AN ACTIVE-BOTTLENECK MECHANISM FOR WEAK-TO-STRONG GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weak-to-strong generalization (W2SG) occurs when a student trained on a teacher's predictions outperforms that teacher. We study when this happens under fully converged, ridgeless two-stage learning, with no early stopping, no explicit regularization, and no assumption that the student is more expressive than the teacher. In two-stage linear regression, a teacher is fit from n labeled examples and a student is trained solely on the teacher's predictions on m fresh, unlabeled inputs. Although both stages share the same hypothesis class and the same training rule, we show that the student outperforms the teacher exactly when m lies in an explicit intermediate range: too few pseudo-labels leave the student without enough signal, too many let it inherit the teacher's noise. Under power-law covariance, we derive this range in closed form as a function of the spectral decay and noise level, including regimes where the improving region splits into two disjoint intervals of m. We then study a random-feature model in which the student has strictly more features than the teacher, and identify two regimes, again given by explicit thresholds: one where improvement occurs only for m in a bounded interval, and one where it occurs only once the student width N_S exceeds an explicit threshold. Both regimes are governed by a single *active-bottleneck* principle: whichever of m or N_S is scarcer controls how much teacher error is filtered out, while increasing the other resource only reduces estimation noise. Together, these results show that finite data and finite width can themselves regularize a two-stage learner, with no explicit mechanism doing so.

1 INTRODUCTION

As AI systems approach and eventually exceed human ability on complex tasks, humans will need to supervise models more capable than themselves. Burns et al. (2023) propose a way to study this problem today: fine-tune a strong pretrained model using only the predictions of a weaker, already fine-tuned one, and ask whether the strong model can outperform its supervisor. They call this *weak-to-strong generalization* (W2SG), and show that it occurs empirically — a GPT-4-level model supervised only by a GPT-2-level model can recover a substantial fraction of the performance gap between them. The finding is counterintuitive: the strong model is trained explicitly to fit the weak model's labels, yet it does not merely imitate them. Beyond predictive accuracy, recent empirical work asks whether W2SG can transfer alignment and trustworthiness properties such as safety, robustness, fairness, reasoning, and privacy, with mixed success across attributes (Ye et al., 2024; Pawelczyk et al., 2024; Yang et al., 2024; Jeon et al., 2026). Related work also identifies *weak-to-strong deception*, where stronger models can hide misaligned behavior in regions that weaker supervisors cannot reliably evaluate (Yang et al., 2025). Existing theoretical explanations for W2SG typically rely on some form of *asymmetry* between the two stages. Either the student is assumed to have strictly greater representational capacity than the teacher (Xue et al., 2025; Dong et al., 2025); or the student is trained with heavier or differently structured regularization than the teacher (Moniri and Hassani, 2025; Wu et al., 2026); or the student is stopped early, before it has had the chance to fully absorb the teacher's errors (Medvedev et al., 2025; Geng et al., 2026); or the analysis assumes that the student can correct pseudo-labels and expand their coverage (Lang et al., 2024; Shin et al., 2025); or teacher and student have access to different feature spaces, such as weak and strong features (Wu and Sahai, 2025); or the student is nonlinear and has greater capacity than a linear teacher (Oh et al., 2026).

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.

Each of these mechanisms gives the student some structural advantage that the teacher does not share. This raises a natural question: is such an asymmetry *necessary*? Can a student with exactly the teacher's capacity, trained by exactly the teacher's rule, run to exact convergence with no ridge penalty at all, still outperform a teacher it was trained only to imitate?

We show that the answer is yes, and that a single, simple resource is enough to cause it: the number of pseudo-labeled examples, m, that the student sees in its own training stage. In our first setting, both the teacher and the student solve ridgeless linear regression over the *same* hypothesis class, and the student is trained to convergence on m fresh inputs labeled by the teacher. Because m is finite, the student's minimum-norm solution is constrained to an m-dimensional subspace of parameter space; it does not have room to reproduce the teacher exactly. This subspace can retain more of the teacher's useful signal than its noise, or the reverse, depending on m. Under power-law covariance, we make this trade-off completely explicit: we solve for the critical values of m — as a fixed threshold, or as a limiting ratio m/p when m scales with the ambient dimension p — that separate the regime where the student is worse than the teacher from the regime where it is better, and we show that this improving region can itself be a single interval, all of $(0, p)$, or two disjoint intervals of m, depending on the noise level and the covariance decay rate.

This single-resource result suggests something more general is at work: a shared resource between the two training stages must be constrained to an intermediate range — tight enough to keep out the teacher's error, loose enough to let the target signal through — for W2SG to occur. We call such a resource, whichever it turns out to be, an *active bottleneck*, and in our second setting we test this idea against a genuine capacity asymmetry. We give the student strictly more random features than the teacher, so that the student is a truly stronger model in the usual sense, and ask whether pseudo-sample size still matters once this asymmetry is present. It does, but its role now shares the stage with a second resource, student width. We identify two regimes, again with explicit thresholds: in one, improvement holds only when m falls inside a bounded interval, which we solve for exactly; in the other, improvement requires the student width N_S to exceed an explicit threshold, and now *more* pseudo-labels are beneficial rather than harmful. The two regimes are governed by a single principle: whichever resource, m or N_S, is scarcer at a given point in parameter space is the one that determines how much of the teacher's error is filtered out; the other resource, once it is not the bottleneck, only reduces estimation noise and can be increased freely. This is the sense in which our two settings tell one story: Case I is the special case of this principle in which only one resource, m, is available to bind, and student and teacher share the same capacity; Case II makes the principle explicit and shows it survives, in modified form, once a genuine capacity asymmetry is introduced.

Our emphasis on an intermediate, rather than merely large, pseudo-sample size is not only a theoretical curiosity: large-scale empirical studies of distillation scaling laws have observed the same non-monotonicity directly in trained language models, where increasing the number of teacher-labeled examples eventually collapses W2SG even between models with a large capacity gap (Busbridge et al., 2025). We discuss the relationship between this empirical finding and our theoretical account in Appendix B. In summary, our contributions are

- **W2SG at ridgeless convergence with no capacity asymmetry.** In two-stage linear regression with identical teacher and student hypothesis classes, we show that a student trained to convergence on the teacher's own predictions can strictly outperform that teacher, with no early stopping and no explicit regularization at either stage.
- **An explicit sample-size phase diagram.** Under power-law covariance, we solve for the critical pseudo-sample size(s) separating improvement from non-improvement, as fixed thresholds and as limiting ratios m/p, and characterize when the improving region is a single interval, extends to the full overparameterized range, or splits into two disjoint intervals of m (Theorem 3.4). Numerical validation in Appendix A further supports our theoretical results.
- **A random-feature model with a genuine student advantage, and an active-bottleneck principle.** We extend the analysis to random features with $N_S > N_T$, and give explicit critical thresholds in m and in N_S separating the sample-bottleneck and feature-bottleneck regimes (Theorem 4.3, Corollary 4.5). Both regimes instantiate the same active-bottleneck principle: the scarcer of the two Stage-II resources controls teacher-error filtering, while the other resource only reduces estimation noise. We also validate this result numerically in Appendix A, finding agreement with our theoretical predictions.

We discuss the relationship of these results to the broader W2SG literature — in particular to alternative mechanisms based on early stopping, asymmetric regularization, representation geometry, and overparameterization, and to empirical scaling-law evidence at LLM scale — in Appendix B.

2 PRELIMINARIES

Two-stage learning. We consider a two-stage prediction pipeline common to both settings, with the hypothesis classes, learning rules, and noise models specialized in Sections 3 and 4. Let $\beta^\star \in \mathbb{R}^p$ be fixed throughout a problem instance, or drawn independently of the remaining randomness, and let $x \sim P_x$ with

$$\psi^\star(x) := x^\top \beta^\star, \qquad y = \psi^\star(x) + \varepsilon, \tag{1}$$

where ε is independent of x. When $\beta^\star$ is random, all expectations below also average over it. *Stage I* produces a teacher $\widehat{h}_T = \mathscr{A}_T(\mathcal{D}_n) \in \mathcal{H}_T$ from n labeled samples $\mathcal{D}_n := \{(x_i, y_i)\}_{i=1}^n$. The class $\mathcal{H}_T$ may depend on auxiliary randomness independent of $\mathcal{D}_n$. *Stage II* uses no labeled data of its own. It draws m fresh inputs $\widetilde{x}_i \overset{i.i.d.}{\sim} P_x$, independently of Stage I, and assigns pseudo-labels $\widetilde{y}_i := \widehat{h}_T(\widetilde{x}_i)$. With $\widetilde{\mathcal{D}}_m := \{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^m$, the student is $\widehat{h}_S := \mathscr{A}_S(\widetilde{\mathcal{D}}_m) \in \mathcal{H}_S$, where $\mathcal{H}_S$ may likewise depend on independent auxiliary randomness. Thus the student observes only teacher predictions, not $\beta^\star$, ε, or the Stage-I data directly.

Population risk and the W2SG gain. For a predictor h, define

$$\mathcal{R}(h; \beta^\star) := \mathbb{E}_{x \sim P_x}\left[\left(h(x) - x^\top \beta^\star\right)^2\right], \tag{2}$$

where the test point is independent of all training and auxiliary randomness. The averaged teacher and student risks are $\overline{\mathcal{R}}_T := \mathbb{E}[\mathcal{R}(\widehat{h}_T; \beta^\star)]$ and $\overline{\mathcal{R}}_S := \mathbb{E}[\mathcal{R}(\widehat{h}_S; \beta^\star)]$, where expectations are over all randomness entering the corresponding predictor, including $\beta^\star$ when random. The *weak-to-strong generalization gain* is $\mathsf{W2SG} := \overline{\mathcal{R}}_T - \overline{\mathcal{R}}_S$, and W2SG occurs when $\mathsf{W2SG} > 0$, i.e., when the student achieves lower expected risk than its teacher.

3 CASE I: RIDGELESS REGRESSION

We first instantiate the two-stage template of Section 2 in the setting where the teacher and student share the same linear hypothesis class, the same feature space, and the same ridgeless training rule. Ridgeless linear regression is not a toy warm-up here: even in this classical setting, minimum-norm interpolation is known to exhibit genuinely surprising behavior — double descent, benign overfitting, and sharp phase transitions in the proportional asymptotic regime — and it has been the object of substantial recent theoretical attention in its own right (Hastie et al., 2022; Cheng and Montanari, 2024; Han and Xu, 2026; Wu and Xu, 2020; Richards et al., 2021; Loureiro et al., 2021). We use it here because it lets us isolate a single mechanism for W2SG in its purest form: with teacher and student identical in every respect except the data each one sees, any improvement of the student over the teacher must come entirely from the *pseudo-sample size* m. Prior analyses of W2SG in ridgeless regression have focused on representation mismatch, scaling behavior, and optimization dynamics between teacher and student (Dong et al., 2025; Wu and Sahai, 2025; Wu et al., 2026). We instead remove every such source of asymmetry: teacher and student share one feature space and one hypothesis class, and both are trained to exact convergence with zero ridge. We show that W2SG nonetheless occurs, governed entirely by m.

3.1 MODEL AND TRAINING PROCEDURE

We specialize the data-generating process of (1) to Gaussian covariates, $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in \mathbb{R}^{p \times p}$ positive definite, and to Gaussian label noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ in Stage I, independent of x. The target coefficient $\beta^\star$ is random, independent of all training and test randomness, with $\mathbb{E}[\beta^\star] = 0$, $\mathbb{E}[\beta^\star \beta^{\star\top}] = I_p$. We do not need to assume a specific distribution for $\beta^\star$ beyond these two moments: every quantity we study — the teacher and student risks, and the W2SG gain — is a linear or quadratic functional of $\beta^\star$ through $\mathbb{E}[\beta^\star \beta^{\star\top}]$ alone, so any isotropic distribution with these moments (Gaussian, uniform on a sphere of the matching radius, or otherwise) yields identical results.

Both stages share the linear hypothesis class $\mathcal{H} := \{h_\beta : x \mapsto x^\top \beta,\ \beta \in \mathbb{R}^p\}$. Given $X = [x_1|\cdots|x_n]^\top \in \mathbb{R}^{n\times p}$ and $y = (y_1, \ldots, y_n)^\top$, the teacher is the minimum-norm interpolator $\widehat{\beta}_T = X^\dagger y$ with $\widehat{h}_T(x) = x^\top \widehat{\beta}_T$, where $X^\dagger$ denotes the Moore–Penrose pseudoinverse of X. Independently of Stage I, draw $\widetilde{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ for $i = 1, \ldots, m$, form the noiseless pseudo-labels $\widetilde{y}_i = \widetilde{x}_i^\top \widehat{\beta}_T$, and write $\widetilde{X} = [\widetilde{x}_1|\cdots|\widetilde{x}_m]^\top$, $\widetilde{y} = (\widetilde{y}_1, \ldots, \widetilde{y}_m)^\top$. The student is again the minimum-norm interpolator of its own (pseudo-labeled) data,

$$\widehat{\beta}_S = \widetilde{X}^\dagger \widetilde{y} = \widetilde{X}^\dagger \widetilde{X} \widehat{\beta}_T, \qquad \widehat{h}_S(x) = x^\top \widehat{\beta}_S. \tag{3}$$

For Gaussian covariates, the population risk (2) of any $h_\beta \in \mathcal{H}$ takes the standard closed form $\mathcal{R}(h_\beta; \beta^\star) = \|\Sigma^{1/2}(\beta - \beta^\star)\|_2^2$, used throughout the ridgeless-regression literature cited above. The averaged risks therefore specialize to

$$\overline{\mathcal{R}}_T = \mathbb{E}_{\beta^\star, X, \varepsilon}\left[\mathcal{R}(\widehat{h}_T; \beta^\star)\right], \qquad \overline{\mathcal{R}}_S = \mathbb{E}_{\beta^\star, X, \varepsilon, \widetilde{X}}\left[\mathcal{R}(\widehat{h}_S; \beta^\star)\right]. \tag{4}$$

3.2 Stage-II Projection Geometry

Because $\widetilde{X}$ is drawn independently of everything determining $\widehat{\beta}_T$ — in particular, independently of X, ε, and $\beta^\star$ — we can study the statistical behavior of the Stage-II design and the statistical behavior of the teacher separately, and only combine them at the level of the final risk. This separation is what makes Lemma 3.2 below useful: it is a statement purely about $\widetilde{X}$, valid for *any* fixed $\widehat{\beta}_T$, and can therefore be applied after averaging over Stage I. The minimum-norm student solution satisfies

$$\widehat{\beta}_S = P_{\widetilde{X}} \widehat{\beta}_T, \qquad P_{\widetilde{X}} := \widetilde{X}^\dagger \widetilde{X}, \tag{5}$$

where $P_{\widetilde{X}}$ is the orthogonal projector onto the row space of $\widetilde{X}$. Under our Gaussian design, once $m \geq p$ this projector is the identity almost surely, so the student exactly reproduces the teacher and the two-stage procedure is statistically inert. The only regime in which Stage II can do anything other than copy the teacher is therefore the overparameterized one, $m < p$, and we restrict attention to it from here on.

Proposition 3.1 (Necessity of a finite-sample bottleneck)**.** *If $m \geq p$, then $\widehat{\beta}_S = \widehat{\beta}_T$ almost surely and* $\mathsf{W2SG} = 0$*. Hence W2SG in Case I requires $m < p$.*

The proof follows immediately by straightforward algebra. When $m < p$, $P_{\widetilde{X}}$ has rank exactly m: Stage II retains the component of $\widehat{\beta}_T$ in an m-dimensional random subspace — the row space of $\widetilde{X}$ — and discards its component in the $(p-m)$-dimensional orthogonal complement. The sample size m controls the *dimension* of this retained subspace; the realized covariates $\widetilde{X}$ control its *orientation*. This is the source of the trade-off at the heart of Case I: increasing m lets more of the teacher's useful signal reach the student, but by the same mechanism it also lets more of the teacher's error reach the student. Covariance geometry biases which directions this random subspace preferentially retains. The next lemma makes this precise *in expectation*: retention is not a hard cutoff at the top m eigenvalues of Σ, but a soft preference for higher-variance directions, with a total retention budget fixed by m.

Lemma 3.2 (Expected spectral retention)**.** *Let $1 \leq m < p$ and $\Sigma = V \operatorname{diag}(\lambda_1, \ldots, \lambda_p) V^\top$, with $V = [v_1, \ldots, v_p]$ orthogonal and $\lambda_1 \geq \cdots \geq \lambda_p > 0$. Define $\ell_i := \mathbb{E}_{\widetilde{X}} \|P_{\widetilde{X}} v_i\|_2^2$. Then*

$$\mathbb{E}_{\widetilde{X}}[P_{\widetilde{X}}] = V \operatorname{diag}(\ell_1, \ldots, \ell_p) V^\top, \qquad 1 > \ell_1 \geq \cdots \geq \ell_p > 0, \qquad \ell_1 + \ldots + \ell_p = m.$$

Proof is outlined in Appendix C (see Lemma C.6). Thus m sets a total expected retention budget across all p directions, while Σ determines how that budget is allocated: higher-variance directions are retained with no smaller expected weight than lower-variance ones, but *every* direction, including the top eigendirection, retains a strictly positive expected fraction below one. A finite-sample bottleneck therefore attenuates both signal and teacher error, never one exclusively; whether the net effect helps or hurts the student is exactly what the risk analysis below resolves.

3.3 Power-Law Covariance and Two Regimes for m

We now specialize Σ to a power-law spectrum. Non-isotropy is essential to the mechanism: under an isotropic covariance, the projection of Lemma 3.2 retains every direction with the same expected

weight m/p, attenuating signal and teacher error identically, so W2SG cannot occur without an explicit ridge. A power-law spectrum gives the graded, coordinate-dependent separation between signal and noise-dominated directions that a finite-sample projection needs to filter one from the other. We use it both for this reason and because it is the standard testbed for learning curves and scaling laws in high-dimensional regression (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Spigler et al., 2020; Cui et al., 2021; Bordelon et al., 2024; Defilippis et al., 2026; Wu et al., 2026), smoother than the spiked-covariance model used for a related purpose by Wu and Sahai (2025). We conjecture the same qualitative behavior holds for other sufficiently graded non-isotropic spectra, though not for every non-isotropic Σ.

Assumption 3.3 (Power-law spectrum and asymptotic scaling)**.** *The population covariance has eigenvalues $\lambda_i(\Sigma) = i^{-\alpha}$ for $i = 1, \dots, p$, with $1 < \alpha < 1 + \sqrt{2}$. As $p \to \infty$, the Stage-I sample size satisfies $p/n \to \phi_n \in (1, \infty)$, and the Stage-II sample size m follows one of two regimes:*

- Proportional regime*: $m \to \infty$ with $p/m \to \phi_m \in (1, \infty)$, so $m = \Theta(p)$;*
- Fixed-sample regime*: m is held at a fixed finite value as $p \to \infty$, so $m/p \to 0$.*

The two regimes capture genuinely different critical scales: the fixed-sample regime can reveal $O(1)$ thresholds in m, whereas the proportional regime captures thresholds of order $\Theta(p)$. Considering both is therefore necessary to describe the full phase diagram. The restriction $\alpha < 1 + \sqrt{2}$ is technical and specific to the phase-diagram analysis; the regime $\alpha > 1 + \sqrt{2}$, which is only partially resolved, is discussed in Appendix C.

