3.677, 3.710, 3.727, 3.735, 3.739$ at

$n_s = 199, 399, 799, 1599, 3199$, against the constant $G(a+1)^2/(a-b+1) = 3.744$ that (2)–(4) predict in the limit, while $\bar{\mathcal{E}}_s\, n_s^2$ — the part-(i) normalization — halves at each step. Corroboration, not a proof step.)

K.3 (II) The noise floor

Fix ratios $(\psi, \bar{\kappa_w}, \bar{\kappa_s})$ and $\sigma^2 > 0$; let $d \to \infty$ with $m = \psi d$, $n_w = m/\bar{\kappa_w}$, $n_s = d/\bar{\kappa_s}$. Two facts remove the naive decay sources. *(a)* $A_m = G\sum_{i>\psi d} i^{-(a+b)} \sim G(\psi d)^{1-a-b}/(a+b-1) \to 0$ (since $a+b>1$), so $s_w^2 \to \sigma^2$: the floor is not an approximation-error artifact. *(b)* Setting $\sigma^2 = 0$ makes both stages noiseless, and by the part-(i) mechanism their bias sums decay to 0; so the only surviving contribution is noise-linear.

$\bar{\mathcal{E}}_s$ is exactly affine in σ^2. σ^2 enters (D.2) only through $s_w^2 = \sigma^2 + A_m$, linearly into Γ_w, hence into v^w, ρ, Γ_s, v^s; the shrinkages and fixed points $\tau_w, \tau_s, \Xi_w, \Xi_s$ are noise-independent. So $\bar{\mathcal{E}}_s = \bar{\mathcal{E}}_s|_{\sigma^2=0} + \sigma^2\,(d\bar{\mathcal{E}}_s/d\sigma^2)$ with vanishing intercept by (b), and the floor is $\sigma^2 \lim(d\bar{\mathcal{E}}_s/d\sigma^2)$.

The noise-transfer coefficient. Differentiating (D.2) in σ^2 (only v^w, v^s depend on it), with $vp_i^w := \lambda_i/((\lambda_i+\tau_w)^2 n_w(1-\Xi_w))$ for $i \le m$, the two resulting terms combine per coordinate into the same student transmission coefficient found in Lemma G.1:

$$\frac{d\bar{\mathcal{E}}_s}{d\sigma^2} = \sum_i vp_i^w \cdot \lambda_i \Big[(F_i^s)^2 + \frac{\Xi_s}{1-\Xi_s}(1-F_i^s)^2\Big]. \tag{K.9}$$

No signal (ω^2, hence b) appears — the source of the b-independence.

Proof of part (ii). Pass (K.9) to the proportional limit via Lemma K.4 with the student-scale variable $x = i/d \in (0,1]$, teacher support $x \in (0,\psi]$: writing $F_s(x) = 1/(1+\theta_s x^a)$ and $\phi_w(x) = 1/(1+\theta_w(x/\psi)^a)$ with $\theta_s = \theta(a,\bar{\kappa_s})$, $\theta_w = \theta(a,\bar{\kappa_w})$, and using $\frac{1}{d}\sum_{i\le m} f(i/d) \to \int_0^\psi f$ with $n_w = \psi d/\bar{\kappa_w}$,

$$\frac{d\bar{\mathcal{E}}_s}{d\sigma^2} \to \frac{\bar{\kappa_w}}{\psi(1-\Xi_w)} \int_0^\psi \phi_w(x)^2 \Big[F_s(x)^2 + \frac{\Xi_s}{1-\Xi_s}(1-F_s(x))^2\Big] dx =: C_{\text{floor}},$$

with $\Xi_w = \Xi_{\text{prop}}(a,\bar{\kappa_w})$, $\Xi_s = \Xi_{\text{prop}}(a,\bar{\kappa_s})$. Every factor is strictly positive, so $C_{\text{floor}} > 0$ and the floor $\sigma^2 C_{\text{floor}} > 0$. Combined with the affine structure, $\bar{\mathcal{E}}_s \to \sigma^2 C_{\text{floor}}(a,\psi,\bar{\kappa_w},\bar{\kappa_s})$. □

Remark K.13 (Mechanism, κ-dependence, and the clean surprise). The floor is nonzero precisely because Ξ_w, Ξ_s tend to constants in $(0,1)$ (Lemma K.4): the single-stage noise variance $V = \sigma^2\Xi/(1-\Xi)$ stays bounded away from 0 as $n \to \infty$, because interpolation refuses to average label noise away. The floor *does* depend on $(\bar{\kappa_w}, \bar{\kappa_s}, \psi)$ (through θ, Ξ and the integration limit); the tempting "floor $= \sigma^2 f(a)$" holds only in the $\kappa \to \infty$ edge. The clean, reportable fact is instead the *exact b-independence*: the floor is a property of the covariance decay a and the sampling ratios alone; the signal geometry b sets only the vanishing imitation part.

The approach to the floor is *not* one-sided, and saying otherwise would be false on a whole b-range. At $\psi = 0.3$, $\bar{\kappa_w} = 3$, $\bar{\kappa_s} = 2$, $\sigma^2 = 1/4$, $r^2 = 1$, along $d = 10^3, 4\cdot10^3, 1.6\cdot10^4, 6.4\cdot10^4, 2.56\cdot10^5, 1.024\cdot10^6$ the value of $\bar{\mathcal{E}}_s$ *decreases* to the floor at $(a,b) = (1.5, 0)$ ($0.363274 \to 0.255721$) and at $(1.5, 0.25)$ ($0.291015 \to 0.252467$) — neither an edge case — while it *increases* to it at $(1.5, 3)$ ($0.250123 \to 0.252240$), and at $(1.5, 1)$ it dips first and rises after (0.252005, 0.251946, 0.252138, …, 0.252240). The limit is the theorem; the side from which it is approached is not part of it.

K.4 (III) No scaling penalty for weak supervision

Supervised optimal-ridge baseline. For a single ridge stage $(\Sigma, n, \omega^2, \sigma^2)$ parametrized by τ: by Lemma K.6 $B_0(\tau) \sim \tau^{(a+b-1)/a}$ ($b < a+1$), and since the optimal τ far exceeds the ridgeless value, $\Xi \to 0$ so $V \sim \sigma^2 J_a \tau^{-1/a}/n$. Minimizing $\text{Risk}(\tau) \sim C_1\tau^{(a+b-1)/a} + C_2\sigma^2\tau^{-1/a}/n$ gives $\tau^* \asymp (\sigma^2/n)^{a/(a+b)}$ and

$$\text{Risk}(\tau^*) \asymp (\sigma^2/n)^{(a+b-1)/(a+b)} \asymp n^{-(a+b-1)/(a+b)}, \tag{K.10}$$

the classical minimax rate for power-law (capacity a, source b) regression. In the well-specified band $1 < b < a+1$ it is the optimal rate of Caponnetto & De Vito (2007), in our symbols as follows: their capacity exponent is our a (their eigenvalue condition reads $\lambda_n \asymp n^{-a}$), and their source parameter c admits the profile $\omega_n^2 \propto n^{-b}$ for every $c < (a+b-1)/a$, their rate exponent $ac/(ac+1)$ increasing to our $(a+b-1)/(a+b)$ as c rises to that value; their qualification ceiling $c \le 2$ is exactly our saturation boundary $b \le a+1$.

Their class also requires $c \ge 1$, i.e. $b \ge 1$, so that attribution covers $1 < b < a+1$ and no more — the scoping clause is not optional. For $b < 1$ the same exponent is attained, with matching upper and lower bounds for one-pass sketched SGD, by Lin et al. (2024). For $b > a+1$ the bias saturates ($B_0 \asymp \tau^2$, Lemma K.8) and the optimum gives $n^{-2a/(2a+1)}$; at $b = a+1$ it saturates only up to a logarithm ($B_0 \asymp \tau^2 \log(e/\tau)$) and the optimum gives $n^{-2a/(2a+1)}(\log n)^{1/(2a+1)}$ — both by Lemma K.10.

Upper bound: the pipeline achieves the rate. The referee-facing point of this half is that "the student is a near-clone" must be turned into a quantitative statement about the *whole* teacher band $i \le m$ — that is where the teacher's fluctuation profile v^{w} lives, and hence where the student's self-noise $\sum_i \lambda_i v_i^{\mathrm{s}}$ is fed from. It is, and by a wide margin; the mechanism is (K.14) below.

Proposition K.14 (The tuned pipeline attains the supervised exponent, in all three bands)**.** *Fix $a > 1$, $b \ge 0$, $\sigma^2 > 0$, $r^2 > 0$, and let the resources scale as Theorem D.14 (iii) prescribes: as $n_{\mathrm{w}} \to \infty$,*

$$\textit{(R1)}\ 0 < c_m^- \le m/n_{\mathrm{w}} \le c_m^+ < \infty, \qquad \textit{(R2)}\ d/n_{\mathrm{s}} \to \bar{\kappa_{\mathrm{s}}} \in (1,\infty), \qquad \textit{(R3)}\ n_{\mathrm{s}}/n_{\mathrm{w}} \to \infty. \tag{K.11}$$

Put

$$\mathcal{R}(n_{\mathrm{w}}) \ := \ \begin{cases} n_{\mathrm{w}}^{-(a+b-1)/(a+b)}, & b < a+1, \\ n_{\mathrm{w}}^{-2a/(2a+1)} (\log n_{\mathrm{w}})^{1/(2a+1)}, & b = a+1, \\ n_{\mathrm{w}}^{-2a/(2a+1)}, & b > a+1. \end{cases}$$

Then there are $C < \infty$, $n_0 < \infty$ depending only on $(a, b, \sigma^2, r^2, c_m^-, c_m^+, \bar{\kappa_{\mathrm{s}}})$ with

$$\bar{\mathcal{E}}_s^*(n_{\mathrm{w}}) \ := \ \inf_{\lambda_{\mathrm{w}} \ge 0,\ \lambda_{\mathrm{s}} \ge 0} \bar{\mathcal{E}}_s \ \le \ C\,\mathcal{R}(n_{\mathrm{w}}) \qquad (n_{\mathrm{w}} \ge n_0).$$

Proof. **Step 0 (the choice of ridges, and that it is admissible).** Set

$$\tau_{\mathrm{w}} \ := \ \begin{cases} (\sigma^2/n_{\mathrm{w}})^{a/(a+b)}, & b < a+1, \\ (\sigma^2/(n_{\mathrm{w}} \log n_{\mathrm{w}}))^{a/(2a+1)}, & b = a+1, \\ (\sigma^2/n_{\mathrm{w}})^{a/(2a+1)}, & b > a+1, \end{cases} \qquad \iota := \tau_{\mathrm{w}}^{-1/a},$$

and take the *minimum-norm* student, $\lambda_{\mathrm{s}} = 0$ (legitimate since $n_{\mathrm{s}} < d$ by (R2)).

A direct computation from these definitions gives, for $n_{\mathrm{w}} \ge n_0(a, b, \sigma^2)$,

$$\Phi_{ab}(\tau_{\mathrm{w}}) \ \asymp \ \mathcal{R}(n_{\mathrm{w}}), \qquad \iota/n_{\mathrm{w}} \ \asymp \ \mathcal{R}(n_{\mathrm{w}}), \qquad \iota \ = \ o(n_{\mathrm{w}}), \tag{K.12}$$

with constants depending only on (a, b, σ^2). The three rate calculations are as follows.

Below saturation ($b < a+1$):

$$\Phi_{ab}(\tau_{\mathrm{w}}) = \tau_{\mathrm{w}}^{(a+b-1)/a} = (\sigma^2/n_{\mathrm{w}})^{(a+b-1)/(a+b)}, \qquad \frac{\iota}{n_{\mathrm{w}}} = \sigma^{-2/(a+b)} n_{\mathrm{w}}^{-(a+b-1)/(a+b)}.$$

Above saturation ($b > a+1$):

$$\Phi_{ab}(\tau_{\mathrm{w}}) = \tau_{\mathrm{w}}^2 = (\sigma^2/n_{\mathrm{w}})^{2a/(2a+1)}, \qquad \frac{\iota}{n_{\mathrm{w}}} = \sigma^{-2/(2a+1)} n_{\mathrm{w}}^{-2a/(2a+1)}.$$

At saturation ($b = a+1$): write $\ell := \log n_{\mathrm{w}}$. Then

$$\frac{\iota}{n_{\mathrm{w}}} = \sigma^{-2/(2a+1)} n_{\mathrm{w}}^{-2a/(2a+1)} \ell^{1/(2a+1)}.$$

Moreover,

$$\log(e/\tau_{\mathrm{w}}) = 1 + \frac{a}{2a+1} \log(n_{\mathrm{w}} \ell/\sigma^2)$$

lies between $\frac{a}{2a+1}\ell$ and $C(a,\sigma^2)\ell$ once $\ell \geq \sigma^2$. Consequently,

$$\Phi_{ab}(\tau_{\mathrm{w}}) = \tau_{\mathrm{w}}^2 \log(e/\tau_{\mathrm{w}}) \asymp \sigma^{4a/(2a+1)} n_{\mathrm{w}}^{-2a/(2a+1)} \ell^{1/(2a+1)}.$$

In each band $\iota/n_{\mathrm{w}} \asymp \mathcal{R}(n_{\mathrm{w}}) \to 0$, whence $\iota = o(n_{\mathrm{w}})$.

Admissibility. By (C.5) the map $\tau \mapsto \tau(1 - \frac{1}{n_{\mathrm{w}}} \sum_{i \leq m} \frac{\lambda_i}{\lambda_i + \tau})$ has derivative $1 - \Xi_{\mathrm{w}}(\tau) > 0$, so it is a continuous strictly increasing bijection onto its range and prescribing τ_{w} and reading off λ_{w} is legitimate; and by Lemma K.1, $\sum_{i \leq m} \lambda_i/(\lambda_i + \tau_{\mathrm{w}}) \leq I_a \iota = o(n_{\mathrm{w}})$, so the resulting $\lambda_{\mathrm{w}} = \tau_{\mathrm{w}}(1 - o(1)) > 0$. In particular the choice is inside the admissible family, so it bounds the infimum.

Step 1: control the quantities entering the risk.