3.4 Asymptotic W2SG Phase Diagram

Under Assumption 3.3, the teacher and student risks admit deterministic equivalents: asymptotically exact high-dimensional approximations, in the sense of Atanasov et al. (2026), that replace the random resolvents governing $\overline{\mathcal{R}}_T$ and $\overline{\mathcal{R}}_S$ by explicit functions of Σ and the two sample ratios (also, see Atanasov et al. (2025) for two-point DE). We write $\mathsf{W2SG}_p^{\mathrm{DE}}$ for the resulting deterministic equivalent of the finite-p gain $\mathsf{W2SG}_p$; the full derivation, including the standard Stage-I risk formula in terms of a self-consistent ridge parameter τ_T and its associated coefficient $\gamma_T \in (0, 1)$ (already known from the ridgeless-regression literature cited above), is given in Appendix C. Here we state only what is needed to read the theorem: γ_T is a scalar, determined by α and ϕ_n alone, that measures how much of the Stage-I estimator's error is variance (driven by σ^2) rather than bias.

The two regimes of Assumption 3.3 yield different phase diagrams, stated here and proved in Appendix C. In the proportional regime, $m = \Theta(p)$ and we study $\mathsf{W2SG}^{\mathrm{DE}} := \lim_{p\to\infty} \mathsf{W2SG}_p^{\mathrm{DE}}$; in the fixed-sample regime, $m = O(1)$ and we study $G_m := \lim_{p\to\infty} \mathsf{W2SG}_p^{\mathrm{DE}}$, which retains its dependence on m.

Theorem 3.4 (W2SG phase diagram, informal)**.** *Suppose Assumption 3.3 holds and $\sigma^2 > 0$.*

(a) Proportional regime ($m = \Theta(p)$).

(i) *If $1 < \alpha \le 2$, then $\mathsf{W2SG}^{\mathrm{DE}} > 0$ for **every** $\phi_m > 1$ — that is, for every pseudo-sample size $m < p$ growing proportionally with p, with no further condition on σ^2 or on ϕ_m.*

(ii) *If $2 < \alpha \le 1 + \sqrt{2}$, there is a unique critical ratio $\phi_{m,\mathrm{crit}}$ (solution of a fixed-point equation, see Appendix C), depending on α, ϕ_n, and σ^2 but requiring no restriction on σ^2, such that $\mathsf{W2SG}^{\mathrm{DE}} > 0$ if and only if $\phi_m < \phi_{m,\mathrm{crit}}$, equivalently $m > m_{\mathrm{crit}}^{\mathrm{FP}}(p) \sim p/\phi_{m,\mathrm{crit}} = \Theta(p)$.*

(b) Fixed-sample regime ($m = O(1)$). *Here $\mathsf{W2SG}_p^{\mathrm{DE}}$ converges, as $p \to \infty$ with m fixed, to an explicit limit G_m depending only on m, α, and $\sigma^2\gamma_T/(1-\gamma_T)$.*

(i) *If $1 < \alpha < 2$: for σ^2 below an explicit threshold $\sigma_c^2(\alpha, \phi_n)$, $G_m < 0$ at $m = 1$; and $G_m \to (2-\alpha)\sigma^2\gamma_T/(1-\gamma_T) > 0$ as $m \to \infty$, so at least one crossing m_* exists. For σ^2 below a second, smaller threshold $\sigma_0^2(\alpha, \phi_n) < \sigma_c^2(\alpha, \phi_n)$, this crossing is unique, and as*

$\sigma^2 \downarrow 0$ *it is located at*

$$m_{\text{crit}}^{\text{HP}} \sim \left[\frac{c_\alpha(1-\gamma_T)}{(2-\alpha)\sigma^2\gamma_T}\right]^{\frac{1}{\alpha-1}}, \qquad c_\alpha := \left[\frac{\pi}{\alpha}\csc\frac{\pi}{\alpha}\right]^\alpha, \tag{6}$$

and $\mathsf{W2SG}_p^{\text{DE}} \to G_m > 0$ *for every* $m > m_{\text{crit}}^{\text{HP}}$.

(ii) *If* $2 < \alpha \leq 5/2$*: there is an explicit threshold* $\sigma^2_{\text{safe}}(\alpha, \phi_n)$ *such that, for* $\sigma^2 \leq \sigma^2_{\text{safe}}(\alpha, \phi_n)$, $G_m \leq 0$ *for* every *fixed* $m \geq 1$*: in this low-noise regime, no fixed pseudo-sample size produces W2SG at all.*

Explicit formulas for $\sigma_{\text{safe}}, \sigma_0, \sigma_c$ *can be found in Appendix C.*

Two qualitative points are worth drawing out explicitly. First, part (a) shows that when the pseudo-sample size scales with the ambient dimension, a single, unconditional statement is possible for $\alpha \leq 2$: W2SG is guaranteed for *every* $m < p$, with no bottleneck fine-tuning needed at all; the bottleneck only becomes a genuine constraint, with a sharp critical threshold, once $\alpha > 2$. Second, part (b) shows that the fixed- and proportional-sample regimes are not merely two proof techniques for the same fact — they can disagree qualitatively. For $2 < \alpha \leq 5/2$ and sufficiently small noise, no *fixed* m ever produces W2SG, yet part (a)(ii) guarantees a $\Theta(p)$-scale pseudo-sample size that does. The bottleneck effect, in this regime, is a genuinely proportional phenomenon: it requires m to grow with p, not merely to be large in absolute terms. See Figures 2 and 3 in Appendix A for numerical validation.

Remark 3.5 (Beyond the small-noise threshold, and disconnected regions). The explicit formula (6) for $m_{\text{crit}}^{\text{HP}}$ is an asymptotic statement as $\sigma^2 \downarrow 0$, though in practice it tracks the true critical m closely even at moderate noise levels. For noise above the thresholds in Theorem 3.4, teacher error can be concentrated enough near specific high-variance directions — rather than spread smoothly across the spectrum — that the Stage-II projection improves on the teacher even at very small m. Combined with the proportional-scale threshold of part (a), the set of pseudo-sample sizes producing W2SG can then split into a window near $m = O(1)$ and a separate window near $m = \Theta(p)$. This is not merely an artifact of two disconnected proof techniques: the intermediate scaling $m \to \infty$ with $m/p \to 0$, analyzed in Appendix C (Section C.6), contains no critical m at all, so our fixed-sample and proportional results together already give a complete picture across every scaling of m with p.

4 Case II: The Random-Feature Model

Case I shows that a representational advantage is not necessary for W2SG: with teacher and student sharing one hypothesis class, the pseudo-sample size m alone was enough to act as the bottleneck. We now ask a different question. Suppose the student genuinely *is* a stronger model, with a richer representation than the teacher. Does the bottleneck mechanism survive this asymmetry, and if so, is m still the bottleneck, or can that role pass to a different resource entirely? This section shows that the mechanism survives, and that the bottleneck can indeed be either resource: the pseudo-sample size m, as in Case I, or the student's feature width N_S, which has no counterpart in Case I at all. Which resource is active, and where the handoff between them occurs, is the content of Theorem 4.3 below.

4.1 Model and Two-Stage Training

We adopt the linear random-feature model of Medvedev et al. (2025), in which teacher and student have independently sampled feature maps with controllable widths. Unlike their early-stopping analysis, we train the student to full convergence and ask whether finite pseudo-sample size and feature width alone can induce W2SG. Thus, in the same underlying model, we isolate a resource-bottleneck mechanism complementary to early stopping.

Target and covariance structure. Let $x \sim \mathcal{N}(0, \Sigma)$, where $\Sigma := \text{diag}(I_k, \delta_p I_{p-k})$: the first k coordinates have unit variance and the remaining $p-k$ have variance $\delta_p \in (0,1)$, a sequence that we allow to depend on p and whose asymptotic rate (relative to $p-k$) will matter once we take $p \to \infty$ in Section 4.3. We refer to the first block as the *strong* coordinates and the second as the *weak* coordinates. The target coefficient $\beta^\star \in \mathbb{R}^p$ is deterministic — unlike Case I, where $\beta^\star$ was

random with isotropic second moment — and is supported entirely on the strong block, so $\beta^\star_i = 0$ for $i > k$: the signal we want the student to recover lives entirely in the high-variance directions, while the weak block is pure background. We normalize $\|\beta^\star\|^2_\Sigma = 1$. There is no label noise in this section ($\varepsilon \equiv 0$ in (1)); the only source of teacher error here is a finite random-feature representation, not observation noise, and isolating this source is the point of the model.

Stage I: teacher. Let $U_T \in \mathbb{R}^{N_T \times p}$ have i.i.d. $\mathcal{N}(0,1)$ entries; U_T is drawn once and then held fixed throughout Stage I — it plays the role of a fixed, randomly initialized feature map, not a parameter that Stage I's learning rule optimizes over. Conditional on U_T, the teacher's feature extractor is $f_T(x) = U_T x \in \mathbb{R}^{N_T}$, and its hypothesis class is $\mathcal{H}_T(U_T) := \{h_{T,w} : x \mapsto x^\top U_T^\top w : w \in \mathbb{R}^{N_T}\}$. Rather than fitting n noisy labeled samples as in Case I, the teacher here is idealized as the *population* risk minimizer within $\mathcal{H}_T(U_T)$ — the $n \to \infty$ limit of Stage I training, which lets us isolate representation bias as the sole source of teacher error, uncontaminated by any Stage-I estimation variance:

$$w^\star_T := \operatorname*{argmin}_{w \in \mathbb{R}^{N_T}} \|w\|_2 \quad \text{subject to} \quad U_T \Sigma U_T^\top w = U_T \Sigma \beta^\star. \tag{7}$$

The induced coefficient in the original input space is $\beta^\star_T := U_T^\top w^\star_T$, and the teacher predictor is $h^\star_T(x) := x^\top \beta^\star_T$. Because $w^\star_T$ solves the exact population problem, $h^\star_T$'s only error, conditional on U_T, is approximation bias from the finiteness of N_T; the only remaining randomness, across different draws of U_T, is randomness in *which* N_T-dimensional representation the teacher happened to receive, not sample noise.

Stage II: student. Independently of U_T, draw fresh covariates $\widetilde{x}_1, \ldots, \widetilde{x}_m \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ and assign pseudo-labels $\widetilde{y}_i := h^\star_T(\widetilde{x}_i) = \widetilde{x}_i^\top \beta^\star_T$, giving the Stage-II dataset $\widetilde{\mathcal{D}}_m$ of Section 2, with design matrix $\widetilde{X}$ and label vector $\widetilde{y}$. The student is given its *own* independent random-feature map, $U_S \in \mathbb{R}^{N_S \times p}$ with i.i.d. $\mathcal{N}(0,1)$ entries, independent of U_T and of $\widetilde{\mathcal{D}}_m$, with feature extractor $f_S(x) = U_S x$ and hypothesis class $\mathcal{H}_S(U_S)$ defined analogously to $\mathcal{H}_T(U_T)$. This is the crucial departure from Case I: the student's representation is no longer tied to the teacher's, and the student's width N_S is a resource entirely under our control, with no counterpart on the teacher side. Once N_S is allowed to differ from N_T — in particular, once $N_S > N_T$ — the student is a genuinely stronger model in the ordinary sense of having more capacity, and it is exactly this second resource, alongside m, that gives Case II its two possible bottlenecks. Because $\widetilde{y}$ is generated by $h^\star_T \notin \mathcal{H}_S(U_S)$ in general, it need not be exactly interpolable by the student; Stage II therefore performs minimum-norm *least-squares* regression, rather than minimum-norm interpolation as in Case I:

$$\widehat{w}_S = \operatorname*{argmin}_{w \in \mathbb{R}^{N_S}} \|w\|_2 \qquad \text{subject to} \qquad w \in \operatorname*{argmin}_{v \in \mathbb{R}^{N_S}} \left\| \widetilde{X} U_S^\top v - \widetilde{y} \right\|_2^2, \tag{8}$$

with induced coefficient $\widehat{\beta}_S := U_S^\top \widehat{w}_S$ and student predictor $\widehat{h}_S(x) := x^\top \widehat{\beta}_S$.

Risks. The population risk (2) again takes the closed form $\mathcal{R}(h_\beta; \beta^\star) = (\beta - \beta^\star)^\top \Sigma (\beta - \beta^\star)$, so the averaged risks specialize to

$$\overline{\mathcal{R}}_T = \mathbb{E}_{U_T}[\mathcal{R}(h^\star_T; \beta^\star)], \qquad \overline{\mathcal{R}}_S = \mathbb{E}_{U_T, U_S, \widetilde{\mathcal{X}}_m}\left[\mathcal{R}(\widehat{h}_S; \beta^\star)\right], \tag{9}$$

where $\overline{\mathcal{R}}_S$ additionally averages over Stage-II sample randomness and over the student's own representation U_S.

4.2 Two Stage-II Resources, Three Regimes

As in Case I, whether Stage II can do anything nontrivial at all is governed by a rank condition — but here there are *two* independent Stage-II quantities, m and N_S, that can each be the scarcer one relative to p.

Definition 4.1 (Stage-II bottleneck regimes)**.** Stage II is in the *sample-bottleneck*, *feature-bottleneck*, or *no-bottleneck* regime when $m < \min\{N_S, p\}$, $N_S < \min\{m, p\}$, or $p < \min\{m, N_S\}$, respectively.

Exactly as in Case I, then, W2SG is only possible once *at least one* Stage-II resource is scarce relative to p — both being abundant just reproduces the teacher exactly. What Case II adds is that this scarcity can come from either resource, and the two subsections below show that it produces genuinely different behavior depending on which one it is.

4.3 Asymptotic W2SG Phase Diagram

We work under the same style of deterministic-equivalent analysis introduced in Section 3.4, now applied jointly to the two Stage-II resources, and let all five relevant quantities grow proportionally with p. In the following, ϑ is the signal-coordinate fraction; the corresponding deterministic-equivalent derivation is given in Appendix D.

Assumption 4.2 (Proportional asymptotic scaling)**.** *Let $d_p := p - k$, and assume*

$$k/p \to \vartheta, \qquad d_p\delta_p \to \eta, \qquad N_T/p \to \rho_T, \qquad m/p \to \rho_m, \qquad N_S/p \to \rho_S, \tag{10}$$

where $\vartheta, \rho_T \in (0,1)$, $\eta \in (0,\infty)$, and $\rho_m, \rho_S \in (0,\infty)$. We consider the regular *fixed-ratio regime, in which $\rho_T \neq \vartheta$, $\rho_m \neq \rho_S$, $\rho_m, \rho_S \neq \vartheta$, and $\rho_m, \rho_S \neq 1$; the nontrivial excluded boundary cases are treated in Section D.6 of Appendix D.*

Theorem 4.3 (W2SG phase diagram)**.** *Suppose Assumption 4.2 holds. Define $\Lambda_T := \vartheta - \rho_T$ if $\rho_T < \vartheta$, and $\Lambda_T := \frac{(1-\vartheta)(\rho_T-\vartheta)}{1-\rho_T}$ if $\rho_T > \vartheta$.*

(i) **Necessary condition.** *If $\min\{\rho_m, \rho_S\} < \vartheta$, then $\mathsf{W2SG}^{\mathrm{DE}} < 0$: if either Stage-II resource falls below the signal fraction, the student cannot even recover the signal, let alone improve on the teacher.*

(ii) **Sample-bottleneck regime** *($\vartheta < \rho_m < \min\{1, \rho_S\}$, so m, not N_S, is the scarcer Stage-II resource). Let $\Delta_{\mathrm{DE}} := (\rho_S + 3\vartheta - \Lambda_T)^2 - 16\vartheta\rho_S$. If $\Delta_{\mathrm{DE}} > 0$, then*

$$\mathsf{W2SG}^{\mathrm{DE}} > 0 \iff \rho_m \in \left(\frac{\rho_S + 3\vartheta - \Lambda_T - \sqrt{\Delta_{\mathrm{DE}}}}{4}, \frac{\rho_S + 3\vartheta - \Lambda_T + \sqrt{\Delta_{\mathrm{DE}}}}{4}\right) \cap \big(\vartheta, \min\{1, \rho_S\}\big). \tag{11}$$

If $\Delta_{\mathrm{DE}} \leq 0$, there is no positive region on this branch. This is the direct analogue of the Case I phase diagram: m must lie in a bounded window *for the student to improve on the teacher.*

(iii) **Feature-bottleneck regime** *($\vartheta < \rho_S < \min\{1, \rho_m\}$, so N_S, not m, is now the scarcer Stage-II resource). Here*

$$\mathsf{W2SG}^{\mathrm{DE}} > 0 \iff \rho_S - \vartheta > \Lambda_T \quad \textit{and} \quad \rho_m > \frac{2\rho_S(\rho_S - \vartheta)}{\rho_S - \vartheta - \Lambda_T}. \tag{12}$$

Unlike (ii), the condition on ρ_m is a half-line*: once N_S is the active bottleneck and satisfies $\rho_S - \vartheta > \Lambda_T$, more pseudo-labels are unambiguously good, and W2SG occurs for every m large enough.*

The handoff between the two regimes in the theorem is governed entirely by *which of m, N_S is smaller relative to p*: regime (ii) is exactly $\rho_m < \rho_S$ (with both above ϑ), and regime (iii) is exactly $\rho_S < \rho_m$. There is no third, intermediate mechanism — whichever resource is scarcer is the one whose value must be tuned into a window (or past a threshold) for W2SG to occur, and the theorem's two branches are this dichotomy made precise.

Remark 4.4 (Reading the phase diagram). Theorem 4.3 yields three main consequences. First, positive W2SG requires $\rho_S > \vartheta + \Lambda_T > \rho_T$, so a student capacity advantage is necessary but not sufficient. Second, in the sample-bottleneck regime, m must lie in a bounded window, whereas in the feature-bottleneck regime it only needs to exceed a threshold. Third, these regimes illustrate the active-bottleneck principle: the scarcer resource controls how much teacher error is filtered, while the non-bottleneck resource primarily reduces estimation noise and can be increased without harming W2SG.

The next corollary rearranges Theorem 4.3 to hold ρ_m fixed and vary ρ_S instead, exhibiting the same two regimes from the feature-width side.