(1a) Omitted tail and effective noise. By Lemma K.7 (d), $A_m = G \sum_{i>m} i^{-(a+b)} \leq \frac{G}{a+b-1} m^{-(a+b-1)} \leq \frac{G}{a+b-1} (c_m^- n_{\mathrm{w}})^{-(a+b-1)}$ by (R1), so $\sigma^2 \leq s_{\mathrm{w}}^2 = \sigma^2 + A_m \leq 2\sigma^2$ for $n_{\mathrm{w}} \geq n_0$; moreover $A_m = O(n_{\mathrm{w}}^{-(a+b-1)}) = o(\mathcal{R}(n_{\mathrm{w}}))$, because $a+b-1$ strictly exceeds the exponent of n_{w} in $\mathcal{R}$ (for $b < a+1$ that exponent is $(a+b-1)/(a+b) < a+b-1$ since $a+b > 1$; for $b \geq a+1$ it is $2a/(2a+1) < 1 \leq 2a \leq a+b-1$).

(1b) Teacher denominator. By Lemma K.1, $\Xi_{\mathrm{w}} = \frac{1}{n_{\mathrm{w}}} \sum_{i \leq m} (1 + \tau_{\mathrm{w}} i^a)^{-2} \leq J_a \iota / n_{\mathrm{w}} \to 0$, so $1 - \Xi_{\mathrm{w}} \geq \frac{1}{2}$ eventually.

(1c) Teacher bias and fluctuation level. Since $m \geq c_m^- n_{\mathrm{w}}$ by (R1), $\iota = o(n_{\mathrm{w}}) = o(m)$, so we have $\tau_{\mathrm{w}} \geq m^{-a}$ eventually, so Lemma K.8 applies at $p = m$ with $c_1 = 1$: $B_0^{(m)}(\tau_{\mathrm{w}}) \leq c_+ G \Phi_{ab}(\tau_{\mathrm{w}}) \to 0$. Combining,

$$\Gamma_{\mathrm{w}} n_{\mathrm{w}} \Xi_{\mathrm{w}} \;=\; \frac{\Xi_{\mathrm{w}}}{1 - \Xi_{\mathrm{w}}} \big(B_0^{(m)}(\tau_{\mathrm{w}}) + s_{\mathrm{w}}^2 \big) \;\leq\; 2 \frac{J_a \iota}{n_{\mathrm{w}}} \big(c_+ G \Phi_{ab}(\tau_{\mathrm{w}}) + 2\sigma^2 \big) \;\leq\; C_1 \mathcal{R}(n_{\mathrm{w}}) \quad \text{(K.13)}$$

by (K.12).

(1d) Student fixed point and denominator. By (R2) and Lemma K.9 applied with $\kappa_0 := (1 + \bar{\kappa_{\mathrm{s}}})/2 > 1$: $c_1 d^{-a} \leq \tau_{\mathrm{s}} \leq I_a^a n_{\mathrm{s}}^{-a} \leq 1$ and $1 - \Xi_{\mathrm{s}} \geq c_2 > 0$.

Step 2 (the near-clone bound, made quantitative). This is the step the one-line "near-clone" assertion elides. The teacher's fluctuation profile v^{w}, and with it the realized profile $\rho_i = (F_i^{\mathrm{w}})^2 \omega_i^2 + v_i^{\mathrm{w}}$ that drives Γ_{s}, vanishes for $i > m$. On the band that survives we have a *uniform* clone bound:

$$\max_{i \leq m} \big(1 - F_i^{\mathrm{s}}\big) \;=\; \max_{i \leq m} \frac{\tau_{\mathrm{s}}}{\lambda_i + \tau_{\mathrm{s}}} \;\leq\; \tau_{\mathrm{s}}\, m^a \;\leq\; I_a^a \Big(\frac{m}{n_{\mathrm{s}}}\Big)^a \;=:\; \eta \;\longrightarrow\; 0, \qquad \text{(K.14)}$$

using $\lambda_i \geq m^{-a}$ on $i \leq m$, then $\tau_{\mathrm{s}} \leq I_a^a n_{\mathrm{s}}^{-a}$ from (1d), and finally $m/n_{\mathrm{s}} = (m/n_{\mathrm{w}})(n_{\mathrm{w}}/n_{\mathrm{s}}) \leq c_m^+ (n_{\mathrm{w}}/n_{\mathrm{s}}) \to 0$ by (R1) and (R3). It is (K.14) — not "$F_i^{\mathrm{s}} \approx 1$ on the teacher-signal band $i \lesssim \iota$" — that controls the self-noise, because v_i^{w} is spread over all of $i \leq m$, and $m/\iota \to \infty$.

Step 3 (the three terms of (D.2)). Write T_1, T_2, T_3 for them in order.

T_1: for $i > m$ we have $F_i^{\mathrm{w}} = 0$ and the summand is $\lambda_i \omega_i^2$, contributing A_m; for $i \leq m$,

$$1 - F_i^{\mathrm{s}} F_i^{\mathrm{w}} = (1 - F_i^{\mathrm{w}}) + F_i^{\mathrm{w}}(1 - F_i^{\mathrm{s}}) \leq (1 - F_i^{\mathrm{w}}) + (1 - F_i^{\mathrm{s}}),$$

so by $(x+y)^2 \leq 2x^2 + 2y^2$,

$$T_1 \;\leq\; 2\, B_0^{(m)}(\tau_{\mathrm{w}}) \;+\; 2\, B_0^{(d)}(\tau_{\mathrm{s}}) \;+\; A_m.$$

T_2: since $F_i^{\mathrm{s}} \leq 1$,

$$T_2 \leq \sum_{i \leq m} \lambda_i v_i^{\mathrm{w}} = \Gamma_{\mathrm{w}} \sum_{i \leq m} (\lambda_i/(\lambda_i + \tau_{\mathrm{w}}))^2 = \Gamma_{\mathrm{w}} n_{\mathrm{w}} \Xi_{\mathrm{w}},$$

bounded by (K.13).

T_3: by (C.6) with zero student label noise,

$$T_3 \;=\; \Gamma_{\rm s}\sum_i \frac{\lambda_i^2}{(\lambda_i+\tau_{\rm s})^2} \;=\; \Gamma_{\rm s}\, n_{\rm s}\, \Xi_{\rm s} \;=\; \frac{\Xi_{\rm s}}{1-\Xi_{\rm s}}\sum_{i\le m}\lambda_i\rho_i\big(1-F_i^{\rm s}\big)^2,$$

the sum running over $i \le m$ only because $\rho_i = 0$ beyond. Split ρ_i: the signal part obeys

$$\sum_{i\le m}\lambda_i (F_i^{\rm w})^2\omega_i^2(1-F_i^{\rm s})^2 \le \sum_{i\le d}\lambda_i\omega_i^2(1-F_i^{\rm s})^2 = B_0^{(d)}(\tau_{\rm s}),$$

while the fluctuation part obeys, by (K.14), $\sum_{i\le m}\lambda_i v_i^{\rm w}(1-F_i^{\rm s})^2 \le \eta^2\,\Gamma_{\rm w} n_{\rm w}\Xi_{\rm w}$. With $\Xi_{\rm s}/(1-\Xi_{\rm s}) \le 1/c_2$ from (1d),

$$T_3 \;\le\; c_2^{-1}\Big[B_0^{(d)}(\tau_{\rm s}) \;+\; \eta^2\,\Gamma_{\rm w}\,n_{\rm w}\,\Xi_{\rm w}\Big].$$

Step 4 (collect). By (1c), $B_0^{(m)}(\tau_{\rm w}) \le c_+ G\Phi_{ab}(\tau_{\rm w}) \le C\mathcal{R}(n_{\rm w})$. By (1d), Lemma K.8 applies at $p = d$ with $\tau_{\rm s} \asymp n_{\rm s}^{-a}$, giving $B_0^{(d)}(\tau_{\rm s}) \le c_+ G\Phi_{ab}(\tau_{\rm s})$; each Φ_{ab} is increasing on $(0,1]$ (for $\tau^2\log(e/\tau)$ the derivative is $\tau(2\log(e/\tau)-1) > 0$ there), and $\Phi_{ab}(c\tau) \le c^2\Phi_{ab}(\tau)$ for $c \ge 1$, $\tau \le 1$; so with $\tau_{\rm s} \le I_a^a n_{\rm s}^{-a}$ and $n_{\rm s} \ge n_{\rm w}$ eventually,

$$\Phi_{ab}(\tau_{\rm s}) \;\le\; C\,\Phi_{ab}(n_{\rm w}^{-a}) \;\le\; C\cdot\begin{cases} n_{\rm w}^{-(a+b-1)}, & b < a+1,\\ n_{\rm w}^{-2a}\log(en_{\rm w}), & b = a+1,\\ n_{\rm w}^{-2a}, & b > a+1,\end{cases}$$

each of which is $o(\mathcal{R}(n_{\rm w}))$, since $\min(a+b-1, 2a)$ strictly exceeds the exponent of $n_{\rm w}$ in $\mathcal{R}$ (same comparison as in (1a); at $b = a+1$ the extra $\log(en_{\rm w})$ is absorbed because $2a - 2a/(2a+1) > 0$).

Finally $A_m = o(\mathcal{R}(n_{\rm w}))$ by (1a), $\Gamma_{\rm w} n_{\rm w}\Xi_{\rm w} \le C_1\mathcal{R}(n_{\rm w})$ by (K.13), and $\eta \to 0$ by (K.14). Summing $T_1 + T_2 + T_3$ gives $\mathcal{E}_s \le C\,\mathcal{R}(n_{\rm w})$, and the choice of Step 0 is admissible, so the same bounds $\mathcal{E}_s^*(n_{\rm w})$. □

Remark K.15 (Where each resource hypothesis is used). (R1) $m = \Theta(n_{\rm w})$ — a two-sided bound, not a limit: the proof nowhere uses convergence of $m/n_{\rm w}$ — is used twice: it makes the sketch's approximation error A_m lower order (Step 1a), and — on the lower side — it is what rules out the degenerate branch in which both ridges are so small that the teacher's effective dimension exceeds its capacity (Case B below). (R2) $d/n_{\rm s} \to \bar{\kappa_{\rm s}} > 1$ is the hypothesis of Lemma K.9: it keeps $1-\Xi_{\rm s}$ bounded below and pins $\tau_{\rm s} \asymp n_{\rm s}^{-a}$, and both uses need the bound away from 1, not merely $n_{\rm s} = \Theta(d)$.

We claim no *sharpness* for it. The mechanism that breaks part (i) at $n_{\rm s}/d \to 1$ (Remark K.12) is a *ridgeless* one — there the student sits at its MNI fixed point, where $\tau_{\rm s} \asymp d^{-(a+1)}$ and $1-\Xi_{\rm s} \asymp 1/n_{\rm s}$ collapse together — whereas (iii) tunes the ridge at both stages, and an admissible tuned ridge is free to hold $\tau_{\rm s}$ above the MNI value. Whether $n_{\rm s} = \Theta(d)$ would in fact suffice here we do not know; (R2) is stated because it is what the proof uses. (R3) $n_{\rm s}/n_{\rm w} \to \infty$ enters only through $\eta \to 0$ in (K.14) — the quantitative form of "near-clone".

No rate of divergence is needed in (R3): the self-noise is controlled by η^2 times a quantity that is *already* at the target rate, not by η^2 times a constant.

Lower bound: the pipeline cannot beat the rate. This half is formula-free. The student output $\widehat{w}_{\rm s}$ is a measurable function of $(\beta_{\rm w}, \widetilde{X})$, where $\beta_{\rm w}$ depends on the data only through the $n_{\rm w}$ labeled teacher samples $D_{\rm w}$, and the fresh unlabeled $\widetilde{X}$ is independent of $(w^*, D_{\rm w})$ (it informs about the known Σ and carries zero information about w^*). So $w^* \to D_{\rm w} \to \widehat{w}_{\rm s}$ is a Markov chain for each fixed $\widetilde{X}$, and

$$\mathbb{E}[\mathcal{E}_{\rm s}] = \mathbb{E}_{\widetilde{X}}\,\mathbb{E}\big[\|\Sigma^{1/2}(\widehat{w}_{\rm s} - w^*)\|^2 \mid \widetilde{X}\big] \ge \mathbb{E}_{\widetilde{X}}\big[\mathrm{BayesRisk}(D_{\rm w})\big] = \mathrm{BayesRisk}(D_{\rm w}),$$

since for each fixed $\widetilde{X}$, $\widehat{w}_{\rm s} = g_{\widetilde{X}}(D_{\rm w})$ is *some* estimator of w^* from $D_{\rm w}$.

Making the bound exact: with Gaussian prior $w^* \sim \mathcal{N}(0, W)$, $W = \mathrm{diag}(\omega_i^2)$, and design X (n_{w} rows $\mathcal{N}(0, \Sigma)$), the posterior of w^* given $D_{\mathrm{w}} = (X, y)$ is Gaussian with covariance $(W^{-1} + X^\top X/\sigma^2)^{-1}$, so

$$\mathrm{BayesRisk}_{\mathrm{exact}}(n_{\mathrm{w}}) = \mathbb{E}_X \,\mathrm{tr}\left[\Sigma(W^{-1} + X^\top X/\sigma^2)^{-1}\right].$$

Now $M \mapsto (W^{-1} + M)^{-1}$ is operator convex on the PSD cone (matrix inversion is operator convex; Bhatia, *Matrix Analysis*, Thm. V.2.5), and $\mathrm{tr}[\Sigma\,\cdot\,]$ is a positive linear functional ($\Sigma \succeq 0$), so $M \mapsto \mathrm{tr}[\Sigma(W^{-1} + M)^{-1}]$ is convex on PSD M.