Corollary 4.5 (Phase diagram in student width)**.** *Suppose the conditions of Theorem 4.3 hold and fix $\rho_m > \vartheta$.*

(i) **Feature-bottleneck regime** *($\vartheta < \rho_S < \min\{1, \rho_m\}$). Let $D_S := (\rho_m + 2\vartheta)^2 - 8\rho_m(\vartheta + \Lambda_T)$. If $D_S > 0$,*

$$\mathsf{W2SG}^{\mathrm{DE}} > 0 \iff \rho_S \in \left(\frac{\rho_m + 2\vartheta - \sqrt{D_S}}{4}, \frac{\rho_m + 2\vartheta + \sqrt{D_S}}{4}\right) \cap \big(\vartheta, \min\{1, \rho_m\}\big); \tag{13}$$

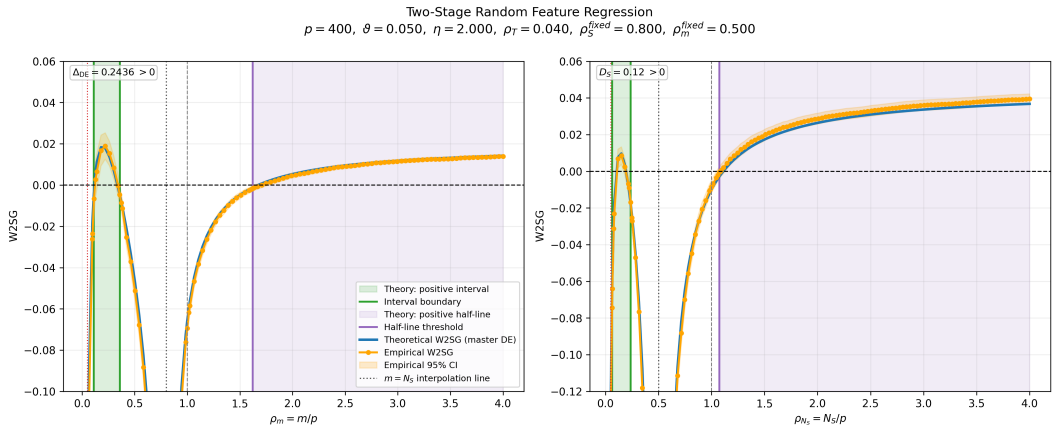


Figure 1: **Two-stage random-feature regression.** Empirical and deterministic-equivalent W2SG for $p = 400$, $\vartheta = 0.05$, $\eta = 2$, $\rho_T = 0.04$, over 200 Monte Carlo trials. Left: W2SG versus $\rho_m = m/p$ at fixed $\rho_S = 0.8$. Right: W2SG versus $\rho_S = N_S/p$ at fixed $\rho_m = 0.5$. Blue: finite-p master deterministic equivalent. Orange: empirical mean with 95% confidence band. Green lines: predicted bounded positive-W2SG interval. Purple line: threshold of the predicted positive half-line. Gray dotted line: the interpolation boundary $m = N_S$. Red dotted line: the signal threshold $\rho = \vartheta$. Gray dashed line: the ambient-rank boundary $\rho = 1$.

if $D_S \le 0$, there is no positive region on this branch. The feature-bottleneck positive region is thus a bounded interval *in ρ_S — too little width starves the student, too much lets it overfit the teacher's bias.*

(ii) **Sample-bottleneck regime** *($\vartheta < \rho_m < \min\{1, \rho_S\}$). If $\rho_m \le 2\vartheta$, there is no positive region on this branch. If $\rho_m > 2\vartheta$,*

$$\mathrm{W2SG}^{\mathrm{DE}} > 0 \iff \rho_S > \frac{\rho_m(2\rho_m - 3\vartheta + \Lambda_T)}{\rho_m - 2\vartheta}. \tag{14}$$

The sample-bottleneck positive region is thus a half-line *in ρ_S: once m is already the active bottleneck, additional student width is purely beneficial.*

Corollary 4.5 is the mirror image of Theorem 4.3: interval versus half-line swaps sides depending on which variable is held fixed and which is scanned, but the underlying division is the same one identified in Remark 4.4 — whichever resource is currently the bottleneck needs a window; the other only needs to clear a threshold. Figure 1 provides numerical validation of both results. Together, Theorem 4.3 and Corollary 4.5 formalize the *active-bottleneck principle*: the scarcer Stage-II resource controls teacher-error filtering and must lie in an appropriate range, while the non-bottleneck resource only reduces estimation noise once it clears its threshold. Case I is the single-resource special case, in which only m can bind and the sample-bottleneck branch of Theorem 4.3 remains.

5 Conclusion

We showed that W2SG can arise from a finite Stage-II resource constrained to an intermediate range, without early stopping or explicit regularization and, in Case I, without any student capacity advantage. In ridgeless linear regression with identical teacher and student hypothesis classes, the pseudo-sample size m alone determines whether the student improves on its teacher, with explicit thresholds in the fixed-sample and proportional regimes. In the random-feature setting, the same trade-off can be governed by either pseudo-sample size or student width, yielding an *active-bottleneck principle*: the scarcer resource controls teacher-error filtering, while the non-binding resource can be increased freely. We conjecture that these phase transitions extend beyond our power-law and two-block covariance models to broader graded non-isotropic spectra, and ask whether the active-bottleneck principle extends to nonlinear models, other resources such as training time, and large-scale W2SG settings exhibiting similar non-monotonic dependence on pseudo-sample size.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery, organization and language editing of the abstract, introduction, and related-work discussion, and to provide feedback on manuscript structure and experimental design. All citations, mathematical claims, proofs, code, and reported results were independently verified by the authors. The authors take responsibility for the final content.

REFERENCES

Alexander Atanasov, Blake Bordelon, Jacob A Zavatone-Veth, Courtney Paquette, and Cengiz Pehlevan. Two-point deterministic equivalence for stochastic gradient dynamics in linear models. *arXiv preprint arXiv:2502.05074*, 2025.

Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *Journal of Statistical Mechanics: Theory and Experiment*, 2026(4): 043404, 2026.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024.

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

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. *arXiv preprint arXiv:2502.08606*, 2025.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational mathematics*, 7(3):331–368, 2007.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. *Advances in neural information processing systems*, 37:126474–126499, 2024.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52 (6):2879–2912, 2024.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. In *International Conference on Learning Representations*, volume 2026, pages 108893–108934, 2026.

Yijun Dong, Yicheng Li, Yunai Li, Jason D Lee, and Qi Lei. Discrepancies are virtue: Weak-to-strong generalization through lens of intrinsic dimension. *arXiv preprint arXiv:2502.05075*, 2025.

Scott Geng, Dutch Hansen, and Jerry Li. Weak-to-strong generalization is nearly inevitable (in linear models). *arXiv preprint arXiv:2605.05742*, 2026.

Qiyang Han and Xiaocong Xu. The distribution of ridgeless least squares interpolators. *Journal of Machine Learning Research*, 27(23):1–94, 2026.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Muhammed Ildiz, Halil Gozeten, Ege 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*, volume 2025, pages 27181–27220, 2025.

Myeongho Jeon, Jan Sobotka, Suhwan Choi, and Maria Brbic. Weak-to-strong generalization under distribution shifts. *Advances in Neural Information Processing Systems*, 38:153420–153458, 2026.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. *Advances in neural information processing systems*, 37:46837–46880, 2024.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-to-strong generalization even in random feature networks, provably. *arXiv preprint arXiv:2503.02877*, 2025.

Behrad Moniri and Hamed Hassani. On the mechanisms of weak-to-strong generalization: A theoretical perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 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.

Junsoo Oh, Jerry Song, and Chulhee Yun. From linear to nonlinear: Provable weak-to-strong generalization through feature learning. *Advances in Neural Information Processing Systems*, 38: 169216–169285, 2026.

Martin Pawelczyk, Lillian Sun, Zhenting Qi, Aounon Kumar, and Himabindu Lakkaraju. Generalizing trust: Weak-to-strong trustworthiness in language models. *arXiv preprint arXiv:2501.00418*, 2024.

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, pages 3889–3897. PMLR, 2021.

Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-centric lens. In *International Conference on Learning Representations*, volume 2025, pages 22039–22077, 2025.

Seamus Somerstep, Felipe Maia Polo, Moulinath Banerjee, Yaacov Ritov, Mikhail Yurochkin, and Yuekai Sun. A transfer learning framework for weak to strong generalization. In *International Conference on Learning Representations*, volume 2025, pages 100714–100748, 2025.

Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

David Wu and Anant Sahai. Provable weak-to-strong generalization via benign overfitting. In *International Conference on Learning Representations*, volume 2025, pages 88280–88316, 2025.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, volume 33, pages 10112–10123, 2020.

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:2603.05691*, 2026.

Gengze Xu, Wei Yao, Ziqiao Wang, and Yong Liu. On the emergence of weak-to-strong generalization: A bias-variance perspective. *arXiv preprint arXiv:2505.24313*, 2025.

Yihao Xue, Jiping Li, and Baharan Mirzasoleiman. Representations shape weak-to-strong generalization: Theoretical insights and empirical predictions. *arXiv preprint arXiv:2502.00620*, 2025.

Wenkai Yang, Shiqi Shen, Guangyao Shen, Wei Yao, Yong Liu, Gong Zhi, Yankai Lin, and Ji-Rong Wen. Super (ficial)-alignment: Strong models may deceive weak models in weak-to-strong generalization. In *International Conference on Learning Representations*, volume 2025, pages 3271–3295, 2025.

Yuqing Yang, Yan Ma, and Pengfei Liu. Weak-to-strong reasoning. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 8350–8367, 2024.

Wei Yao, Wang Zhaoyang, Gengze Xu, Chen Qian, Dongrui Liu, Ziqiao Wang, Yong Liu, and Yunbei Xu. On the blessing of pre-training in weak-to-strong generalization. *arXiv preprint arXiv:2605.05710*, 2026.

Ruimeng Ye, Yang Xiao, and Bo Hui. Weak-to-strong generalization beyond accuracy: a pilot study in safety, toxicity, and legal reasoning. *arXiv preprint arXiv:2410.12621*, 2024.

A Experimental Results

We validate the theory against Monte Carlo simulations by comparing the finite-p deterministic equivalents with empirically measured W2SG. The experiments are organized around the two theoretical settings studied in the paper. For Case I, we examine the ridgeless-regression phase diagram on the two sides of the critical exponent $\alpha = 2$ and then study the finite-size behavior associated with the fixed-m and proportional scalings. For Case II, we validate the random-feature phase diagram, separate the sample- and feature-bottleneck mechanisms, and include a negative-control experiment in which the theory predicts no admissible positive-W2SG region.

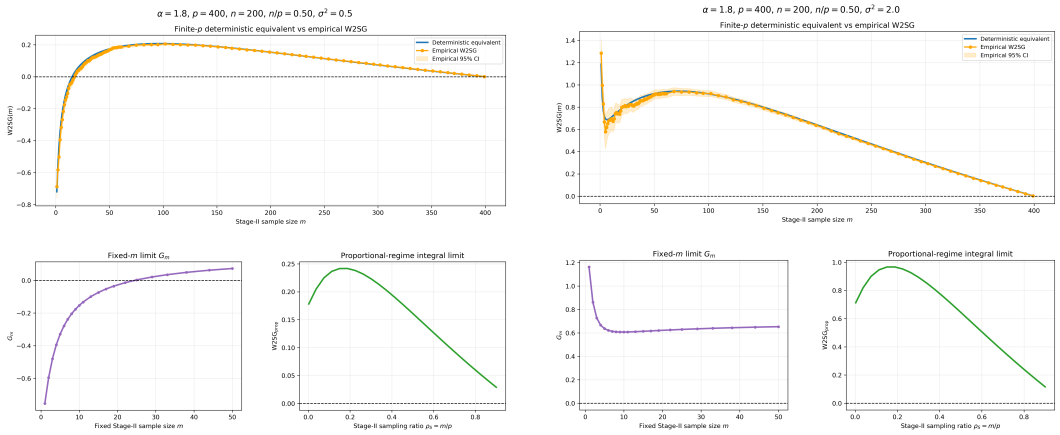
A.1 Case I: Ridgeless Regression

We first consider the ridgeless-regression setting of Theorem 3.4. We examine the W2SG phase behavior on the two sides of the critical exponent $\alpha = 2$, comparing the finite-p deterministic equivalent with Monte Carlo simulations and with the fixed-m and proportional asymptotic limits. We then study finite-size scaling to illustrate how these two asymptotic regimes emerge under different scalings of the Stage-II sample size m.

A.1.1 Phase Diagram

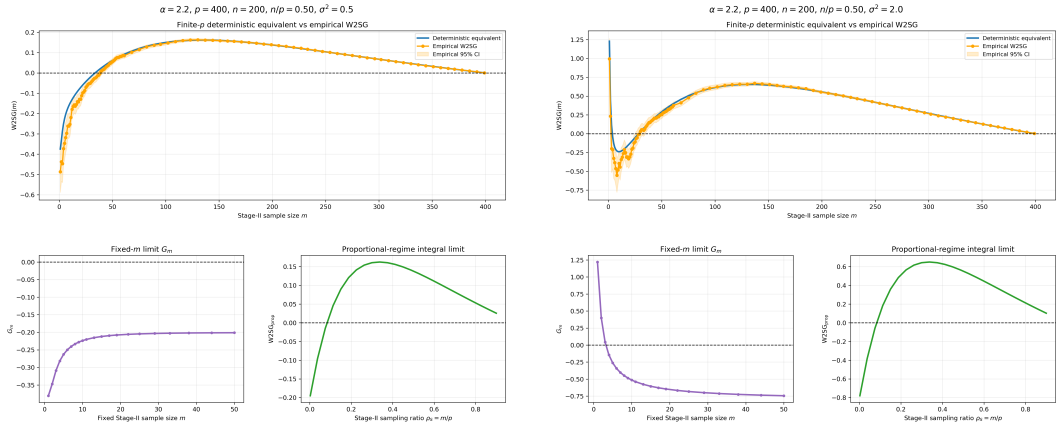
We first validate the Case I deterministic equivalent and phase diagram at $p = 400$ and $n/p = 0.5$. We consider $\alpha = 1.8$ and $\alpha = 2.2$, representing the two sides of the critical exponent $\alpha = 2$, and use $\sigma^2 \in \{0.5, 2\}$. For each setting, we compare the finite-p deterministic equivalent with empirical Monte Carlo W2SG and also display the fixed-m and proportional asymptotic limits.

Regime $1 < \alpha < 2$. Figure 2 considers $\alpha = 1.8$ and contrasts two noise levels. At $\sigma^2 = 0.5$ (Figure 2a), the finite-p deterministic equivalent is negative for the smallest pseudo-sample sizes and crosses zero at approximately $m = 15.25$; the empirical curve crosses at approximately $m = 17.39$. The fixed-sample limit exhibits the same eventual sign change, but at the larger value $m \approx 24.47$. By contrast, the proportional-regime limit is positive throughout the displayed interval of fixed sampling ratios $m/p > 0$. These observations illustrate the distinction between the two asymptotic scalings in Theorem 3.4: small fixed m can remain unfavorable even though every fixed positive proportional ratio is favorable when $\alpha \leq 2$. At the higher noise level $\sigma^2 = 2$ (Figure 2b), the gain is positive throughout all three displayed regimes. The finite-p curve is nevertheless strongly non-monotone: it decreases sharply at very small m, rises over an intermediate range, and then decreases toward zero as m approaches p. The fixed-m and proportional limits display the same broad spectral trade-off at their respective scales. In both noise settings, the finite-p deterministic equivalent follows the Monte Carlo mean closely, including through the small-m transition and the non-monotone region.



(a) $\sigma^2 = 0.5$. (b) $\sigma^2 = 2$.

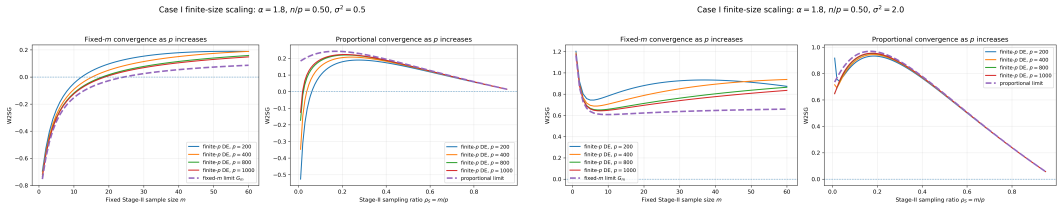
Figure 2: **Case I with $1 < \alpha < 2$.** Numerical validation at $\alpha = 1.8$, $p = 400$, and $n = 200$ using 1000 Monte Carlo trials. In each subfigure, the top panel compares the finite-p deterministic equivalent with empirical W2SG and its 95% confidence interval; the bottom-left panel shows the fixed-m limit G_m; and the bottom-right panel shows the proportional-regime limit as a function of the Stage-II sampling ratio.



(a) $\sigma^2 = 0.5$. (b) $\sigma^2 = 2$.

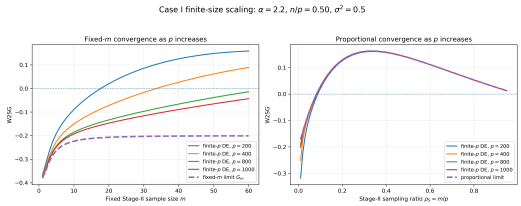
Figure 3: **Case I with $\alpha > 2$.** Numerical validation at $\alpha = 2.2$, $p = 400$, and $n = 200$ using 1000 Monte Carlo trials. The panel layout is the same as in Figure 2: finite-p deterministic equivalent versus empirical W2SG on top, fixed-m limit on the bottom left, and proportional-regime limit on the bottom right.

Regime $\alpha > 2$. Figure 3 repeats the comparison at $\alpha = 2.2$. At $\sigma^2 = 0.5$ (Figure 3a), the fixed-m limit G_m remains negative throughout the displayed range, whereas the proportional limit is negative for small sampling ratios and becomes positive only after a critical ratio $\rho_{m,\mathrm{crit}} \approx 0.0840$. At $p = 400$, this predicts a critical pseudo-sample size $p\rho_{m,\mathrm{crit}} \approx 33.6$, essentially coinciding with the finite-p deterministic-equivalent crossing $m \approx 33.64$. The empirical crossing, $m \approx 38.36$, is slightly shifted at this finite dimension but exhibits the same transition. This is the clearest numerical example of the qualitative disagreement between the fixed-sample and proportional regimes for $\alpha > 2$: a fixed number of pseudo-samples remains unfavorable, while a sufficiently large $\Theta(p)$-scale pseudo-sample size yields positive W2SG. At $\sigma^2 = 2$ (Figure 3b), the phase structure is richer. The finite-p deterministic equivalent is positive at the very smallest m, becomes negative over an intermediate small-m interval, and becomes positive again, with zero crossings at approximately $m = 3.73$ and $m = 28.38$. The empirical curve shows the same disconnected structure, with crossings near $m = 2.54$ and $m = 29.14$. The fixed-m limit likewise changes sign near $m = 3.23$

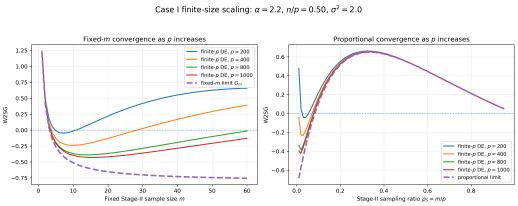


(a) $\alpha = 1.8$, $\sigma^2 = 0.5$.

(b) $\alpha = 1.8$, $\sigma^2 = 2$.



(c) $\alpha = 2.2$, $\sigma^2 = 0.5$.



(d) $\alpha = 2.2$, $\sigma^2 = 2$.

Figure 4: **Finite-size convergence in Case I.** Finite-p deterministic equivalents for $p \in \{200, 400, 800, 1000\}$ with $n/p = 0.5$, compared with the two asymptotic limits in Theorem 3.4. In each subfigure, the left panel keeps the absolute Stage-II sample size m fixed and compares $\mathsf{W2SG}_p^{\mathrm{DE}}(m)$ with the fixed-m limit G_m (purple dashed curve). The right panel keeps the sampling ratio $\rho_m = m/p$ fixed and compares the finite-p curves with the proportional-regime limit (purple dashed curve). The four subfigures use the same (α, σ^2) pairs as the main Case I experiments.

and is negative thereafter over the displayed range, whereas the proportional limit again crosses from negative to positive near $\rho_m \approx 0.084$. This is the finite-dimensional manifestation of the disconnected positive-W2SG regions discussed in Remark 3.5: improvement can occur at very small m and again at proportional scale, while an intermediate range remains unfavorable.