Jensen over the random Gram, with $\mathbb{E}_X[X^\top X] = n_{\mathrm{w}}\Sigma$,

$$\mathrm{BayesRisk}_{\mathrm{exact}}(n_{\mathrm{w}}) \geq \mathrm{tr}\left[\Sigma(W^{-1} + n_{\mathrm{w}}\Sigma/\sigma^2)^{-1}\right] = \sum_i \frac{\sigma^2\lambda_i\omega_i^2}{\sigma^2 + n_{\mathrm{w}}\lambda_i\omega_i^2} =: \mathrm{SeqRisk}(n_{\mathrm{w}}). \tag{K.15}$$

So the sequence-model formula is a rigorous *lower* bound on the exact Bayes risk, not an approximation of it. Its rate is elementary: with the transition at $i_0 = (n_{\mathrm{w}}G/\sigma^2)^{1/(a+b)}$, the head ($i < i_0$) contributes $\sim i_0\sigma^2/n_{\mathrm{w}}$ and the tail ($i > i_0$) contributes $\sim G i_0^{1-(a+b)}$, both $\Theta(n_{\mathrm{w}}^{-(a+b-1)/(a+b)})$.

Chaining,

$$\mathbb{E}[\mathcal{E}_{\mathrm{s}}] \geq \mathrm{BayesRisk}_{\mathrm{exact}}(n_{\mathrm{w}}) \geq \mathrm{SeqRisk}(n_{\mathrm{w}}) = \Theta\big(n_{\mathrm{w}}^{-(a+b-1)/(a+b)}\big).$$

The student may denoise the teacher (so $\bar{\mathcal{E}}_s < \bar{\mathcal{E}}_w$ is possible), but it can never manufacture information about w^* absent from the n_{w} labels.

Two objects appear when closing the sandwich: the upper bound is on the limit *formula* $\bar{\mathcal{E}}_s^*$, while the Bayes bound above is on $\mathbb{E}[\mathcal{E}_{\mathrm{s}}]$ of the *actual* pipeline — and probability convergence transfers expectations the wrong way for a lower bound. They need no separate machinery, because one and the same $\mathrm{SeqRisk}(n_{\mathrm{w}})$ bounds both. For $\mathbb{E}[\mathcal{E}_{\mathrm{s}}]$ that is (K.15); for the formula it is the next three lines.

The formula-level lower half. Write $h_i := F_i^{\mathrm{s}}F_i^{\mathrm{w}}$ for the filter the pipeline applies to the truth, so that $h_i = 0$ and $v_i^{\mathrm{w}} = 0$ for $i > m$. A teacher's fluctuation profile and its shrinkage are the same object up to a factor λ_i,

$$v_i^{\mathrm{w}} = \frac{\lambda_i\Gamma_{\mathrm{w}}}{(\lambda_i + \tau_{\mathrm{w}})^2} = \frac{(F_i^{\mathrm{w}})^2}{\lambda_i}\,\Gamma_{\mathrm{w}},$$

so the passed-noise term of (D.2) loses its τ_{w} entirely when written in h:

$$\lambda_i(F_i^{\mathrm{s}})^2 v_i^{\mathrm{w}} \;=\; h_i^2\,\Gamma_{\mathrm{w}} \;\geq\; h_i^2\,\frac{\sigma^2}{n_{\mathrm{w}}} \qquad \text{(every } i\text{, every } \tau_{\mathrm{w}}\text{)}, \tag{K.16}$$

the floor being $\Gamma_{\mathrm{w}} \geq s_{\mathrm{w}}^2/(n_{\mathrm{w}}(1 - \Xi_{\mathrm{w}})) \geq \sigma^2/n_{\mathrm{w}}$, by (C.6) with $\Xi_{\mathrm{w}} \in (0, 1)$ and $s_{\mathrm{w}}^2 = \sigma^2 + A_m \geq \sigma^2$.

Dropping the nonnegative student self-noise and minimizing coordinatewise over $h_i \in \mathbb{R}$ — *unconstrained* —

$$\bar{\mathcal{E}}_s \;\geq\; \sum_i \left[\lambda_i\omega_i^2(1 - h_i)^2 + \frac{\sigma^2}{n_{\mathrm{w}}}\,h_i^2\right] \;\geq\; \sum_i \frac{\sigma^2\lambda_i\omega_i^2}{\sigma^2 + n_{\mathrm{w}}\lambda_i\omega_i^2} \;=\; \mathrm{SeqRisk}(n_{\mathrm{w}}), \tag{K.17}$$

since $\min_{h\in\mathbb{R}}\{A(1 - h)^2 + Bh^2\} = AB/(A + B)$ at $h^* = A/(A + B)$, here with $A = \lambda_i\omega_i^2$ and $B = \sigma^2/n_{\mathrm{w}}$. This holds for *every* $(\tau_{\mathrm{w}}, \tau_{\mathrm{s}})$, every m and every summable spectrum.

Two features are worth naming. The minimizer $h_i^* = n_{\mathrm{w}}\lambda_i\omega_i^2/(\sigma^2 + n_{\mathrm{w}}\lambda_i\omega_i^2)$ already lies in $[0, 1]$, so the feasibility constraint $h_i \leq F_i^{\mathrm{w}}$ that a pipeline does obey is never needed; and the $\tau_{\mathrm{w}} \to \infty$ escape one might fear — a student dodging the teacher bias — is closed by (K.16), which cancels τ_{w} outright, not by any constraint. Discarding the factor $(F_i^{\mathrm{w}})^{-2}$ in that display is exactly what would force a constrained minimization; keeping it is what makes the bound τ_{w}-free.

The rate is the head/tail computation already made after (K.15), now run as a lower bound with an explicit constant: for $i \leq i_0 = (n_{\mathrm{w}}G/\sigma^2)^{1/(a+b)}$ one has $n_{\mathrm{w}}\lambda_i\omega_i^2 \geq \sigma^2$, so each such summand of (K.17) is $\geq \sigma^2/(2n_{\mathrm{w}})$, and those terms alone give, once $i_0 \geq 2$,

$$\bar{\mathcal{E}}_s \;\geq\; \mathrm{SeqRisk}(n_{\mathrm{w}}) \;\geq\; \tfrac{1}{4}\,G^{1/(a+b)}\big(\sigma^2/n_{\mathrm{w}}\big)^{(a+b-1)/(a+b)} \qquad (b \geq 0). \tag{K.18}$$

For $b < a+1$ this is the upper bound's exponent, so part (iii) is closed there by (K.17) alone — the supervised head/tail computation rerun, with the noise side counted only up to i_0, i.e. capped by the signal, exactly as in the sequence form (K.15).

Saturated band and the boundary: what (K.17) cannot do, and what closes them. The same computation is also its ceiling, and that ceiling is exact rather than numerical: $\mathrm{SeqRisk}(n_{\mathrm{w}}) \asymp n_{\mathrm{w}}^{-(a+b-1)/(a+b)}$ for every $b \geq 0$, which for $b > a+1$ decays *strictly faster* than the claimed $n_{\mathrm{w}}^{-2a/(2a+1)}$, while at $b = a+1$ the two exponents agree but SeqRisk is log-free. So (K.17) is provably insufficient on $b \geq a+1$: it is an information bound, and what it gives away there is exactly what a Bayes estimator gains over a ridge student. The reason is structural: a free per-coordinate gain may mute head coordinates where noise exceeds signal; a ridge student cannot, and exactly this must be retained.

Write $\underline{v}_i := \sigma^2\lambda_i/((\lambda_i+\tau_{\mathrm{w}})^2 n_{\mathrm{w}}) \leq v_i^{\mathrm{w}}$ (the same floor $\Gamma_{\mathrm{w}} \geq \sigma^2/n_{\mathrm{w}}$) and $i_\tau := \tau^{-1/a}$. Fix *any* admissible pair $(\tau_{\mathrm{w}}, \tau_{\mathrm{s}})$ — no optimality is assumed anywhere below — and set $\tau_{\mathrm{e}} := \max(\tau_{\mathrm{w}}, \tau_{\mathrm{s}})$. Since $F_i^{\mathrm{s}}, F_i^{\mathrm{w}} \in [0,1]$ and $\tau \mapsto \tau/(\lambda+\tau)$ is increasing,

$$1 - F_i^{\mathrm{s}} F_i^{\mathrm{w}} \;\geq\; \max(1-F_i^{\mathrm{s}},\, 1-F_i^{\mathrm{w}}) \;=\; \tau_{\mathrm{e}}/(\lambda_i+\tau_{\mathrm{e}}) \qquad (i \leq m),$$

and $1 - F_i^{\mathrm{s}}F_i^{\mathrm{w}} = 1 \geq \tau_{\mathrm{e}}/(\lambda_i+\tau_{\mathrm{e}})$ for $i > m$; while on the joint head $i \leq \min(i_{\tau_{\mathrm{e}}}, m)$ one has $\lambda_i \geq \tau_{\mathrm{e}} \geq \tau_{\mathrm{s}}$, hence $(F_i^{\mathrm{s}})^2 \geq \frac{1}{4}$, and $\lambda_i \geq \tau_{\mathrm{e}} \geq \tau_{\mathrm{w}}$, hence $\lambda_i \underline{v}_i \geq \sigma^2/(4n_{\mathrm{w}})$.

Keeping the teacher-noise term as $(F_i^{\mathrm{s}})^2 v_i^{\mathrm{w}} \geq (F_i^{\mathrm{s}})^2 \underline{v}_i$ and discarding the (nonnegative) student self-noise,

$$\bar{\mathcal{E}}_s \;\geq\; B_0^{(d)}(\tau_{\mathrm{e}}) \;+\; \frac{\sigma^2}{16\, n_{\mathrm{w}}} \min\big(\lfloor i_{\tau_{\mathrm{e}}} \rfloor,\, m\big), \tag{K.19}$$

with $B_0^{(d)}$ the truncated bias sum of Lemma K.8: the supervised one-ridge objective in the *single* variable τ_{e} — the precise sense in which the lower half reruns the supervised computation. Because (K.19) holds at every $(\tau_{\mathrm{w}}, \tau_{\mathrm{s}})$, minimizing its right side over $\tau_{\mathrm{e}} \in (0,\infty)$ bounds $\bar{\mathcal{E}}_s^*$ from below. Three cases exhaust that range; two are degenerate, and it is exactly there that the resource hypotheses (K.11) are needed.

Case A: $\tau_{\mathrm{e}} \geq \lambda_1 = 1$. Then $\tau_{\mathrm{e}}/(\lambda_i+\tau_{\mathrm{e}}) \geq \frac{1}{2}$ for every i, so $B_0^{(d)}(\tau_{\mathrm{e}}) \geq \frac{1}{4}\sum_{i\leq d}\lambda_i\omega_i^2 \geq r^2/8$ for d large: a constant, hence far above $\mathcal{R}(n_{\mathrm{w}})$.

Case B: $\tau_{\mathrm{e}} < 1$ and $\lfloor i_{\tau_{\mathrm{e}}} \rfloor \geq m$. Then the noise term of (K.19) is at least $\sigma^2 m/(16 n_{\mathrm{w}}) \geq \sigma^2 c_m^-/16$ by (R1): again a constant. (This is the branch in which both ridges are so small that the teacher's effective dimension exceeds its own capacity; it is ruled out by a lower bound on the risk, not by fiat.)

Case C: $\tau_{\mathrm{e}} < 1$ and $\lfloor i_{\tau_{\mathrm{e}}} \rfloor < m$. Then $i_{\tau_{\mathrm{e}}} < m$, so $\tau_{\mathrm{e}} > m^{-a} \geq d^{-a}$ and Lemma K.8 applies at $p = d$ with $c_1 = 1$: $B_0^{(d)}(\tau_{\mathrm{e}}) \geq c_- G \Phi_{ab}(\tau_{\mathrm{e}})$. Also $i_{\tau_{\mathrm{e}}} \geq 1$, so $\min(\lfloor i_{\tau_{\mathrm{e}}} \rfloor, m) = \lfloor i_{\tau_{\mathrm{e}}} \rfloor \geq \frac{1}{2} i_{\tau_{\mathrm{e}}}$. Hence

$$\bar{\mathcal{E}}_s \;\geq\; c_-\, G\, \Phi_{ab}(\tau_{\mathrm{e}}) \;+\; \frac{\sigma^2}{32\, n_{\mathrm{w}}}\, \tau_{\mathrm{e}}^{-1/a},$$

and Lemma K.10 with $A := c_- G$, $B := \sigma^2/(32 n_{\mathrm{w}})$ — so that $A \geq B$ and $\ell = 1 + \log(32 c_- G n_{\mathrm{w}}/\sigma^2) \asymp \log n_{\mathrm{w}}$ for n_{w} large — yields

$$\bar{\mathcal{E}}_s \;\geq\; \kappa \cdot \begin{cases} G^{1/(a+b)}(\sigma^2/n_{\mathrm{w}})^{(a+b-1)/(a+b)}, & b < a+1, \\ G^{1/(2a+1)}(\sigma^2/n_{\mathrm{w}})^{2a/(2a+1)}(\log n_{\mathrm{w}})^{1/(2a+1)}, & b = a+1, \\ G^{1/(2a+1)}(\sigma^2/n_{\mathrm{w}})^{2a/(2a+1)}, & b > a+1, \end{cases}$$

with $\kappa = \kappa(a,b) > 0$; i.e. $\bar{\mathcal{E}}_s \geq \kappa' \mathcal{R}(n_{\mathrm{w}})$.

Cases A and B are larger still, so $\bar{\mathcal{E}}_s^*(n_{\mathrm{w}}) \geq \kappa' \mathcal{R}(n_{\mathrm{w}})$: the extra boundary factor of Theorem D.14 (iii) survives on the lower side, and the saturated band $b > a+1$ is closed — both through the composed-ridge structure, neither through the per-coordinate bound (K.17) alone.

Numerically (formula-level minimization, slopes per doubling of n_{w}): at $(a,b) = (2,1)$ both the full formula and the composed-ridge bound measure -0.66 against $-\frac{2}{3}$; at $(1.5, 0.5)$, -0.49 against $-\frac{1}{2}$;

at $(2,4)$ they sit at -0.80 (target $-\frac{4}{5}$), while $\mathrm{SeqRisk}$ decays at the strictly faster Bayes exponent $-\frac{5}{6}$ — the leak of (K.17) on $b \geq a+1$ is exact, not a numerical drift; at the boundary $(1.5, 2.5)$ the composed-ridge minimum grows with $L^{1/4}$ (ratio flat to 3%); and pointwise on 31×31 $(\tau_{\mathrm{w}}, \tau_{\mathrm{s}})$ grids at eight (a,b) corners, both $\bar{\mathcal{E}}_s \geq \mathrm{SeqRisk}(n_{\mathrm{w}})$ and $\bar{\mathcal{E}}_s \geq$ the composed bound hold with no violation.