A.1.2 Finite-Size Scaling

Figure 4 checks directly that the two asymptotic formulas arise from two different large-p scalings of the same finite-p deterministic equivalent. In the left panels, m is held fixed as p increases. The finite-p curves move toward the fixed-sample limit G_m; for $\alpha = 1.8$ and $\sigma^2 = 0.5$, this is visible in the rightward motion of the finite-p zero crossing toward the crossing of G_m. For $\alpha = 2.2$ and $\sigma^2 = 0.5$, increasing p instead drives the fixed-m curve downward toward the negative limiting branch, reinforcing that fixed m does not access the positive proportional regime. In the right panels, m/p is held fixed. Away from the small-ρ_m boundary, the finite-p curves collapse rapidly onto the proportional-limit prediction as p increases. For $\alpha = 1.8$, the limiting proportional curve remains positive for both noise levels, in agreement with Theorem 3.4(a)(i). For $\alpha = 2.2$, the curves converge to the proportional phase transition predicted by Theorem 3.4(a)(ii). The $\sigma^2 = 2$ setting has the largest finite-size corrections near $\rho_m = 0$, precisely where the proportional scaling approaches the fixed-m/sparse boundary. Taken together, the four scaling experiments suggest that the asymptotic behavior of the deterministic equivalent depends critically on how the Stage-II sample size m scales with the ambient dimension p. In particular, fixed-m and proportional-m limits capture different high-dimensional regimes, highlighting the need to analyze distinct scalings of m when characterizing W2SG asymptotically.

A.2 Case II: Random-Feature Regression

We next examine the random-feature setting of Theorem 4.3 and Corollary 4.5, where the Stage-II learner is controlled by both the pseudo-sample size m and the student feature width N_S. We first validate the predicted phase diagram, then isolate the two bottleneck mechanisms through a branch-separation experiment, and finally consider a negative-control setting in which the theory predicts no admissible positive-W2SG region.

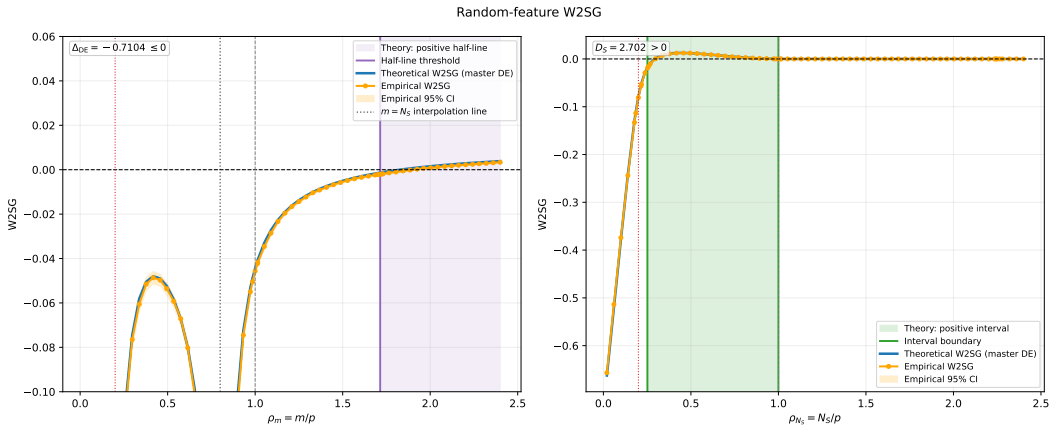


Figure 5: **Case II branch separation.** A configuration in which the two scans realize different components of the phase diagram. Left: the bounded positive interval in ρ_m is absent, while the positive half-line remains. Right: a bounded positive interval in ρ_S survives. The plotting conventions are the same as in Figure 1. Results use 500 Monte Carlo trials.

A.2.1 Phase Diagram

Figure 1 tests both branches of Theorem 4.3 and Corollary 4.5 at $p = 400$, $\vartheta = 0.05$, $\eta = 2$, $\rho_T = 0.04$. Sweeping ρ_m at fixed $\rho_S = 0.8$ (left panel) recovers the sample-bottleneck branch: a bounded positive-W2SG interval, exactly as in Theorem 4.3(ii). Sweeping ρ_S at fixed $\rho_m = 0.5$ (right panel) recovers the feature-bottleneck branch instead: a bounded interval in ρ_S per Corollary 4.5(i), with positivity resuming as a half-line once ρ_S grows large enough to make m the active bottleneck again. In both panels, the empirical curve tracks the finite-p master deterministic equivalent closely away from the internal interpolation boundary $m = N_S$, where the risk is singular and finite-dimensional fluctuations are strongly amplified, exactly as predicted by the exact interpolation classification of Appendix D.

A.2.2 Separation of the Two Phase Mechanisms

Figure 5 separates the two phase mechanisms by choosing a configuration in which one bounded interval disappears while another branch survives. The headline setting displays several positive branches simultaneously. To verify that these branches are genuinely distinct rather than different representations of the same phenomenon, we take $p = 400$, $\vartheta = 0.2$, $\eta = 2$, and $\rho_T = 0.16$, with $\rho_S = 0.8$ in the ρ_m-scan and $\rho_m = 2.25$ in the ρ_S-scan.

For the left scan in Figure 5, the interval discriminant satisfies $\Delta_{\mathrm{DE}} < 0$, so no bounded positive-ρ_m interval exists. Nevertheless, the half-line branch remains admissible, with the asymptotic threshold at $\rho_m \simeq 1.714$. The finite-p deterministic-equivalent and empirical first zero crossings occur at approximately $\rho_m^{\mathrm{DE}} \simeq 1.827$ and $\rho_m^{\mathrm{emp}} \simeq 1.902$, respectively. In contrast, the right scan has $D_S > 0$ and retains a bounded interval $\rho_S \in (0.252, 1)$. The corresponding finite-p and empirical transitions, approximately 0.294 and 0.296, are nearly coincident. This experiment therefore isolates the two geometric components of the phase diagram: the bounded interval can disappear without eliminating the positive half-line, while the complementary scan can simultaneously retain its bounded positive branch.

A.2.3 Negative Control

Figure 6 provides a negative control in which the branch constraints exclude all positive-W2SG regions. We retain $p = 400$, $\vartheta = 0.05$, $\eta = 2$, and $\rho_T = 0.04$, but fix both Stage-II resources below the signal threshold, $\rho_S^{\mathrm{fixed}} = \rho_m^{\mathrm{fixed}} = 0.025 < \vartheta$. Although the raw quadratic discriminants are positive in the configuration of Figure 6, their formal roots lie outside the admissible bottleneck branches once the signal-threshold and active-resource constraints are enforced. Consequently, neither scan contains a positive-W2SG interval or half-line. Consistent with this prediction, both the

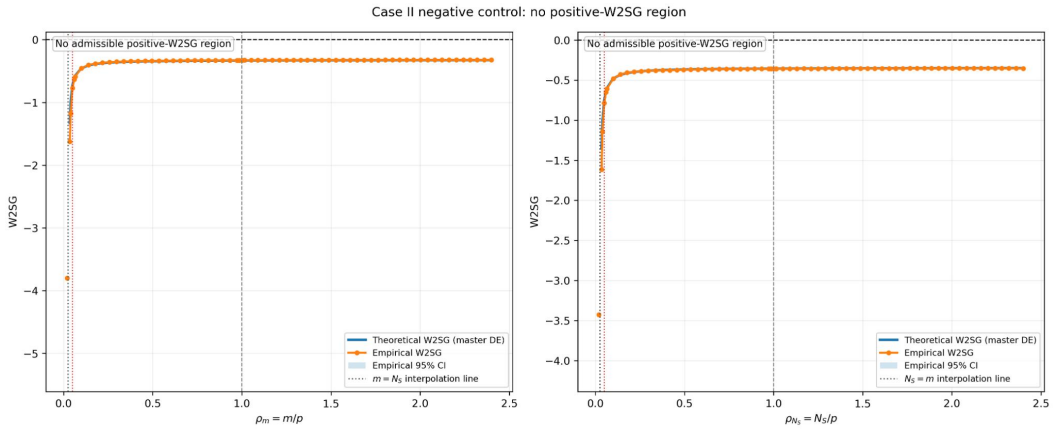


Figure 6: **Case II negative control.** A configuration in which the fixed Stage-II resource lies below the signal threshold, so neither scan admits an admissible positive-W2SG region. Left: W2SG as a function of $\rho_m = m/p$ at fixed ρ_S. Right: W2SG as a function of $\rho_S = N_S/p$ at fixed ρ_m. In both scans the finite-p master deterministic equivalent remains strictly negative throughout the displayed range and closely follows the empirical Monte Carlo mean. The gray dotted line denotes the interpolation boundary $m = N_S$ and the red dotted line denotes the signal threshold $\rho = \vartheta$. Shaded bands are empirical 95% confidence intervals.

deterministic equivalent and the empirical Monte Carlo curves remain strictly below zero throughout the scanned domains. This negative control is useful because it shows that the phase theory does not merely fit observed sign changes: it also predicts settings in which improvement is absent altogether.

B Related Work

A growing theoretical literature explains W2SG through different lenses. We organize it along three axes bearing directly on our contribution: *what mechanism* drives improvement, *whether the student has a capacity advantage* over the teacher, and *which resource*, if any, is identified as the source of the effect.

Mechanisms proposed for W2SG. One line of work gives structural conditions guaranteeing improvement. Charikar et al. (2024) relate improvement to the student's *misfit* to the teacher's labels under realizability and convexity; Mulgund and Pabbaraju (2025) extend this to Bregman divergences, and Xu et al. (2025) give a bias–variance decomposition connecting improvement to how well the student approximates a posterior-mean teacher. Lang et al. (2024) use expansion-type conditions under which the student corrects and extends the teacher's pseudo-label coverage, and Shin et al. (2025) study how easy/hard data overlap supports this. Somerstep et al. (2025) cast W2SG as latent-concept transfer. These frameworks isolate *properties of a fixed predictor pair* — misfit, coverage, alignment — sufficient for improvement. We instead fix a standard training procedure at both stages and ask how the *resources* supplied to it determine improvement, without assuming any property of the resulting predictors.

A second line explains W2SG through training dynamics. Medvedev et al. (2025) show that early-stopped population gradient flow in random-feature regression yields improvement, because finite training time suppresses teacher error before the student fully absorbs it; Geng et al. (2026) show a related effect for finite-time logistic training. Both rely on stopping *before* convergence. We study the complementary regime: our student always converges, and the constraint that produces improvement is on the data or features it receives, not on training time.

Symmetric versus asymmetric capacity. Several analyses assume the student has an advantage the teacher lacks. Moniri and Hassani (2025) show a student can compensate for a teacher that is under-regularized relative to it, so asymmetric regularization is itself the source of improvement. Xue et al. (2025) characterize teacher-error propagation through principal-kernel geometry, allow-

ing the student's representation to differ favorably from the teacher's. Dong et al. (2025) derive an inherited-variance decomposition governed by teacher–student representation overlap, again asymmetric. Oh et al. (2026) let the student be a nonlinear network trained beyond a linear convolutional teacher. In each case, some asymmetry — regularization, representation quality, or nonlinearity — does part of the work. Our Case I removes this axis entirely: teacher and student share one hypothesis class, one ridgeless rule, and one convergence guarantee. Case II reintroduces capacity asymmetry deliberately, via $N_S > N_T$, to test whether a resource-based mechanism survives alongside it.

Resource-based explanations and scaling. Closest to our framing are analyses treating sample size or feature count as explicit, finite parameters. Ildiz et al. (2025) characterize surrogate predictors that improve a downstream ridgeless learner relative to training directly on the target; their benchmark is improvement over direct target training, whereas ours is improvement over the teacher Stage I actually produces. Wu et al. (2026) derive deterministic equivalents for random-feature *ridge* regression and find improved scaling exponents via a joint choice of sample size, width, and ridge; larger sample sizes and widths are generally beneficial once the ridge is tuned. Our Case II student is ridgeless throughout, and on the sample-bottleneck branch W2SG holds only inside a bounded interval of m — more pseudo-labels eventually destroy it. Wu and Sahai (2025) study a structurally similar two-stage ridgeless pipeline, but their source of improvement is a built-in asymmetry: the student's covariates are strictly stronger than the teacher's throughout. Our Case I needs no such asymmetry at all, and our Case II treats a representational advantage as a separate, explicit parameter analyzed jointly with m. Yao et al. (2026) show a shared pretrained initialization can induce W2SG in a single-index model — a different, complementary source of advantage rather than a competing explanation. None of these works identifies a single scarce resource governing a sharp transition between improvement and its absence.

Overparameterization: helpful, harmful, or beside the point? A further group discusses overparameterization directly without explaining why it helps or hurts. Wu et al. (2026) show that, with ridge regularization, the interplay between the ridge and the degree of overparameterization can improve W2SG scaling laws. Dong et al. (2025) find, in kernel regression with an assumed teacher–student discrepancy, that W2SG occurs both overparameterized-ridgeless and underparameterized-with-ridge. Both leave open *why* overparameterization itself helps. Moniri and Hassani (2025) argue the opposite: without regularization, an overparameterized student loses Stage-I information that a regularized student would retain, which they treat as harmful.

Our Case I agrees that information is lost — this is exactly Proposition 3.1 — but shows the loss is not intrinsically harmful. Whether it helps depends on which directions of parameter space it comes from, which depends on the covariance structure. Under power-law covariance, Theorem 3.4 shows this same loss produces $\mathsf{W2SG} > 0$ throughout the overparameterized range for $\alpha \leq 2$, and on an explicit sub-range for $\alpha > 2$. Overparameterization is neither uniformly good nor bad here: it is the mechanism, and its effect depends on where the discarded information lives spectrally.

Evidence from LLM-scale distillation. Busbridge et al. (2025) report an extensive empirical study of distillation scaling laws and find that the number of teacher-labeled examples m has a non-monotone effect: beyond a point, more pseudo-labels cause overfitting to the teacher and drive W2SG toward zero, even with a large capacity gap remaining. This matches the upper half of our own phase diagrams and is, to our knowledge, the first evidence that this shape holds at LLM scale rather than only in linear models. We do not see this as being in tension with our interval characterization: they report the shape empirically, with no mechanism or explicit threshold, which is exactly what Theorems 3.4 and 4.3 supply; and an interval with a finite upper endpoint already implies "large m is eventually bad," so the two claims agree rather than compete.

Alignment and trustworthiness. A complementary empirical literature asks whether W2SG extends beyond predictive accuracy to practically important alignment and trustworthiness properties, including safety, toxicity, legal reasoning, fairness, adversarial and out-of-distribution robustness, reasoning, and privacy (Ye et al., 2024; Pawelczyk et al., 2024; Yang et al., 2024; Jeon et al., 2026). These studies generally find that W2SG can transfer several such properties, often with auxiliary losses or regularization, but that safety-related behavior is substantially harder to recover than standard task performance. Complementing these results, Yang et al. (2025) identify *weak-to-strong*

deception: stronger models supervised by weaker ones can concentrate conflict-induced misaligned behavior in regions that the supervisor cannot reliably evaluate. This effect becomes more pronounced as the capability gap grows, while supervision through an intermediate model partially mitigates it.

What is new here. No prior analysis combines all three features of our setting: (i) teacher and student trained by the *same procedure* to *exact convergence*, with no early stopping or ridge; (ii) in Case I, no capacity or regularization asymmetry; and (iii) an explicit, closed-form characterization of the resource level — m in Case I, jointly m and N_S in Case II — separating improvement from its absence, rather than a sufficient condition or scaling exponent. A resource constraint shared between the two stages can itself regularize the student, and this constraint must be *active* — neither too loose nor too tight — for improvement to occur; the non-monotonicity of Busbridge et al. (2025) suggests this picture is not an artifact of the linear models we analyze.

C Two-Stage Minimum-Norm Interpolation Theory

This section develops the theoretical analysis of the two-stage minimum-norm interpolation model used in Section 3. In this setting, the teacher and student operate in the same ambient linear hypothesis space, but are trained on different samples: Stage I fits a ridgeless estimator to noisy labeled data, while Stage II fits a second ridgeless estimator to fresh covariates pseudo-labeled by the Stage-I teacher. The central question is whether this second interpolation step can improve population prediction risk despite receiving no information beyond the teacher-generated labels.

We first develop the deterministic-equivalent description of the two-stage procedure in the proportional high-dimensional regime. After recording the resolvent deterministic equivalence used throughout the analysis, we derive the Stage-I prediction-risk formula and then analyze Stage II conditionally on the realized teacher. This separates the student risk into the Stage-I error carried into the second stage, the additional error generated by Stage-II interpolation, and the cross term produced by the interaction between the two stages. A weighted ridgeless pseudoinverse deterministic equivalent then allows the remaining Stage-I randomness to be evaluated. Under the additional isotropic random-teacher assumption, these calculations yield a fully deterministic expression for the Stage-I-to-Stage-II risk improvement in terms of the population spectrum and the effective Stage-I and Stage-II resolution parameters.

We then specialize the deterministic-equivalent formula to a power-law covariance spectrum and study its asymptotic sign. The fixed-point equations for the effective resolutions are reduced to limiting integral equations, and the trace terms in the W2SG gain are analyzed separately before being combined into a scalar sign criterion. This leads to the proportional W2SG phase diagram and reveals a qualitative change at the spectral-decay exponent $\alpha = 2$: the regime $1 < \alpha \leq 2$ exhibits positive deterministic-equivalent gain throughout the nontrivial proportional Stage-II range, whereas for $\alpha > 2$ the gain can change sign as the Stage-II sampling ratio varies. The subsequent analysis characterizes the existence and, where established, uniqueness of these proportional transitions.

Finally, we move beyond fixed proportional Stage-II sampling. We first consider a sparse-growing Stage-II sample size and derive the corresponding deterministic-equivalent limit. We then study the ultra-sparse regime in which the Stage-II sample count is fixed while the ambient dimension diverges. Because this regime lies outside the standard proportional deterministic-equivalence theorem, its analysis is stated under an explicit continuation assumption. Within that framework, we derive the fixed-m gain, identify noise- and spectral-decay-dependent sign transitions, obtain critical and large-m asymptotics, and give explicit conditions for W2SG and non-W2SG behavior. Together, these results describe how the benefit of the second interpolation stage depends on spectral decay, label noise, and the amount of pseudo-labeled Stage-II data across proportional, sparse-growing, and fixed-sample regimes.

C.1 Deterministic Equivalence

We begin by stating the deterministic-equivalence results that form the basis of our analysis Atanasov et al. (2026; 2025). These results apply in proportional high-dimensional asymptotic

regimes, where the sample size and feature dimension grow at comparable rates. Any additional assumptions required in the subsequent derivations are stated explicitly.