The upper half of Proposition K.14 was probed the same way: at $(\bar{\kappa_{\mathrm{s}}}, \sigma^2, r^2) = (2, 0.25, 1)$ with $m = 2n_{\mathrm{w}}$, $d = n_{\mathrm{w}}^2$, $n_{\mathrm{s}} = d/2$ and n_{w} up to 800, the evaluated $\bar{\mathcal{E}}_s$ at the ridge of Step 0 tracks $\mathcal{R}(n_{\mathrm{w}})$ with a flat ratio, the measured slope reaching -0.667 at $(2,1)$ (target $-\frac{2}{3}$) and -0.771 at $(2,3)$ against the predicted $-\frac{4}{5} + \frac{1}{5}/\log n_{\mathrm{w}} = -0.770$; the student self-noise T_3 sits at 10^{-15} of the total, as (K.14) predicts. Corroboration, never a proof step.

Together, for every $b \geq 0$,

$$\bar{\mathcal{E}}_s^*(n_{\mathrm{w}}) \asymp \mathcal{R}(n_{\mathrm{w}}) = \begin{cases} n_{\mathrm{w}}^{-(a+b-1)/(a+b)}, & b < a+1, \\ n_{\mathrm{w}}^{-2a/(2a+1)} (\log n_{\mathrm{w}})^{1/(2a+1)}, & b = a+1, \\ n_{\mathrm{w}}^{-2a/(2a+1)}, & b > a+1, \end{cases}$$

which is Theorem D.14 (iii) in full. The upper half is Proposition K.14 — an explicit teacher ridge and a minimum-norm student, with the "near-clone" step discharged by (K.14). The lower half is reached twice over: by the τ_{w}-free sequence bound (K.17), which closes $b < a+1$ using no resource hypothesis at all, and by the composed-ridge bound (K.19), which closes all three bands (and is what supplies the saturated exponent and the boundary logarithm, where (K.17) provably leaks).

Neither uses any convergence theory. For $b < a+1$ the exponent is moreover unimprovable by *any* pipeline and not merely by the formulas: (K.15) bounds $\mathbb{E}[\mathcal{E}_{\mathrm{s}}]$ of the actual estimator, and the exact random-design Bayes risk reduces to the same sequence formula by operator convexity — no external minimax theorem is load-bearing.

Remark K.16 (What each part is a statement about). Every part is an exact asymptotic of the deterministic limit formula $\bar{\mathcal{E}}_s$ of (D.2), and the shape-limits taken below ($n_{\mathrm{s}}/d \to 0$ in (i), optimal ridge in (iii)) are limits of those formulas. Those formulas are the channel calculus of Definition C.2, defined for every summable spectrum, so the power law (C.2) lies inside their domain throughout. What Assumption C.1 governs — and hence the range in which Theorem D.1 identifies $\bar{\mathcal{E}}_s$ with the limit of the random risk $\mathcal{E}_{\mathrm{s}}$ — is bounded condition number, and we make that identification nowhere else.

Two features of part (iii) are worth recording at this level. First, it tunes the ridge, while our convergence engines are written at the ridgeless fixed point; the ridge variant (τ bounded below, all resolvents uniformly regular) is strictly easier but not transcribed, so (iii) is consumed exactly as it is stated, as an asymptotic of the formulas. Second, its optimality (Corollary D.15) is formula-independent and unconditional — the Bayes bound uses only the Markov structure and the Gaussian prior — so the matching of exponents there rests on no convergence theorem at all.

L The Gaussian sketch (isotropic features)

Let $S \in \mathbb{R}^{m \times d}$ have i.i.d. $\mathcal{N}(0, 1/d)$ entries, $\Sigma = I_d$, isotropic prior. Write $C = SS^\top$, whose spectrum converges to the Marchenko–Pastur law μ_ψ of ratio ψ (mean 1, support $[(1-\sqrt{\psi})^2, (1+\sqrt{\psi})^2]$).

Reduction. For Gaussian x, conditioning on S: the best sketched predictor is $\theta^* = C^{-1} S w^*$ with $\mathbb{E}[\theta^*(\theta^*)^\top \mid S] = \frac{r^2}{d} C^{-1}$, the approximation error is $A_S = \|w^*\|^2 (1 - \frac{m}{d}) + o_{\mathbb{P}}(1)$, which is $r^2(1-\psi) + o_{\mathbb{P}}(1)$ under the moment condition of Remark G.4 (the realized $\|w^*\|^2$ need not concentrate for the variance-only prior class), and the residual is Gaussian and independent of Sx.

The teacher is therefore the channel of Definition C.2 on the (random, quenched) spectrum of C with target profile $u_i^2 = \frac{r^2}{d} c_i^{-1}$ in the eigenbasis of C and noise $s^2 = \sigma^2 + r^2(1-\psi)$; all channel sums become integrals against μ_ψ, e.g. the fixed point reads $\int \frac{c}{c+\tau_{\mathrm{w}}} \, d\mu_\psi(c) = 1/\kappa_{\mathrm{w}}$, explicitly solvable through the Stieltjes transform $\int \frac{d\mu_\psi(c)}{c+\tau} = \frac{\sqrt{(1-\psi+\tau)^2 + 4\psi\tau} - (1-\psi+\tau)}{2\psi\tau}$, and the inverse moments $\int c^{-1} d\mu_\psi = (1-\psi)^{-1}$, $\int c^{-2} d\mu_\psi = (1-\psi)^{-3}$ enter the target-profile sums.

The student composition is immediate because with $\Sigma = I$ only $\|\beta_{\mathrm{w}}\|^2$ and $\langle \beta_{\mathrm{w}}, w^* \rangle$ survive (Corollary D.3 applies): $\mathbb{E}\,\mathcal{E}_{\mathrm{s}} = \frac{n_{\mathrm{s}}}{d}\mathbb{E}\|\beta_{\mathrm{w}} - w^*\|^2 + (1 - \frac{n_{\mathrm{s}}}{d})r^2$. Simulation agrees to $< 1\%$ at $d = 1500$.

Random vs. aligned capacity. At equal $(\psi, \kappa_{\mathrm{w}}, \sigma^2)$ the Gaussian-sketch teacher is worse than the truncation teacher even though $\Sigma = I$ has no preferred directions: the truncation teacher's in-sketch covariance is I_m (condition number 1) while the Gaussian sketch's is the spread-out $C \sim \mu_\psi$, which inflates the interpolation fixed point's variance response. Numerically at $(\psi, \kappa_{\mathrm{w}}, \sigma^2, r^2) = (0.3, 3, 0.25, 1)$: $\bar{\mathcal{E}}_w = 1.448$ (Gaussian sketch) vs. 1.375 (truncation). Random feature budgets pay an ill-conditioning tax before any signal is even lost.

M Finite-sample information bounds

M.1 The observation model in Proposition 2.2

Step 1: preserve the observations while reversing a hidden sign. Let $\mathcal{O}$ be the sigma-field generated by the teacher output, independent fresh input batches, and independent random seeds. For each $i > m$, the measure- preserving transformation

$$(w_i^*, X_{:,i}) \longmapsto (-w_i^*, -X_{:,i})$$

leaves $\mathcal{O}$ invariant.