Lemma C.1 (Resolvent deterministic equivalence)**.** *We adopt the notation and definition of* strong deterministic equivalence *used by (Atanasov et al., 2026, Sec. II.E.3). In particular, the symbol $\simeq$ below has the same test-matrix meaning as in that reference.*

Let $\{A_p\}_{p\geq 1}$ be a sequence of random $p\times p$ matrices and let $\{B_p\}_{p\geq 1}$ be a sequence of deterministic $p \times p$ matrices. Following (Atanasov et al., 2026, Sec. II.E.3), we write

$$A_p \simeq B_p$$

if, for every deterministic test matrix M_p satisfying

$$\sup_p \|M_p\|_{\mathrm{op}} < \infty,$$

one has

$$\frac{\mathrm{tr}(A_p M_p)}{\mathrm{tr}(B_p M_p)} \xrightarrow[p\to\infty]{\mathbb{P}} 1.$$

The same notion is used in (Atanasov et al., 2026, Eqs. (18)–(20)) to obtain strong deterministic equivalents for empirical-covariance resolvents. For the sample covariance matrix, this framework yields the following deterministic equivalents in the proportional high-dimensional regime:

$$(\widehat{\Sigma}_p + \lambda I)^{-1} \simeq \frac{\tau}{\lambda}(\Sigma_p + \tau I)^{-1}. \tag{DE1}$$

$$\widehat{\Sigma}_p(\widehat{\Sigma}_p + \lambda I)^{-1} \simeq \Sigma_p(\Sigma_p + \tau I)^{-1}. \tag{DE2}$$

where

$$x_i \sim \mathcal{N}(0, \Sigma_p), \qquad \widehat{\Sigma}_p = \frac{1}{n_p}\sum_{i=1}^{n_p} x_i x_i^\top, \qquad p, n_p \to \infty, \qquad \frac{p}{n_p} \to \phi \in (0, \infty).$$

Here $\lambda > 0$ is the ridge regularization parameter and $\tau = \tau(\lambda)$ denotes the corresponding renormalized ridge parameter in the proportional regime. Its precise characterization will be introduced later in the derivation. The ridgeless regime is obtained, when appropriate, by subsequently taking the limit $\lambda \downarrow 0$. In the overparameterized regime $\phi > 1$, $\tau(\lambda)$ may remain nonzero in this limit, whereas in the underparameterized regime $\phi < 1$, one has $\tau(\lambda) \to 0$.

We now introduce the two-stage ridgeless linear-regression model with Gaussian covariates that will be used throughout the subsequent analysis.

C.2 Two-Stage Linear Regression with Gaussian Covariates

Let

- n be the number of data points in Stage I,
- m be the number of data points in Stage II,
- p be the dimension of the feature space.

We consider the linear model

$$y = X\beta^\star + \varepsilon, \tag{15}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix, $\beta^\star \in \mathbb{R}^p$ is the true parameter vector, and $\varepsilon \in \mathbb{R}^n$ is the label-noise vector.

For the time being, $\beta^\star$ is treated as deterministic; later, when a random-teacher model is required, the corresponding distributional assumptions will be stated explicitly.

The rows $x_i^\top$ of X satisfy $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p\times p}$ is the population covariance matrix, assumed positive definite, $\Sigma \succ 0$. The noise components are independent of the design and satisfy $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. For an independent test covariate $x \sim \mathcal{N}(0, \Sigma)$ and any $\widehat{\beta} \in \mathbb{R}^p$, we define the population prediction risk by

$$\mathcal{R}(h_{\widehat{\beta}}) := \mathbb{E}_x\left[\left(x^\top\widehat{\beta} - x^\top\beta^\star\right)^2\right] = (\widehat{\beta} - \beta^\star)^\top\Sigma(\widehat{\beta} - \beta^\star) = \left\|\Sigma^{1/2}(\widehat{\beta} - \beta^\star)\right\|_2^2. \tag{16}$$

C.3 Stage I: Ridgeless Regression

Stage-I ridgeless regression. In Stage I, we fit the ridgeless least-squares estimator

$$\widehat{\beta}_T := X^\dagger y, \tag{17}$$

where $X^\dagger$ denotes the Moore–Penrose pseudoinverse of X.

For Gaussian covariates, the design matrix has maximal rank almost surely. Therefore, when $n \geq p$,

$$\widehat{\beta}_T = (X^\top X)^{-1} X^\top y, \tag{18}$$

which is the ordinary least-squares estimator. When $p > n$,

$$\widehat{\beta}_T = X^\top (X X^\top)^{-1} y, \tag{19}$$

which is the unique minimum-Euclidean-norm solution of $X\beta = y$. Thus, in the overparameterized regime $p > n$, the ridgeless least-squares estimator coincides with the minimum-norm interpolator.

Define the empirical covariance matrix $\widehat{\Sigma} := \frac{1}{n} X^\top X \in \mathbb{R}^{p \times p}$. For fixed p, the law of large numbers gives $\widehat{\Sigma} \to \Sigma$ as $n \to \infty$ under the usual moment assumptions. In the proportional high-dimensional regime considered below, however, $\widehat{\Sigma}$ does not in general converge to Σ in operator norm, and its effect must instead be described through deterministic equivalents.

An equivalent covariance-form representation of (17) is

$$\widehat{\beta}_T = \widehat{\Sigma}^\dagger \left(\frac{1}{n} X^\top y \right), \tag{20}$$

where $\widehat{\Sigma}^\dagger$ denotes the Moore–Penrose pseudoinverse. A substantial body of recent work has developed both asymptotic and non-asymptotic risk characterizations for the minimum ℓ_2-norm interpolator, corresponding to ridgeless regression in the overparameterized regime $p \geq n$ (Hastie et al., 2022; Cheng and Montanari, 2024; Han and Xu, 2026; Wu and Xu, 2020; Richards et al., 2021; Loureiro et al., 2021). To analyze the Stage-I population risk in the proportional high-dimensional regime, we use the strong deterministic-equivalence framework introduced in Lemma C.1, following Atanasov et al. (2026). In particular, the random resolvents involving $\widehat{\Sigma}$ that appear in the risk calculation are replaced by their deterministic equivalents in the bounded-test-matrix sense. Since $\mathcal{R}(h_{\widehat{\beta}_T})$ already averages over an independent test covariate, we further average over the randomness of the training sample and define

$$\overline{\mathcal{R}}_T := \mathbb{E}_{X,\varepsilon}\left[\mathcal{R}(h_{\widehat{\beta}_T}) \right].$$

This quantity corresponds to the average generalization error considered in (Atanasov et al., 2026, Sec. III.A–B). Applying the deterministic-equivalent resolvent identities of Lemma C.1 yields

$$\overline{\mathcal{R}}_T \overset{\mathrm{DE}}{\simeq} \overline{\mathcal{R}}_T^{\mathrm{DE}},$$

where $\overset{\mathrm{DE}}{\simeq}$ at the scalar level denotes asymptotic equality in the proportional high-dimensional limit, following the notation used in (Atanasov et al., 2026, Eq. (24)). The explicit expression for $\overline{\mathcal{R}}_T^{\mathrm{DE}}$ derived below is the ridgeless specialization of the ridge-regression generalization-error formula in (Atanasov et al., 2026, Sec. III.B, Eqs. (24)–(25)).

Theorem C.2 (Stage-I risk deterministic equivalent)**.** *The averaged Stage-I population risk admits the deterministic equivalent*

$$\overline{\mathcal{R}}_T^{\mathrm{DE}} = \frac{1}{1 - \gamma_T} \left[\tau_T^2 \, \beta^{\star\top} \Sigma (\Sigma + \tau_T I_p)^{-2} \beta^\star + \sigma^2 \, \gamma_T \right], \tag{21}$$

where

$$\gamma_T := \frac{p}{n} \frac{1}{p} \sum_{i=1}^{p} \frac{\lambda_i^2}{(\lambda_i + \tau_T)^2} = \frac{1}{n} \sum_{i=1}^{p} \frac{\lambda_i^2}{(\lambda_i + \tau_T)^2},$$

and, in the ridgeless overparameterized regime, $\tau_T > 0$ is determined by

$$\sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \tau_T} = n.$$

Proof. This follows directly from the ridgeless specialization of the generalization-error deterministic equivalent in (Atanasov et al., 2026, Sec. III.B, Eqs. (24)–(25)). □

C.4 Stage II Risk Derivation (Distilled Stage-II Sample)

Stage-II data model. Draw a fresh independent dataset

$$\{\widetilde{x}_i\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\Sigma), \qquad \widetilde{X} \in \mathbb{R}^{m\times p},$$

and define the *distilled* labels, with no additional label noise, by $\widetilde{y} := \widetilde{X}\widehat{\beta}_T$.

Stage-II ridgeless estimator. In Stage II, we fit the ridgeless least-squares estimator

$$\widehat{\beta}_S = \begin{cases} \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{m}\|\widetilde{X}\beta - \widetilde{y}\|_2^2, & m \ge p, \\ \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{\|\beta\|_2 : \widetilde{X}\beta = \widetilde{y}\right\}, & m < p. \end{cases}$$

Using $\widetilde{y} = \widetilde{X}\widehat{\beta}_T$, the student estimator can be written as

$$\widehat{\beta}_S = \begin{cases} \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{m}\left\|\widetilde{X}\beta - \widetilde{X}\widehat{\beta}_T\right\|_2^2 = (\widetilde{X}^\top\widetilde{X})^{-1}\widetilde{X}^\top\widetilde{X}\,\widehat{\beta}_T = \widehat{\beta}_T, & m \ge p, \\ \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{\|\beta\|_2 : \widetilde{X}\beta = \widetilde{X}\widehat{\beta}_T\right\} = \widetilde{X}^\top(\widetilde{X}\widetilde{X}^\top)^{-1}\widetilde{X}\,\widehat{\beta}_T, & m < p. \end{cases}$$

When $\widetilde{\Sigma}$ is singular (e.g., $m < p$), $\widehat{\beta}_S$ is understood as the minimum-norm interpolating solution. Since $\widehat{\beta}_S$ depends on both the Stage-I training sample (X,ε) and the fresh Stage-II design $\widetilde{X}$, we define the averaged Stage-II population risk by

$$\overline{\mathcal{R}}_S := \mathbb{E}_{X,\varepsilon,\widetilde{X}}\left[\mathcal{R}(h_{\widehat{\beta}_S})\right].$$

To analyze the Stage-II population risk in the proportional high-dimensional regime, we again use the strong deterministic-equivalence framework introduced in Lemma C.1, following Atanasov et al. (2026). In particular, conditional on $\widehat{\beta}_T$, the random resolvents involving $\widetilde{\Sigma}$ are replaced by their deterministic equivalents in the bounded-test-matrix sense. We therefore write

$$\overline{\mathcal{R}}_S \overset{\text{DE}}{\simeq} \overline{\mathcal{R}}_S^{\text{DE}},$$

where $\overline{\mathcal{R}}_S^{\text{DE}}$ denotes the deterministic equivalent of the averaged Stage-II population risk.

We first derive the Stage-II risk while leaving the remaining expectation over the Stage-I estimator $\widehat{\beta}_T$ unevaluated. The resulting expression depends on moments of $\widehat{\beta}_T$. In the subsequent derivation, these moments are rewritten as weighted matrix functionals that are compatible with the deterministic-equivalence framework.

Theorem C.3 (Stage-II risk deterministic equivalent)**.** *The Stage-II deterministic-equivalent risk is*

$$\overline{\mathcal{R}}_S^{\text{DE}} = \mathbb{E}_{\widehat{\beta}_T}\Bigg[\underbrace{\|\widehat{\beta}_T - \beta^\star\|_\Sigma^2}_{\text{Stage-I error carried over}} + \underbrace{\frac{\tau_S^2}{1-\gamma_S}\sum_{i=1}^p \frac{\lambda_i(\widehat{\beta}_{T,i})^2}{(\lambda_i+\tau_S)^2}}_{\text{Stage-II distillation error}} - \underbrace{2\sum_{i=1}^p \lambda_i \frac{\tau_S}{\lambda_i+\tau_S}\widehat{\beta}_{T,i}(\widehat{\beta}_{T,i} - \beta_i^\star)}_{\text{cross term}}\Bigg], \tag{22}$$

where $\{\lambda_i\}_{i=1}^p$ are the eigenvalues of Σ, $\widehat{\beta}_{T,i}$ and $\beta_i^\star$ denote the corresponding coordinates in the eigenbasis of Σ, and

$$\gamma_S := \frac{p}{m}\cdot\frac{1}{p}\sum_{i=1}^p \frac{\lambda_i^2}{(\lambda_i+\tau_S)^2} = \frac{1}{m}\sum_{i=1}^p \frac{\lambda_i^2}{(\lambda_i+\tau_S)^2}.$$

In the ridgeless overparameterized regime, $\tau_S > 0$ is determined by

$$\sum_{i=1}^p \frac{\lambda_i}{\lambda_i+\tau_S} = m.$$

Consequently, define the averaged Stage-I-to-Stage-II risk improvement by

$$\mathsf{W2SG}_p := \overline{\mathcal{R}}_T - \overline{\mathcal{R}}_S \overset{\mathrm{DE}}{\simeq} \mathsf{W2SG}_p^{\mathrm{DE}},$$

where

$$\mathsf{W2SG}_p^{\mathrm{DE}} := \overline{\mathcal{R}}_T^{\mathrm{DE}} - \overline{\mathcal{R}}_S^{\mathrm{DE}}.$$

At the deterministic-equivalent level, the Stage-I error carried over cancels, yielding

$$\mathsf{W2SG}_p^{\mathrm{DE}} = \mathbb{E}_{\widehat{\beta}_T}\left[2\sum_{i=1}^{p}\lambda_i\frac{\tau_S}{\lambda_i+\tau_S}\widehat{\beta}_{T,i}\big(\widehat{\beta}_{T,i}-\beta_i^\star\big) - \frac{\tau_S^2}{1-\gamma_S}\sum_{i=1}^{p}\frac{\lambda_i(\widehat{\beta}_{T,i})^2}{(\lambda_i+\tau_S)^2}\right]. \tag{23}$$

Proof. Define the Stage-II empirical covariance matrix by

$$\widetilde{\Sigma} := \frac{1}{m}\widetilde{X}^\top\widetilde{X} \in \mathbb{R}^{p\times p}.$$

For the deterministic-equivalent calculation, we first introduce the auxiliary ridge estimator

$$\widehat{\beta}_{S,\lambda_S} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{m}\|\widetilde{X}\beta-\widetilde{y}\|_2^2 + \lambda_S\|\beta\|_2^2\right\}.$$

Using $\widetilde{y} = \widetilde{X}\widehat{\beta}_T$, this is equivalently

$$\widehat{\beta}_{S,\lambda_S} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{m}\|\widetilde{X}(\beta-\widehat{\beta}_T)\|_2^2 + \lambda_S\|\beta\|_2^2\right\}.$$

Although our Stage-II estimator is ridgeless, we keep $\lambda_S > 0$ throughout the deterministic-equivalent calculation, following Atanasov et al. (2026), and take the limit $\lambda_S \downarrow 0$ only at the end. For $\lambda_S > 0$,

$$\widehat{\beta}_{S,\lambda_S} = (\widetilde{X}^\top\widetilde{X} + m\lambda_S I_p)^{-1}\widetilde{X}^\top\widetilde{y} = (\widetilde{X}^\top\widetilde{X} + m\lambda_S I_p)^{-1}\widetilde{X}^\top\widetilde{X}\,\widehat{\beta}_T = (\widetilde{\Sigma}+\lambda_S I_p)^{-1}\widetilde{\Sigma}\,\widehat{\beta}_T. \tag{24}$$

Equivalently,

$$\widehat{\beta}_{S,\lambda_S} = \Big(I_p - \lambda_S(\widetilde{\Sigma}+\lambda_S I_p)^{-1}\Big)\widehat{\beta}_T.$$

Hence, defining the Stage-II update by $\Delta_{S,\lambda_S} := \widehat{\beta}_{S,\lambda_S} - \widehat{\beta}_T$, we obtain $\Delta_{S,\lambda_S} = -\lambda_S(\widetilde{\Sigma} + \lambda_S I_p)^{-1}\widehat{\beta}_T$.

Two-stage population risk decomposition. Write

$$\widehat{\beta}_{S,\lambda_S} - \beta^\star = \Delta_{S,\lambda_S} + (\widehat{\beta}_T - \beta^\star).$$

Expanding the population risk and conditioning first on $\widehat{\beta}_T$ gives

$$\overline{\mathcal{R}}_{S,\lambda_S} = \mathbb{E}_{\widehat{\beta}_T}\Big[\|\widehat{\beta}_T-\beta^\star\|_\Sigma^2 + \mathbb{E}_{\widetilde{X}}\big[\|\Delta_{S,\lambda_S}\|_\Sigma^2 \mid \widehat{\beta}_T\big] + 2\,\mathbb{E}_{\widetilde{X}}\big[(\widehat{\beta}_T-\beta^\star)^\top\Sigma\Delta_{S,\lambda_S} \mid \widehat{\beta}_T\big]\Big]. \tag{25}$$

Deterministic equivalence for Stage II conditional on $\widehat{\beta}_T$. Condition on $\widehat{\beta}_T$ and view Stage II as ridge regression with fixed teacher $\widehat{\beta}_T$ and zero label noise. In the proportional limit, applying the deterministic-equivalence framework of Lemma C.1 to the Stage-II resolvent terms, in the same manner as in the Stage-I risk derivation, yields

$$\mathbb{E}_{\widetilde{X}}\Big[(\widehat{\beta}_{S,\lambda_S}-\widehat{\beta}_T)^\top\Sigma(\widehat{\beta}_{S,\lambda_S}-\widehat{\beta}_T)\,\Big|\,\widehat{\beta}_T\Big] \overset{\mathrm{DE}}{\simeq} \frac{\tau_S^2}{1-\gamma_S}\sum_{i=1}^{p}\frac{\lambda_i(\widehat{\beta}_{T,i})^2}{(\lambda_i+\tau_S)^2},$$

where

$$\gamma_S = \frac{p}{m}\cdot\frac{1}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{(\lambda_i+\tau_S)^2}.$$

The renormalized ridge parameter τ_S is related to the explicit ridge parameter λ_S through the corresponding fixed-point equation

$$\tau_S \left(1 - \frac{1}{m}\sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \tau_S}\right) = \lambda_S.$$

Thus, in the ridgeless overparameterized limit $\lambda_S \downarrow 0$ with $\tau_S > 0$,

$$\sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \tau_S} = m.$$

For the cross term, Lemma C.1 gives the strong deterministic equivalence

$$\widetilde{\Sigma}(\widetilde{\Sigma} + \lambda_S I_p)^{-1} \overset{\mathrm{DE}}{\simeq} \Sigma(\Sigma + \tau_S I_p)^{-1}.$$

Hence, conditional on $\widehat{\beta}_T$, the Stage-II estimator is a ridge-regression estimator with fixed teacher $\widehat{\beta}_T$. Applying the corresponding deterministic equivalent from Atanasov et al. (2026) gives

$$\mathbb{E}_{\widetilde{X}}\left[\widehat{\beta}_{S,\lambda_S} \mid \widehat{\beta}_T\right] \overset{\mathrm{DE}}{\simeq} \Sigma(\Sigma + \tau_S I_p)^{-1}\widehat{\beta}_T.$$

Consequently,

$$\mathbb{E}_{\widetilde{X}}\left[\Delta_{S,\lambda_S} \mid \widehat{\beta}_T\right] \overset{\mathrm{DE}}{\simeq} -\tau_S(\Sigma + \tau_S I_p)^{-1}\widehat{\beta}_T.$$

Therefore,

$$\mathbb{E}_{\widetilde{X}}\left[(\widehat{\beta}_T - \beta^\star)^\top \Sigma \Delta_{S,\lambda_S} \mid \widehat{\beta}_T\right] \overset{\mathrm{DE}}{\simeq} -\tau_S(\widehat{\beta}_T - \beta^\star)^\top \Sigma(\Sigma + \tau_S I_p)^{-1}\widehat{\beta}_T = -\sum_{i=1}^{p} \lambda_i \frac{\tau_S}{\lambda_i + \tau_S}\widehat{\beta}_{T,i}(\widehat{\beta}_{T,i} - \beta_i^\star). \tag{26}$$

Final Stage-II risk. Substituting the deterministic equivalents above into (25) and then taking the ridgeless limit $\lambda_S \downarrow 0$ yields

$$\overline{\mathcal{R}}_S \overset{\mathrm{DE}}{\simeq} \mathbb{E}_{\widehat{\beta}_T}\Bigg[\underbrace{\|\widehat{\beta}_T - \beta^\star\|_\Sigma^2}_{\text{Stage-I error carried over}} + \underbrace{\frac{\tau_S^2}{1-\gamma_S}\sum_{i=1}^{p}\frac{\lambda_i(\widehat{\beta}_{T,i})^2}{(\lambda_i+\tau_S)^2}}_{\text{Stage-II distillation error}} - \underbrace{2\sum_{i=1}^{p}\lambda_i\frac{\tau_S}{\lambda_i+\tau_S}\widehat{\beta}_{T,i}(\widehat{\beta}_{T,i} - \beta_i^\star)}_{\text{cross term}}\Bigg]. \tag{27}$$

This is precisely the deterministic equivalent $\overline{\mathcal{R}}_S^{\mathrm{DE}}$ stated in (22). □

Lemma C.4 (Weighted ridgeless pseudoinverse deterministic equivalent)**.** *Let $B = B_p$ be a deterministic matrix with uniformly bounded operator norm.*

Let $\tau_T > 0$ be determined by

$$\frac{1}{n}\operatorname{Tr}\left[\Sigma(\Sigma + \tau_T I_p)^{-1}\right] = 1, \tag{28}$$

and define

$$\gamma_T := \frac{1}{n}\operatorname{Tr}\left[\Sigma^2(\Sigma + \tau_T I_p)^{-2}\right]. \tag{29}$$

Then

$$\mathbb{E}_X \operatorname{Tr}\left[B\widehat{\Sigma}^\dagger\right] \overset{\mathrm{DE}}{\simeq} \frac{1}{1-\gamma_T}\operatorname{Tr}\left[B\Sigma(\Sigma + \tau_T I_p)^{-2}\right]. \tag{30}$$

Proof. We start from Lemma C.1. For $\lambda > 0$, that lemma gives

$$(\widehat{\Sigma} + \lambda I_p)^{-1} \overset{\mathrm{DE}}{\simeq} \frac{\tau(\lambda)}{\lambda}(\Sigma + \tau(\lambda) I_p)^{-1}, \tag{31}$$

where the renormalized ridge $\tau(\lambda)$ is determined by

$$\tau\left\{1-\frac{1}{n}\operatorname{Tr}\left[\Sigma(\Sigma+\tau I_p)^{-1}\right]\right\}=\lambda. \tag{32}$$

Weighted pseudoinverse deterministic equivalent. Starting from the finite-ridge relation (DE2), differentiate with respect to λ. Since

$$-\frac{d}{d\lambda}\left[\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^{-1}\right]=\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^{-2},$$

and, by the chain rule,

$$-\frac{d}{d\lambda}\left[\Sigma(\Sigma+\tau(\lambda)I_p)^{-1}\right]=\frac{d\tau}{d\lambda}\Sigma(\Sigma+\tau(\lambda)I_p)^{-2},$$

testing against B gives

$$\mathbb{E}_X\operatorname{Tr}\left[B\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^{-2}\right]\overset{\mathrm{DE}}{\simeq}\frac{d\tau}{d\lambda}\operatorname{Tr}\left[B\Sigma(\Sigma+\tau(\lambda)I_p)^{-2}\right]. \tag{33}$$

It remains to evaluate $d\tau/d\lambda$. From (32),

$$\lambda=\tau-\frac{\tau}{n}\operatorname{Tr}\left[\Sigma(\Sigma+\tau I_p)^{-1}\right].$$

Differentiating with respect to τ gives

$$\frac{d\lambda}{d\tau}=1-\frac{1}{n}\operatorname{Tr}\left[\Sigma(\Sigma+\tau I_p)^{-1}\right]+\frac{\tau}{n}\operatorname{Tr}\left[\Sigma(\Sigma+\tau I_p)^{-2}\right].$$

Using

$$\Sigma(\Sigma+\tau I_p)^{-1}-\tau\Sigma(\Sigma+\tau I_p)^{-2}=\Sigma^2(\Sigma+\tau I_p)^{-2},$$

this simplifies to

$$\frac{d\lambda}{d\tau}=1-\frac{1}{n}\operatorname{Tr}\left[\Sigma^2(\Sigma+\tau I_p)^{-2}\right]. \tag{34}$$

Hence, as $\lambda\downarrow 0$ and $\tau(\lambda)\to\tau_T$,

$$\left.\frac{d\tau}{d\lambda}\right|_{\lambda\downarrow 0}=\frac{1}{1-\gamma_T},$$

where

$$\gamma_T=\frac{1}{n}\operatorname{Tr}\left[\Sigma^2(\Sigma+\tau_T I_p)^{-2}\right].$$

Finally, by spectral calculus,

$$\lim_{\lambda\downarrow 0}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^{-2}=\widehat{\Sigma}^{\dagger}.$$

Taking $\lambda\downarrow 0$ in (33) therefore yields

$$\mathbb{E}_X\operatorname{Tr}\left[B\widehat{\Sigma}^{\dagger}\right]\overset{\mathrm{DE}}{\simeq}\frac{1}{1-\gamma_T}\operatorname{Tr}\left[B\Sigma(\Sigma+\tau_T I_p)^{-2}\right],$$

which proves the lemma. □

We next derive a fully deterministic expression for the Stage-I-to-Stage-II risk improvement. A related two-stage ridgeless-regression setting is studied in (Ildiz et al., 2025, Definition 3), where the Stage-I surrogate sample and Stage-II target sample may have different population covariances and noise levels. Their analysis therefore accommodates a more general distribution-shift setting and also provides a generalization bound under additional boundedness assumptions. Some of those assumptions, including boundedness conditions involving the inverse covariance, are not imposed in the present analysis. Accordingly, we work instead with deterministic-equivalent formulas under the spectral assumptions stated above and do not claim a corresponding non-asymptotic generalization bound.

Theorem C.5 (Fully deterministic Stage-I-to-Stage-II risk improvement)**.** *Assume, in addition, that the teacher is random and independent of the design and label noise and is isotropic:*

$$\mathbb{E}[\beta^\star] = 0, \qquad \mathbb{E}[\beta^\star \beta^{\star\top}] = I_p.$$

Then the teacher-averaged Stage-I-to-Stage-II risk improvement admits the deterministic equivalent

$$\mathbb{E}_{\beta^\star}[\mathsf{W2SG}_p] \overset{\mathrm{DE}}{\simeq} \mathsf{W2SG}_p^{\mathrm{DE}},$$

where

$$\begin{aligned}\mathsf{W2SG}_p^{\mathrm{DE}} = & -\frac{\tau_S^2}{1-\gamma_S}\operatorname{Tr}\left[\Sigma^2(\Sigma+\tau_S I_p)^{-2}(\Sigma+\tau_T I_p)^{-1}\right] \\ & +\frac{2\sigma^2\tau_S}{n(1-\gamma_T)}\operatorname{Tr}\left[\Sigma^2(\Sigma+\tau_S I_p)^{-1}(\Sigma+\tau_T I_p)^{-2}\right] \\ & -\frac{\sigma^2\tau_S^2}{n(1-\gamma_T)(1-\gamma_S)}\operatorname{Tr}\left[\Sigma^2(\Sigma+\tau_S I_p)^{-2}(\Sigma+\tau_T I_p)^{-2}\right]. \end{aligned} \tag{35}$$

Since all matrices in (35) are functions of Σ, they are simultaneously diagonalizable. Thus, the same deterministic equivalent can be written in the eigenbasis of Σ as

$$\begin{aligned}\mathsf{W2SG}_p^{\mathrm{DE}} = & -\frac{\tau_S^2}{1-\gamma_S}\sum_{i=1}^p \frac{\lambda_i^2}{(\lambda_i+\tau_S)^2(\lambda_i+\tau_T)} \\ & +\frac{2\sigma^2\tau_S}{n(1-\gamma_T)}\sum_{i=1}^p \frac{\lambda_i^2}{(\lambda_i+\tau_S)(\lambda_i+\tau_T)^2} \\ & -\frac{\sigma^2\tau_S^2}{n(1-\gamma_T)(1-\gamma_S)}\sum_{i=1}^p \frac{\lambda_i^2}{(\lambda_i+\tau_S)^2(\lambda_i+\tau_T)^2}. \end{aligned} \tag{36}$$

Proof. To complete the derivation of the Stage-I-to-Stage-II risk improvement, we must average the Stage-II deterministic-equivalent expression over the remaining Stage-I randomness. Direct coordinate-wise application of deterministic equivalence to the random coordinates $\widehat{\beta}_{T,i}$ is not justified by Lemma C.1, which is formulated in terms of deterministic test matrices and trace functionals. We therefore rewrite the coordinate expression in terms of weighted quadratic forms that are compatible with the deterministic-equivalence framework of Lemma C.1.

Since all matrices appearing in the Stage-II deterministic equivalent are functions of the population covariance Σ, define

$$A_S := \tau_S\Sigma(\Sigma+\tau_S I_p)^{-1}, \tag{37}$$

and

$$D_S := 2\tau_S\Sigma(\Sigma+\tau_S I_p)^{-1} - \frac{\tau_S^2}{1-\gamma_S}\Sigma(\Sigma+\tau_S I_p)^{-2}. \tag{38}$$

Indeed, in the eigenbasis of Σ,

$$[A_S]_{ii} = \lambda_i\frac{\tau_S}{\lambda_i+\tau_S},$$

whereas

$$[D_S]_{ii} = 2\lambda_i\frac{\tau_S}{\lambda_i+\tau_S} - \frac{\tau_S^2}{1-\gamma_S}\frac{\lambda_i}{(\lambda_i+\tau_S)^2}.$$

Therefore, the Stage-II expression (23) can be written as

$$\mathbb{E}_{\beta^\star}[\mathsf{W2SG}_p] \overset{\mathrm{DE}}{\simeq} \mathbb{E}\left[\widehat{\beta}_T^\top D_S\widehat{\beta}_T\right] - 2\mathbb{E}\left[\beta^{\star\top}A_S\widehat{\beta}_T\right]. \tag{39}$$

Recall that $\widehat{\Sigma} = \frac{1}{n}X^\top X$, $P_T = \widehat{\Sigma}^\dagger\widehat{\Sigma}$ and the Stage-I labels satisfy

$$y = X\beta^\star + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I_n. \tag{40}$$

The Stage-I ridgeless estimator can be written as

$$\widehat{\beta}_T = \widehat{\Sigma}^\dagger \frac{X^\top y}{n} = \widehat{\Sigma}^\dagger \widehat{\Sigma} \beta^\star + \widehat{\Sigma}^\dagger \frac{X^\top \varepsilon}{n}. \tag{41}$$

Since $\widehat{\Sigma}$ is symmetric positive semidefinite, P_T is an orthogonal projector and hence $P_T^\top = P_T, P_T^2 = P_T$. Thus,

$$\widehat{\beta}_T = P_T \beta^\star + \widehat{\Sigma}^\dagger \frac{X^\top \varepsilon}{n}. \tag{42}$$

We first evaluate $\mathbb{E}\left[\widehat{\beta}_T^\top D_S \widehat{\beta}_T\right]$. Conditioning on X and using (42), we have

$$\widehat{\beta}_T \widehat{\beta}_T^\top = P_T \beta^\star \beta^{\star\top} P_T + P_T \beta^\star \varepsilon^\top \frac{X}{n} \widehat{\Sigma}^\dagger + \widehat{\Sigma}^\dagger \frac{X^\top \varepsilon}{n} \beta^{\star\top} P_T + \widehat{\Sigma}^\dagger \frac{X^\top \varepsilon \varepsilon^\top X}{n^2} \widehat{\Sigma}^\dagger. \tag{43}$$

Because $\beta^\star$ is independent of the design and noise, $\mathbb{E}[\beta^\star \beta^{\star\top}] = I_p$, and $\mathbb{E}[\varepsilon] = 0$, the two mixed terms vanish. Therefore,

$$\mathbb{E}_{\beta^\star,\varepsilon}\left[\widehat{\beta}_T \widehat{\beta}_T^\top \mid X\right] = P_T + \frac{\sigma^2}{n^2} \widehat{\Sigma}^\dagger X^\top X \widehat{\Sigma}^\dagger = P_T + \frac{\sigma^2}{n} \widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger = P_T + \frac{\sigma^2}{n} \widehat{\Sigma}^\dagger, \tag{44}$$

where the last equality uses the Moore–Penrose identity $\widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger = \widehat{\Sigma}^\dagger$.

Consequently,

$$\mathbb{E}\left[\widehat{\beta}_T^\top D_S \widehat{\beta}_T\right] = \mathbb{E}_X \operatorname{Tr}\left[D_S \mathbb{E}_{\beta^\star,\varepsilon}\left[\widehat{\beta}_T \widehat{\beta}_T^\top \mid X\right]\right] = \mathbb{E}_X \operatorname{Tr}(D_S P_T) + \frac{\sigma^2}{n} \mathbb{E}_X \operatorname{Tr}\left[D_S \widehat{\Sigma}^\dagger\right]. \tag{45}$$

From (42),

$$\mathbb{E}\left[\beta^{\star\top} A_S \widehat{\beta}_T\right] = \mathbb{E}\left[\beta^{\star\top} A_S P_T \beta^\star\right] + \mathbb{E}\left[\beta^{\star\top} A_S \widehat{\Sigma}^\dagger \frac{X^\top \varepsilon}{n}\right]. \tag{46}$$

The second term vanishes because the label noise is independent and zero mean. Hence,

$$\mathbb{E}\left[\beta^{\star\top} A_S \widehat{\beta}_T\right] = \mathbb{E}_X \mathbb{E}_{\beta^\star}\left[\beta^{\star\top} A_S P_T \beta^\star \mid X\right] = \mathbb{E}_X \operatorname{Tr}\left[A_S P_T \mathbb{E}[\beta^\star \beta^{\star\top}]\right] = \mathbb{E}_X \operatorname{Tr}(A_S P_T). \tag{47}$$

Substituting (45) and (47) into (39) gives

$$\mathbb{E}_{\beta^\star}[\mathsf{W2SG}_p] \overset{\mathrm{DE}}{\simeq} \mathbb{E}_X \operatorname{Tr}(D_S P_T) + \frac{\sigma^2}{n} \mathbb{E}_X \operatorname{Tr}\left[D_S \widehat{\Sigma}^\dagger\right] - 2\mathbb{E}_X \operatorname{Tr}(A_S P_T). \tag{48}$$

We now evaluate the three remaining Stage-I expectations in (48): $\mathbb{E}_X \operatorname{Tr}(D_S P_T), \mathbb{E}_X \operatorname{Tr}(D_S \widehat{\Sigma}^\dagger), \mathbb{E}_X \operatorname{Tr}(A_S P_T)$. At this stage, A_S and D_S are deterministic matrices because they depend only on Σ and on the Stage-II deterministic-equivalent parameters τ_S and γ_S. Thus, all remaining randomness is contained in the Stage-I design X through $\widehat{\Sigma} = \frac{1}{n} X^\top X, P_T = \widehat{\Sigma}^\dagger \widehat{\Sigma}$.

Recall that $\tau_T > 0$ is determined in the ridgeless overparameterized regime by

$$\frac{1}{n} \operatorname{Tr}\left[\Sigma(\Sigma + \tau_T I_p)^{-1}\right] = 1, \tag{49}$$

or equivalently $\sum_{i=1}^p \frac{\lambda_i}{\lambda_i + \tau_T} = n$. Also recall

$$\gamma_T := \frac{1}{n} \operatorname{Tr}\left[\Sigma^2(\Sigma + \tau_T I_p)^{-2}\right] = \frac{1}{n} \sum_{i=1}^p \left(\frac{\lambda_i}{\lambda_i + \tau_T}\right)^2. \tag{50}$$

For a fixed proportional Stage-II aspect ratio $\phi_S > 1$, $1 - \gamma_S$ remains bounded away from zero. Moreover, $0 \preceq \tau_S \Sigma(\Sigma + \tau_S I_p)^{-1} \preceq \Sigma$, and $0 \preceq \tau_S^2 \Sigma(\Sigma + \tau_S I_p)^{-2} \preceq \Sigma$. Hence, under the assumed uniform bound on $\|\Sigma\|_{\mathrm{op}}$, both A_S and D_S have uniformly bounded operator norm.