Step 2: eliminate the cross term. Consequently, for every bounded measurable h,

$$\mathbb{E}[w_i^* h(\mathcal{O})] = 0.$$

Truncation extends the identity to square-integrable estimators. Thus

$$\mathbb{E}[w_i^* \mid \mathcal{O}] = 0, \qquad \mathbb{E}(\widehat{w}_i - w_i^*)^2 = \mathbb{E}\widehat{w}_i^2 + \omega_i^2.$$

Multiplying by λ_i and summing over $i > m$ proves (2.3).

Step 3: identify which downstream observations are covered. Adaptive queries to the fixed linear teacher do not change the conclusion: their transcript is measurable with respect to its parameter and the query-generation randomness. Any finite chain is also such a measurable transformation. The argument fails if the student is additionally given the hidden coordinates of the original labeled design, signal-dependent side information, or a prior that breaks the required sign symmetry.

M.2 Gaussian Bayes lower bound

Step 1: give the oracle all labeled coordinates. The oracle observes X and y from n_{w} independent noisy labeled examples, including coordinates hidden from the weak teacher. Under the independent Gaussian prior its conditional mean minimizes squared prediction loss, and its conditional Bayes risk is

$$\mathrm{tr}\Big\{\Sigma\big(\Omega^{-1} + \sigma^{-2}X^\top X\big)^{-1}\Big\}.$$

Every pipeline considered here is a randomized function of these labeled data and independent unlabeled data. Providing the oracle more observations than the pipeline can only lower its Bayes risk.

Step 2: average the posterior covariance. Taking expectations and using operator Jensen gives (6.1).

Step 3: count the coordinates that determine the rate. For $\lambda_i\omega_i^2 = Gi^{-s}$, $s = a + b > 1$, let

$$J = \lfloor c_0 n_{\mathrm{w}}^{1/s} \rfloor, \qquad 0 < c_0 < c.$$

For $i \le J$ the denominator ratio $(n_{\mathrm{w}}/\sigma^2)Gi^{-s}$ is bounded below by a positive constant, so each summand is at least c_1/n_{w}. Summing J terms gives

$$c_2 n_{\mathrm{w}}^{-(s-1)/s}.$$

The normalization $G = r^2/\sum_{i\le d} i^{-s}$ is bounded above and below by positive constants. This proves the rate for fixed $r^2, \sigma^2 > 0$, with constants depending on the displayed parameters and c. The Gaussian prior is substantive: its covariance alone does not determine the Bayes difficulty for other priors.

N Numerical checks and reproducible artifacts

This appendix collects the finite-dimensional checks behind Figures 1 and 2, their comparison protocols, and the computer-assisted inequalities used in the paper. The package includes per-trial records, seeds, an environment record, and scripts using fresh Gaussian inputs. The synthetic covariance and signal distributions were specified before each recorded run.

N.1 Finite-dimensional checks of the spiked example

The opening example predicts a gain that disappears at exact cloning. Figure 1 tests that prediction with finite-dimensional estimators. In 200 paired spiked-model trials, the tradeoff is visible; at the smallest budget, the discrepancy from the formula exceeds two Monte Carlo standard errors. The figure retains that discrepancy: the shape and its finite-size calibration are separate questions.

The paired spiked experiment in Figure 1 fixes the signal magnitudes and resamples independent Rademacher signs. This makes its channel second-moment profile also the realized squared profile, avoiding a mismatch between a fixed signal and a prior-averaged prediction. Paired gains reduce common teacher variation. Error bars are two Monte Carlo standard errors across independent full trials, and are not simultaneous confidence bands. The clone endpoint is inserted from the exact identity, not estimated by dividing the singular channel formula.

N.2 Paired chain comparison

The chain comparison tests whether a richer filter can earn its extra fitting cost. For the power-law instance in (5.6), Table 3 reports 40 independent paired trials. The paired improvement is 0.001984 ± 0.000447 (two standard errors), against the formula prediction 0.001890 for those fixed estimators. The empirical single-student baseline uses a numerically selected ridge; the separate formula certificate supplies the comparison against all single ridge choices.

Within each chain trial, both methods share the teacher and a pool of 2000 unlabeled inputs. The single student uses the entire pool; the chain uses disjoint batches of 1546 and 454, with ridges 0.0219 and 0.0173. The single-student ridge is approximately 0.048627. All were fixed from risk formulas before drawing simulation data. One risk difference per trial estimates the mean gain and its standard error, retaining the dependence induced by the shared teacher and inputs. The spike experiment likewise shares a teacher and nested input pools across budgets. Seeds, protocols, and per-trial results accompany the paper.

Power-law chain comparison	Single student	Two stages	Paired improvement
Monte Carlo mean, 40 trials	0.017708	0.015724	0.001984
Two standard errors of improvement			0.000447
Predicted improvement for fixed parameters			0.001890

Table 3: Fresh paired comparison at $(d, m, n_{\mathrm{w}}, N) = (4000, 1200, 400, 2000)$, $(a, b) = (1.2, 2)$, $r^2 = 1$, $\sigma^2 = 0.25$. Each trial resamples the signal, teacher data and noise, and a fresh student input pool. The chain splits that pool; the single student uses the full pool. The single ridge is selected numerically. A separate directed-rounding certificate compares a fixed chain witness against all single-ridge choices at the deterministic-formula level.

N.3 Symbolic checks and interval certificates

The script `audit/check_identities.py` performs twelve exact checks with SymPy 1.14.0: the two-atom fixed-point quadratics, the subtraction giving (I.6), homogeneity of the limiting gain, the extra-stage derivative, the coordinate fixed-point gap, the noiseless collapse, noise affinity and transfer, and three exponent identities. Each check records its two expressions and admissible domain. Rational simplification does not prove the existence of the limiting quantities or the asymptotic estimates in which those identities are used. The ledger results are checked separately by `audit/check_ledger.py`.

The scripts `chainA_iv_cert.py` and `l05e_iv_cert.py` in `certificates/` use directed-rounding interval arithmetic (`mpmath.iv`, version 1.3.0) at 60 decimal digits; their rerun logs are in `certificates/logs/`. For the instance in (5.6), the outward-safe bounds are

$$\inf_{t\in[0,\infty]} R_1^{\mathrm{pop}}(t) > 0.01721744696, \qquad R_2^{\mathrm{fin}} < 0.01630177680,$$

with a separately enclosed gap of at least 0.000915670175564611. The script encloses the teacher and both student fixed points, evaluates the full two-stage recursion at the quoted feasible witness, and covers every single-stage parameter using interval cells, a local second-derivative bound, and a monotone unbounded tail. The nonnegative fitting cost makes the population lower bound valid for every finite-budget single stage. For Theorem D.12, the second script certifies a global one-stage minimum and a feasible two-stage value separated by at least $2.20937164288882 \times 10^{-10}$, together with the required positive transversal criterion. These are rigorous arithmetic enclosures for two specified comparisons in the deterministic calculus; they are neither grid searches nor finite-sample probability bounds.