Next, observe that

$$P_T = \widehat{\Sigma}^\dagger \widehat{\Sigma} = \lim_{\lambda \downarrow 0} \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{-1}.$$

Taking the ridgeless limit in the shrinkage deterministic equivalent (DE2) therefore gives, for every deterministic matrix B with uniformly bounded operator norm,

$$\mathbb{E}_X \operatorname{Tr}(BP_T) \overset{\text{DE}}{\simeq} \operatorname{Tr}\left[B\Sigma(\Sigma + \tau_T I_p)^{-1}\right]. \tag{51}$$

Applying (51) with $B = D_S$ gives

$$\mathbb{E}_X \operatorname{Tr}(D_S P_T) \overset{\text{DE}}{\simeq} \operatorname{Tr}\left[D_S\Sigma(\Sigma + \tau_T I_p)^{-1}\right]. \tag{52}$$

Similarly, taking $B = A_S$ gives

$$\mathbb{E}_X \operatorname{Tr}(A_S P_T) \overset{\text{DE}}{\simeq} \operatorname{Tr}\left[A_S\Sigma(\Sigma + \tau_T I_p)^{-1}\right]. \tag{53}$$

Finally, applying Lemma C.4 with $B = D_S$ gives

$$\mathbb{E}_X \operatorname{Tr}\left[D_S\widehat{\Sigma}^\dagger\right] \overset{\text{DE}}{\simeq} \frac{1}{1-\gamma_T} \operatorname{Tr}\left[D_S\Sigma(\Sigma + \tau_T I_p)^{-2}\right]. \tag{54}$$

Therefore,

$$\frac{\sigma^2}{n}\mathbb{E}_X \operatorname{Tr}\left[D_S\widehat{\Sigma}^\dagger\right] \overset{\text{DE}}{\simeq} \frac{\sigma^2}{n(1-\gamma_T)} \operatorname{Tr}\left[D_S\Sigma(\Sigma + \tau_T I_p)^{-2}\right]. \tag{55}$$

Substituting (52), (55), and (53) into (48) yields

$$\mathbb{E}_{\beta^\star}[\mathsf{W2SG}_p] \overset{\text{DE}}{\simeq} \operatorname{Tr}\left[D_S\Sigma(\Sigma + \tau_T I_p)^{-1}\right] + \frac{\sigma^2}{n(1-\gamma_T)} \operatorname{Tr}\left[D_S\Sigma(\Sigma + \tau_T I_p)^{-2}\right] - 2\operatorname{Tr}\left[A_S\Sigma(\Sigma + \tau_T I_p)^{-1}\right]. \tag{56}$$

Recall that

$$A_S = \tau_S\Sigma(\Sigma + \tau_S I_p)^{-1},$$

and

$$D_S = 2\tau_S\Sigma(\Sigma + \tau_S I_p)^{-1} - \frac{\tau_S^2}{1-\gamma_S}\Sigma(\Sigma + \tau_S I_p)^{-2}.$$

Therefore,

$$D_S - 2A_S = -\frac{\tau_S^2}{1-\gamma_S}\Sigma(\Sigma + \tau_S I_p)^{-2}. \tag{57}$$

The first and third terms of (56) can therefore be combined as

$$\begin{aligned}&\operatorname{Tr}\left[D_S\Sigma(\Sigma + \tau_T I_p)^{-1}\right] - 2\operatorname{Tr}\left[A_S\Sigma(\Sigma + \tau_T I_p)^{-1}\right] = \operatorname{Tr}\left[(D_S - 2A_S)\Sigma(\Sigma + \tau_T I_p)^{-1}\right] \\ &= -\frac{\tau_S^2}{1-\gamma_S} \operatorname{Tr}\left[\Sigma^2(\Sigma + \tau_S I_p)^{-2}(\Sigma + \tau_T I_p)^{-1}\right].\end{aligned} \tag{58}$$

For the noise contribution, substituting the explicit form of D_S into (55) gives

$$\begin{aligned}&\frac{\sigma^2}{n(1-\gamma_T)} \operatorname{Tr}\left[D_S\Sigma(\Sigma + \tau_T I_p)^{-2}\right] \\ =&\frac{2\sigma^2\tau_S}{n(1-\gamma_T)} \operatorname{Tr}\left[\Sigma^2(\Sigma + \tau_S I_p)^{-1}(\Sigma + \tau_T I_p)^{-2}\right] - \frac{\sigma^2\tau_S^2}{n(1-\gamma_T)(1-\gamma_S)} \operatorname{Tr}\left[\Sigma^2(\Sigma + \tau_S I_p)^{-2}(\Sigma + \tau_T I_p)^{-2}\right].\end{aligned} \tag{59}$$

Combining (58) and (59) gives

$$\begin{aligned}\mathbb{E}_{\beta^\star}[\mathsf{W2SG}_p] \overset{\text{DE}}{\simeq}& -\frac{\tau_S^2}{1-\gamma_S} \operatorname{Tr}\left[\Sigma^2(\Sigma + \tau_S I_p)^{-2}(\Sigma + \tau_T I_p)^{-1}\right] + \frac{2\sigma^2\tau_S}{n(1-\gamma_T)} \operatorname{Tr}\left[\Sigma^2(\Sigma + \tau_S I_p)^{-1}(\Sigma + \tau_T I_p)^{-2}\right] \\ &-\frac{\sigma^2\tau_S^2}{n(1-\gamma_T)(1-\gamma_S)} \operatorname{Tr}\left[\Sigma^2(\Sigma + \tau_S I_p)^{-2}(\Sigma + \tau_T I_p)^{-2}\right].\end{aligned} \tag{60}$$

The right-hand side is precisely $\mathsf{W2SG}_p^{\text{DE}}$ in (35).

Finally, because every matrix in this expression is a function of Σ, diagonalizing Σ gives (36). □

Lemma C.6 (Average spectral preference)**.** *Assume $1 \le m < p$. Let*

$$\Sigma = V \operatorname{diag}(\lambda_1, \dots, \lambda_p) V^\top, \qquad \lambda_1 \ge \cdots \ge \lambda_p > 0,$$

where $V = [v_1, \dots, v_p]$ is orthogonal. Define

$$P_m := \widetilde{X}^\dagger \widetilde{X}, \qquad \widehat{\beta}_S := P_m \widehat{\beta}_T.$$

Then the following statements hold: The expected projector satisfies

$$\mathbb{E}_{\widetilde{X}}[P_m] = V \operatorname{diag}(\ell_1, \dots, \ell_p) V^\top,$$

where

$$\ell_i := \mathbb{E}_{\widetilde{X}}[v_i^\top P_m v_i] = \mathbb{E}_{\widetilde{X}} \|P_m v_i\|_2^2$$

obey

$$1 > \ell_1 \ge \cdots \ge \ell_p > 0, \qquad \sum_{i=1}^{p} \ell_i = m.$$

Thus higher-variance covariance eigendirections have greater expected retention, but are not subject to an exact coordinate cutoff.

Proof. Write $\widetilde{X} V = G \Lambda^{1/2}$, where $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ and $G = [g_1, \dots, g_p]$ has independent $\mathcal{N}(0, I_m)$ columns. In the covariance eigenbasis,

$$V^\top P_m V = \Lambda^{1/2} G^\top (G \Lambda G^\top)^{-1} G \Lambda^{1/2}.$$

Sign symmetry of each Gaussian column implies that every off-diagonal entry of this matrix has zero expectation.

For a diagonal entry, define

$$A_{-i} := \sum_{j \ne i} \lambda_j g_j g_j^\top, \qquad a_i := g_i^\top A_{-i}^{-1} g_i.$$

Because $p - 1 \ge m$, A_{-i} is positive definite almost surely. The Sherman–Morrison identity gives

$$(V^\top P_m V)_{ii} = \frac{\lambda_i a_i}{1 + \lambda_i a_i}.$$

Hence $0 < \ell_i < 1$. For fixed G, this expression increases with λ_i and decreases with any λ_j, $j \ne i$. Comparing two configurations with λ_i and λ_j interchanged, and using exchangeability of the Gaussian columns, therefore gives $\lambda_i \ge \lambda_j \Rightarrow \ell_i \ge \ell_j$. Finally,

$$\sum_{i=1}^{p} \ell_i = \mathbb{E}[\operatorname{Tr}(P_m)] = m,$$

because P_m is a rank-m orthogonal projector. □

Interpretation. The pseudo-labeled sample size m determines the dimension of the teacher-parameter subspace retained by Stage-II training: an m-dimensional, sample-dependent subspace is preserved, while its $(p - m)$-dimensional orthogonal complement is eliminated. The retained subspace is the row space of $\widetilde{X}$, whereas the eliminated subspace is its null space. Importantly, these are sample-dependent directions and do not generally coincide with the covariance eigenvectors. In the covariance eigenbasis, the same mechanism appears as preferential retention of higher-variance directions in expectation, with total expected retention

$$\sum_{i=1}^{p} \ell_i = m.$$

Thus, m determines the dimension of the retained subspace, while the sampled covariates and the population covariance determine its orientation.

Assumption C.7 (Power-law covariance)**.** *We assume that the population covariance follows the power-law model*

$$\Sigma = \operatorname{diag}\left(1, 2^{-\alpha}, \ldots, p^{-\alpha}\right), \qquad \lambda_i = i^{-\alpha}, \qquad \alpha > 1, \tag{61}$$

where $\{\lambda_i\}_{i=1}^p$ *are the eigenvalues of* Σ*.*

Assumption C.8 (Proportional high-dimensional asymptotics)**.** *We work in the proportional high-dimensional regime*

$$\phi_{T,p} := \frac{p}{n} \longrightarrow \phi_T > 1, \qquad \phi_{S,p} := \frac{p}{m} \longrightarrow \phi_S > 1, \tag{62}$$

as $p, n, m \to \infty$*. Thus both Stage I and Stage II remain overparameterized, with asymptotically nonzero sample-to-dimension ratios. In this regime, the deterministic-equivalence framework of Lemma C.1 applies under the corresponding proportional-asymptotic assumptions.*

The preceding spectral-preference result provides intuition for how Stage II can modify the Stage-I estimator even in the absence of explicit regularization: the Stage-II row-space projection preferentially retains directions associated with larger population variance.

We now turn from this qualitative interpretation to a quantitative analysis. Under Assumptions C.7 and C.8, we specialize the deterministic-equivalent expression (56) to the power-law covariance model.

C.5 Proportional high-dimensional asymptotics

Theorem C.9 (Power-law limit of the deterministic-equivalent risk improvement)**.** *Under Assumptions C.7 and C.8, let* $\mathsf{W2SG}_p^{\mathrm{DE}}$ *denote the finite-*p *deterministic-equivalent Stage-I-to-Stage-II risk improvement. Then*

$$\mathsf{W2SG}_p^{\mathrm{DE}} \longrightarrow \frac{\sigma^2 \phi_T s_S}{1-\gamma_T} \int_0^1 \frac{x^\alpha \left[2 + \left(2 - \frac{1}{1-\gamma_S}\right) s_S x^\alpha\right]}{(1+s_S x^\alpha)^2 (1+s_T x^\alpha)^2}\, dx. \tag{63}$$

Here $s_T, s_S > 0$ *are the unique solutions of*

$$\int_0^1 \frac{dx}{1+s_a x^\alpha} = \frac{1}{\phi_a}, \qquad a \in \{T, S\}, \tag{64}$$

and

$$\gamma_a = \phi_a \int_0^1 \frac{dx}{(1+s_a x^\alpha)^2}, \qquad a \in \{T, S\}. \tag{65}$$

Proof. For clarity, throughout this proof we use $\tau_{a,p}, \gamma_{a,p}, a \in \{T, S\}$, to denote the corresponding finite-p quantities. Below, we define the rescaled ridge $s_{a,p} := p^\alpha \tau_{a,p}$, whose limit is denoted by s_a, while γ_a denotes the limit of $\gamma_{a,p}$. Since the ridgeless fixed-point equation is

$$\sum_{i=1}^p \frac{i^{-\alpha}}{i^{-\alpha} + \tau_{a,p}} = \frac{p}{\phi_{a,p}},$$

using $s_{a,p} = p^\alpha \tau_{a,p}$ and dividing by p gives

$$\frac{1}{p} \sum_{i=1}^p \frac{i^{-\alpha}}{i^{-\alpha} + \tau_{a,p}} = \frac{1}{p} \sum_{i=1}^p \frac{1}{1 + s_{a,p}(i/p)^\alpha} = \frac{1}{\phi_{a,p}}. \tag{66}$$

Hence $s_{a,p} \longrightarrow s_a,\ a \in \{T, S\}$, where $s_a > 0$ is the unique solution of

$$\int_0^1 \frac{dx}{1+s_a x^\alpha} = \frac{1}{\phi_a}. \tag{67}$$

Consequently,

$$\tau_{a,p} = s_a\, p^{-\alpha} + o(p^{-\alpha}). \tag{68}$$

Similarly,

$$\gamma_{a,p} = \frac{\phi_{a,p}}{p} \sum_{i=1}^{p} \left(\frac{i^{-\alpha}}{i^{-\alpha} + \tau_{a,p}} \right)^2 = \phi_{a,p} \frac{1}{p} \sum_{i=1}^{p} \frac{1}{(1 + s_{a,p}(i/p)^{\alpha})^2}. \tag{69}$$

Therefore,

$$\gamma_{a,p} \longrightarrow \gamma_a := \phi_a \int_0^1 \frac{dx}{(1 + s_a x^{\alpha})^2}. \tag{70}$$

Using (67),

$$1 - \gamma_a = \phi_a \int_0^1 \left[\frac{1}{1 + s_a x^{\alpha}} - \frac{1}{(1 + s_a x^{\alpha})^2} \right] dx = \phi_a s_a \int_0^1 \frac{x^{\alpha}}{(1 + s_a x^{\alpha})^2} \, dx > 0. \tag{71}$$

Hence $(1 - \gamma_{a,p})^{-1}$ remains bounded for all sufficiently large p. For the remainder of the proof, let $x_{i,p} := \frac{i}{p}$. Then $i^{-\alpha} = p^{-\alpha} x_{i,p}^{-\alpha}$, and $\frac{i^{-\alpha}}{i^{-\alpha} + \tau_{a,p}} = \frac{1}{1 + s_{a,p} x_{i,p}^{\alpha}}$. Consider

$$\operatorname{Tr}\left[D_S \Sigma (\Sigma + \tau_{T,p} I_p)^{-1} \right].$$

Since all matrices are diagonal in the eigenbasis of Σ,

$$\begin{aligned} \operatorname{Tr}\left[D_S \Sigma (\Sigma + \tau_{T,p} I_p)^{-1} \right] &= 2\tau_{S,p} \sum_{i=1}^{p} \frac{i^{-\alpha}}{i^{-\alpha} + \tau_{S,p}} \frac{i^{-\alpha}}{i^{-\alpha} + \tau_{T,p}} \\ &- \frac{\tau_{S,p}^2}{1 - \gamma_{S,p}} \sum_{i=1}^{p} \frac{i^{-\alpha}}{(i^{-\alpha} + \tau_{S,p})^2} \frac{i^{-\alpha}}{i^{-\alpha} + \tau_{T,p}}. \end{aligned} \tag{72}$$

The two summands satisfy

$$\tau_{S,p} \frac{i^{-\alpha}}{i^{-\alpha} + \tau_{S,p}} \frac{i^{-\alpha}}{i^{-\alpha} + \tau_{T,p}} = p^{-\alpha} \frac{s_{S,p}}{(1 + s_{S,p} x_{i,p}^{\alpha})(1 + s_{T,p} x_{i,p}^{\alpha})}, \tag{73}$$

and

$$\tau_{S,p}^2 \frac{i^{-\alpha}}{(i^{-\alpha} + \tau_{S,p})^2} \frac{i^{-\alpha}}{i^{-\alpha} + \tau_{T,p}} = p^{-\alpha} \frac{s_{S,p}^2 x_{i,p}^{\alpha}}{(1 + s_{S,p} x_{i,p}^{\alpha})^2 (1 + s_{T,p} x_{i,p}^{\alpha})}. \tag{74}$$

Therefore,

$$\begin{aligned} &\operatorname{Tr}\left[D_S \Sigma (\Sigma + \tau_{T,p} I_p)^{-1} \right] \\ &= p^{1-\alpha} \left[2 s_{S,p} \frac{1}{p} \sum_{i=1}^{p} \frac{1}{(1 + s_{S,p} x_{i,p}^{\alpha})(1 + s_{T,p} x_{i,p}^{\alpha})} - \frac{s_{S,p}^2}{1 - \gamma_{S,p}} \frac{1}{p} \sum_{i=1}^{p} \frac{x_{i,p}^{\alpha}}{(1 + s_{S,p} x_{i,p}^{\alpha})^2 (1 + s_{T,p} x_{i,p}^{\alpha})} \right]. \end{aligned} \tag{75}$$

By Riemann-sum convergence,

$$= p^{1-\alpha} \left[2 s_S \int_0^1 \frac{dx}{(1 + s_S x^{\alpha})(1 + s_T x^{\alpha})} - \frac{s_S^2}{1 - \gamma_S} \int_0^1 \frac{x^{\alpha}}{(1 + s_S x^{\alpha})^2 (1 + s_T x^{\alpha})} \, dx \right] + o(p^{1-\alpha}). \tag{76}$$

Since $\alpha > 1$, this term vanishes as $p \to \infty$.

We next consider

$$\frac{\sigma^2}{n(1 - \gamma_{T,p})} \operatorname{Tr}\left[D_S \Sigma (\Sigma + \tau_{T,p} I_p)^{-2} \right].$$

Expanding D_S gives

$$\frac{\sigma^2}{n(1 - \gamma_{T,p})} \operatorname{Tr}\left[D_S \Sigma (\Sigma + \tau_{T,p} I_p)^{-2} \right] = \frac{2\sigma^2 \tau_{S,p}}{n(1 - \gamma_{T,p})} \sum_{i=1}^{p} \frac{i^{-\alpha}}{i^{-\alpha} + \tau_{S,p}} \frac{i^{-\alpha}}{(i^{-\alpha} + \tau_{T,p})^2}$$

$$-\frac{\sigma^2\tau_{S,p}^2}{n(1-\gamma_{T,p})(1-\gamma_{S,p})}\sum_{i=1}^{p}\frac{i^{-\alpha}}{(i^{-\alpha}+\tau_{S,p})^2}\frac{i^{-\alpha}}{(i^{-\alpha}+\tau_{T,p})^2}. \tag{77}$$

Since

$$\frac{i^{-\alpha}}{(i^{-\alpha}+\tau_{T,p})^2}=p^{\alpha}\frac{x_{i,p}^{\alpha}}{(1+s_{T,p}x_{i,p}^{\alpha})^2}, \tag{78}$$

we have

$$\tau_{S,p}\frac{i^{-\alpha}}{i^{-\alpha}+\tau_{S,p}}\frac{i^{-\alpha}}{(i^{-\alpha}+\tau_{T,p})^2}=s_{S,p}\frac{x_{i,p}^{\alpha}}{(1+s_{S,p}x_{i,p}^{\alpha})(1+s_{T,p}x_{i,p}^{\alpha})^2}, \tag{79}$$

and

$$\tau_{S,p}^2\frac{i^{-\alpha}}{(i^{-\alpha}+\tau_{S,p})^2}\frac{i^{-\alpha}}{(i^{-\alpha}+\tau_{T,p})^2}=s_{S,p}^2\frac{x_{i,p}^{2\alpha}}{(1+s_{S,p}x_{i,p}^{\alpha})^2(1+s_{T,p}x_{i,p}^{\alpha})^2}. \tag{80}$$

Using $\frac{1}{n}=\frac{\phi_{T,p}}{p}$, we obtain

$$\begin{aligned}\frac{\sigma^2}{n(1-\gamma_{T,p})}\operatorname{Tr}\left[D_S\Sigma(\Sigma+\tau_{T,p}I_p)^{-2}\right]&=\frac{2\sigma^2\phi_{T,p}s_{S,p}}{1-\gamma_{T,p}}\frac{1}{p}\sum_{i=1}^{p}\frac{x_{i,p}^{\alpha}}{(1+s_{S,p}x_{i,p}^{\alpha})(1+s_{T,p}x_{i,p}^{\alpha})^2}\\&-\frac{\sigma^2\phi_{T,p}s_{S,p}^2}{(1-\gamma_{T,p})(1-\gamma_{S,p})}\frac{1}{p}\sum_{i=1}^{p}\frac{x_{i,p}^{2\alpha}}{(1+s_{S,p}x_{i,p}^{\alpha})^2(1+s_{T,p}x_{i,p}^{\alpha})^2}.\end{aligned} \tag{81}$$

Taking $p\to\infty$ yields

$$\begin{aligned}\frac{\sigma^2}{n(1-\gamma_{T,p})}\operatorname{Tr}\left[D_S\Sigma(\Sigma+\tau_{T,p}I_p)^{-2}\right]&\longrightarrow\frac{2\sigma^2\phi_T s_S}{1-\gamma_T}\int_0^1\frac{x^{\alpha}}{(1+s_Sx^{\alpha})(1+s_Tx^{\alpha})^2}\,dx\\&-\frac{\sigma^2\phi_T s_S^2}{(1-\gamma_T)(1-\gamma_S)}\int_0^1\frac{x^{2\alpha}}{(1+s_Sx^{\alpha})^2(1+s_Tx^{\alpha})^2}\,dx.\end{aligned} \tag{82}$$

Thus, unlike the signal-dependent contribution, the Stage-I noise contribution remains of order one.

Finally,

$$\begin{aligned}-2\operatorname{Tr}\left[A_S\Sigma(\Sigma+\tau_{T,p}I_p)^{-1}\right]&=-2\tau_{S,p}\sum_{i=1}^{p}\frac{i^{-\alpha}}{i^{-\alpha}+\tau_{S,p}}\frac{i^{-\alpha}}{i^{-\alpha}+\tau_{T,p}}\\&=-2p^{1-\alpha}s_{S,p}\left[\frac{1}{p}\sum_{i=1}^{p}\frac{1}{(1+s_{S,p}x_{i,p}^{\alpha})(1+s_{T,p}x_{i,p}^{\alpha})}\right].\end{aligned} \tag{83}$$

Hence

$$-2\operatorname{Tr}\left[A_S\Sigma(\Sigma+\tau_{T,p}I_p)^{-1}\right]=-2s_Sp^{1-\alpha}\int_0^1\frac{dx}{(1+s_Sx^{\alpha})(1+s_Tx^{\alpha})}+o(p^{1-\alpha}), \tag{84}$$

and therefore this term also vanishes because $\alpha>1$. Moreover, the leading $2s_S$ contribution in (76) cancels with (84). Thus, their combined contribution is

$$\begin{aligned}&\operatorname{Tr}\left[D_S\Sigma(\Sigma+\tau_{T,p}I_p)^{-1}\right]-2\operatorname{Tr}\left[A_S\Sigma(\Sigma+\tau_{T,p}I_p)^{-1}\right]\\&=-\frac{s_S^2}{1-\gamma_S}p^{1-\alpha}\int_0^1\frac{x^{\alpha}}{(1+s_Sx^{\alpha})^2(1+s_Tx^{\alpha})}\,dx+o(p^{1-\alpha}).\end{aligned} \tag{85}$$

Since $\alpha>1$, the entire signal-dependent contribution vanishes in the proportional limit.

C.5.1 Limiting Stage-I-to-Stage-II risk improvement

Combining the preceding calculations gives

$$\mathsf{W2SG}_p^{\mathrm{DE}}=\frac{2\sigma^2\phi_T s_S}{1-\gamma_T}\int_0^1\frac{x^{\alpha}}{(1+s_Sx^{\alpha})(1+s_Tx^{\alpha})^2}\,dx-\frac{\sigma^2\phi_T s_S^2}{(1-\gamma_T)(1-\gamma_S)}\int_0^1\frac{x^{2\alpha}}{(1+s_Sx^{\alpha})^2(1+s_Tx^{\alpha})^2}\,dx$$

$$-\frac{s_S^2}{1-\gamma_S}p^{1-\alpha}\int_0^1 \frac{x^\alpha}{(1+s_S x^\alpha)^2(1+s_T x^\alpha)}\,dx + o(1). \tag{86}$$

Because $\alpha > 1$, $p^{1-\alpha} \longrightarrow 0$, and therefore

$$\mathsf{W2SG}_p^{\mathrm{DE}} \longrightarrow \frac{2\sigma^2\phi_T s_S}{1-\gamma_T}\int_0^1 \frac{x^\alpha}{(1+s_S x^\alpha)(1+s_T x^\alpha)^2}\,dx - \frac{\sigma^2\phi_T s_S^2}{(1-\gamma_T)(1-\gamma_S)}\int_0^1 \frac{x^{2\alpha}}{(1+s_S x^\alpha)^2(1+s_T x^\alpha)^2}\,dx. \tag{87}$$

Combining the two terms over the common denominator $(1+s_S x^\alpha)^2$ gives (63). Define

$$A_y := \int_0^1 \frac{x^{y\alpha}}{(1+s_S x^\alpha)^2(1+s_T x^\alpha)^2}\,dx. \tag{88}$$

Then

$$\mathsf{W2SG}_p^{\mathrm{DE}} \longrightarrow \mathsf{W2SG}^{\mathrm{DE}} := \frac{\sigma^2\phi_T s_S}{1-\gamma_T}\left[2A_1 + \left(2 - \frac{1}{1-\gamma_S}\right)s_S A_2\right]. \tag{89}$$

Define

$$J_\alpha := 2A_1 + \left(2 - \frac{1}{1-\gamma_S}\right)s_S A_2. \tag{90}$$

Since $\phi_T > 1, s_S > 0, 1-\gamma_T > 0$, the sign of the limiting deterministic-equivalent gain satisfies $\operatorname{sgn}\big(\mathsf{W2SG}^{\mathrm{DE}}\big) = \operatorname{sgn}(J_\alpha)$ whenever $\sigma^2 > 0$. Equivalently, for $\sigma^2 > 0$,

$$\mathsf{W2SG}^{\mathrm{DE}} > 0 \quad \Longleftrightarrow \quad J_\alpha > 0.$$

If $\sigma^2 = 0$, then instead $\mathsf{W2SG}^{\mathrm{DE}} = 0$.

Under the power-law covariance model, the proportional limit of the deterministic-equivalent risk improvement is therefore determined by α, ϕ_T, ϕ_S, and σ^2 through s_T, s_S, γ_T, and γ_S. □

Lemma C.10 (Monotonicity and limiting values of γ_S)**.** *For every $\alpha > 1$, the function γ_S is strictly increasing as a function of $\phi_S^{-1} \in (0,1)$. Moreover,*

$$\lim_{\phi_S^{-1}\downarrow 0}\gamma_S = \frac{\alpha-1}{\alpha}, \qquad \lim_{\phi_S^{-1}\uparrow 1}\gamma_S = 1. \tag{91}$$

Consequently:

1. *if $1 < \alpha < 2$, there exists a unique $\phi_{S,\gamma}^{-1}(\alpha) \in (0,1)$ such that*

$$\gamma_S < \frac{1}{2} \quad \textit{for } 0 < \phi_S^{-1} < \phi_{S,\gamma}^{-1}(\alpha),$$

and

$$\gamma_S > \frac{1}{2} \quad \textit{for } \phi_{S,\gamma}^{-1}(\alpha) < \phi_S^{-1} < 1;$$

2. *if $\alpha = 2$,*

$$\gamma_S > \frac{1}{2} \qquad \forall\, \phi_S^{-1} \in (0,1),$$

while

$$\gamma_S \downarrow \frac{1}{2} \qquad (\phi_S^{-1} \downarrow 0);$$

3. *if $\alpha > 2$,*

$$\gamma_S > \frac{1}{2} \qquad \forall\, \phi_S^{-1} \in (0,1).$$

Proof. Recall that $s_S > 0$ is determined implicitly by

$$\int_0^1 \frac{dx}{1 + s_S x^\alpha} = \frac{1}{\phi_S}, \tag{92}$$

and

$$\gamma_S = \phi_S \int_0^1 \frac{dx}{(1 + s_S x^\alpha)^2}. \tag{93}$$

We first derive a convenient closed-form relation for γ_S. Observe that

$$\frac{d}{dx}\left(\frac{x}{1 + sx^\alpha}\right) = \frac{1}{1 + sx^\alpha} - \frac{\alpha s x^\alpha}{(1 + sx^\alpha)^2}.$$

Integrating over $x \in [0, 1]$ gives

$$\frac{1}{1+s} = \int_0^1 \frac{dx}{1 + sx^\alpha} - \alpha \int_0^1 \frac{sx^\alpha}{(1 + sx^\alpha)^2}\, dx.$$

Using $\frac{sx^\alpha}{(1+sx^\alpha)^2} = \frac{1}{1+sx^\alpha} - \frac{1}{(1+sx^\alpha)^2}$, we obtain

$$\int_0^1 \frac{dx}{(1 + sx^\alpha)^2} = \frac{\alpha - 1}{\alpha} \int_0^1 \frac{dx}{1 + sx^\alpha} + \frac{1}{\alpha(1+s)}. \tag{94}$$

Setting $s = s_S$ and using (92) and (93), we find

$$\gamma_S = \phi_S \left[\frac{\alpha - 1}{\alpha}\frac{1}{\phi_S} + \frac{1}{\alpha(1 + s_S)}\right] = \frac{\alpha - 1}{\alpha} + \frac{\phi_S}{\alpha(1 + s_S)}. \tag{95}$$

Equivalently, since $\phi_S = \left(\int_0^1 \frac{dx}{1+s_S x^\alpha}\right)^{-1}$, we may write

$$\gamma_S = \frac{\alpha - 1}{\alpha} + \frac{1}{\alpha}\left[(1 + s_S)\int_0^1 \frac{dx}{1 + s_S x^\alpha}\right]^{-1}. \tag{96}$$

We next establish monotonicity. First,

$$\frac{d}{ds}\int_0^1 \frac{dx}{1 + sx^\alpha} = -\int_0^1 \frac{x^\alpha}{(1 + sx^\alpha)^2}\, dx < 0.$$

Hence $\phi_S^{-1} = \int_0^1 \frac{dx}{1+s_S x^\alpha}$ is a strictly decreasing function of s_S. It therefore remains to show that γ_S is also strictly decreasing as a function of s_S. By (96), it is enough to show that

$$\left[(1 + s)\int_0^1 \frac{dx}{1 + sx^\alpha}\right]^{-1}$$

is strictly decreasing in s. Its logarithmic derivative is

$$\frac{d}{ds}\log\left[\left((1 + s)\int_0^1 \frac{dx}{1 + sx^\alpha}\right)^{-1}\right] = -\frac{1}{1+s} + \frac{\displaystyle\int_0^1 \frac{x^\alpha}{(1 + sx^\alpha)^2}\, dx}{\displaystyle\int_0^1 \frac{dx}{1 + sx^\alpha}}. \tag{97}$$

Now,

$$\int_0^1 \frac{dx}{1 + sx^\alpha} - (1 + s)\int_0^1 \frac{x^\alpha}{(1 + sx^\alpha)^2}\, dx = \int_0^1 \frac{1 + sx^\alpha - (1 + s)x^\alpha}{(1 + sx^\alpha)^2}\, dx = \int_0^1 \frac{1 - x^\alpha}{(1 + sx^\alpha)^2}\, dx > 0. \tag{98}$$

Therefore,

$$\frac{\displaystyle\int_0^1 \frac{x^\alpha}{(1 + sx^\alpha)^2}\, dx}{\displaystyle\int_0^1 \frac{dx}{1 + sx^\alpha}} < \frac{1}{1+s}.$$

Substituting this into (97) shows that the logarithmic derivative is strictly negative. Hence γ_S is strictly decreasing in s_S. Since both γ_S and ϕ_S^{-1} are strictly decreasing functions of s_S, it follows that γ_S is a strictly increasing function of $\phi_S^{-1} \in (0,1)$.

We now determine the endpoint limits. As $s_S \downarrow 0$, $\int_0^1 \frac{dx}{1+s_S x^\alpha} \longrightarrow 1$, so $\phi_S^{-1} \uparrow 1$. Moreover, from (93), $\gamma_S = \phi_S \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2} \longrightarrow 1$. Thus, $\lim_{\phi_S^{-1}\uparrow 1} \gamma_S = 1$. Next consider $s_S \to \infty$. With the substitution $u = s_S^{1/\alpha} x$, we have

$$\int_0^1 \frac{dx}{1+s_S x^\alpha} = s_S^{-1/\alpha} \int_0^{s_S^{1/\alpha}} \frac{du}{1+u^\alpha}. \tag{99}$$

Since $\alpha > 1$, Define $C_\alpha := \int_0^\infty \frac{du}{1+u^\alpha} < \infty$, and therefore

$$\int_0^1 \frac{dx}{1+s_S x^\alpha} \sim C_\alpha s_S^{-1/\alpha}.$$

In particular, $\phi_S^{-1} \downarrow 0$. Furthermore,

$$(1+s_S) \int_0^1 \frac{dx}{1+s_S x^\alpha} \sim C_\alpha s_S^{1-1/\alpha} \longrightarrow \infty,$$

because $\alpha > 1$. Hence, by (96), $\gamma_S \longrightarrow \frac{\alpha-1}{\alpha}$. Thus, $\lim_{\phi_S^{-1}\downarrow 0} \gamma_S = \frac{\alpha-1}{\alpha}$.

It remains to distinguish the three cases. If $1 < \alpha < 2$, then $\frac{\alpha-1}{\alpha} < \frac{1}{2} < 1$. Since γ_S is continuous and strictly increasing in ϕ_S^{-1}, there exists a unique $\phi_{S,\gamma}^{-1}(\alpha) \in (0,1)$ such that $\gamma_S = \frac{1}{2}$. Consequently,

$$\gamma_S < \frac{1}{2} \qquad \text{for } 0 < \phi_S^{-1} < \phi_{S,\gamma}^{-1}(\alpha),$$

whereas

$$\gamma_S > \frac{1}{2} \qquad \text{for } \phi_{S,\gamma}^{-1}(\alpha) < \phi_S^{-1} < 1.$$

If $\alpha = 2$, then $\lim_{\phi_S^{-1}\downarrow 0} \gamma_S = \frac{1}{2}$. Strict monotonicity therefore implies

$$\gamma_S > \frac{1}{2} \qquad \forall\, \phi_S^{-1} \in (0,1),$$

with

$$\gamma_S \downarrow \frac{1}{2} \qquad \text{as } \phi_S^{-1} \downarrow 0.$$

Finally, if $\alpha > 2$, then $\frac{\alpha-1}{\alpha} > \frac{1}{2}$. Since this is already the lower endpoint limit of γ_S, strict monotonicity gives

$$\gamma_S > \frac{1}{2} \qquad \forall\, \phi_S^{-1} \in (0,1).$$

This proves all three cases. □

Lemma C.11 (Removal of the Stage-I weight)**.** *For every $s_T, s_S > 0$,*

$$\frac{A_2}{A_1} \le \frac{B_2(s_S)}{B_1(s_S)}. \tag{100}$$

where

$$B_y(s_S) = \int_0^1 \frac{x^{y\alpha}}{(1+s_S x^\alpha)^2}\, dx \tag{101}$$

Proof. Define

$$w(x) := \frac{x^\alpha}{(1+s_S x^\alpha)^2}, \qquad h(x) := \frac{1}{(1+s_T x^\alpha)^2},$$

and let μ be the probability measure on $[0,1]$ defined by

$$d\mu(x) = \frac{w(x)\,dx}{\int_0^1 w(t)\,dt}.$$

Then, by the definitions of A_1 and A_2,

$$\frac{A_2}{A_1} = \frac{\mathbb{E}_\mu[x^\alpha h(x)]}{\mathbb{E}_\mu[h(x)]}.$$

Since x^α is increasing on $[0,1]$, whereas $h(x)$ is decreasing, the two functions are oppositely monotone. Hence $\mathrm{Cov}_\mu(x^\alpha, h(x)) \le 0$, and therefore $\mathbb{E}_\mu[x^\alpha h(x)] \le \mathbb{E}_\mu[x^\alpha]\,\mathbb{E}_\mu[h(x)]$. Since $\mathbb{E}_\mu[h(x)] > 0$, division gives

$$\frac{A_2}{A_1} \le \mathbb{E}_\mu[x^\alpha].$$

Finally,

$$\mathbb{E}_\mu[x^\alpha] = \frac{\int_0^1 \frac{x^{2\alpha}}{(1+s_S x^\alpha)^2}\,dx}{\int_0^1 \frac{x^\alpha}{(1+s_S x^\alpha)^2}\,dx} = \frac{B_2(s_S)}{B_1(s_S)}.$$

This proves (100). □

Lemma C.12 (Scalar positivity up to the critical exponent)**.** *Let $1 < \alpha \le 2$. Then, for every $s > 0$,*

$$G_\alpha(s) := \int_0^1 \frac{dx}{1+sx^\alpha} - \left(\int_0^1 \frac{dx}{(1+sx^\alpha)^2}\right)\left(2 - \int_0^1 \frac{dx}{1+sx^\alpha}\right) > 0. \tag{102}$$

Proof. We first derive an identity relating the two integrals appearing in G_α. Since

$$\frac{d}{dx}\left(\frac{x}{1+sx^\alpha}\right) = \frac{1}{1+sx^\alpha} - \frac{\alpha s x^\alpha}{(1+sx^\alpha)^2},$$

integration over $[0,1]$ gives

$$\frac{1}{1+s} = \int_0^1 \frac{dx}{1+sx^\alpha} - \alpha \int_0^1 \frac{sx^\alpha}{(1+sx^\alpha)^2}\,dx.$$

Using

$$\frac{sx^\alpha}{(1+sx^\alpha)^2} = \frac{1}{1+sx^\alpha} - \frac{1}{(1+sx^\alpha)^2},$$

we obtain

$$\int_0^1 \frac{dx}{(1+sx^\alpha)^2} = \frac{\alpha-1}{\alpha}\int_0^1 \frac{dx}{1+sx^\alpha} + \frac{1}{\alpha(1+s)}. \tag{103}$$

Moreover,

$$\frac{d}{ds}\int_0^1 \frac{dx}{1+sx^\alpha} = -\int_0^1 \frac{x^\alpha}{(1+sx^\alpha)^2}\,dx,$$

and hence

$$s\frac{d}{ds}\int_0^1 \frac{dx}{1+sx^\alpha} = \int_0^1 \frac{dx}{(1+sx^\alpha)^2} - \int_0^1 \frac{dx}{1+sx^\alpha}. \tag{104}$$

Combining this with (103) gives

$$s\frac{d}{ds}\int_0^1 \frac{dx}{1+sx^\alpha} = \frac{1}{\alpha}\left[\frac{1}{1+s} - \int_0^1 \frac{dx}{1+sx^\alpha}\right]. \tag{105}$$

We first examine G_α near $s = 0$. Expanding the integrands gives

$$\int_0^1 \frac{dx}{1+sx^\alpha} = 1 - \frac{s}{\alpha+1} + \frac{s^2}{2\alpha+1} + O(s^3), \tag{106}$$

$$\int_0^1 \frac{dx}{(1+sx^\alpha)^2} = 1 - \frac{2s}{\alpha+1} + \frac{3s^2}{2\alpha+1} + O(s^3). \tag{107}$$

Therefore,

$$G_\alpha(s) = \left[\frac{2}{(\alpha+1)^2} - \frac{1}{2\alpha+1}\right] s^2 + O(s^3) = \frac{2-(\alpha-1)^2}{(\alpha+1)^2(2\alpha+1)} s^2 + O(s^3). \tag{108}$$

For $1 < \alpha \le 2$, $2-(\alpha-1)^2 > 0$, and consequently

$$G_\alpha(s) > 0 \qquad \text{for all sufficiently small } s > 0. \tag{109}$$

Suppose, toward a contradiction, that G_α has a positive zero, and let $s_\star > 0$ denote its first positive zero. Then

$$G_\alpha(s_\star) = 0, \qquad G_\alpha(s) > 0 \quad \text{for } 0 < s < s_\star.$$

Define $u := 1 - \int_0^1 \frac{dx}{1+s_\star x^\alpha}$. Since $s_\star > 0$, $0 < u < 1$. Thus, $\int_0^1 \frac{dx}{1+s_\star x^\alpha} = 1-u$. The condition $G_\alpha(s_\star) = 0$ then implies

$$\int_0^1 \frac{dx}{(1+s_\star x^\alpha)^2} = \frac{1-u}{1+u}. \tag{110}$$

Define $b := \frac{1-(\alpha-1)u}{\alpha}$. Combining (103) and (110) gives

$$1 + s_\star = \frac{1+u}{\alpha(1-u)b}. \tag{111}$$

Equivalently,

$$s_\star = \frac{u(1+b)}{(1-u)b}. \tag{112}$$

We next note that $b > 0$. If $1 < \alpha < 2$, then

$$b = \frac{1-(\alpha-1)u}{\alpha} > \frac{2-\alpha}{\alpha} > 0,$$

whereas, if $\alpha = 2$, $b = \frac{1-u}{2} > 0$. Hence

$$b > 0 \qquad \text{for all } 1 < \alpha \le 2. \tag{113}$$

Differentiating the definition of G_α and using (103), (105), (110), (111), and (112), a direct simplification yields

$$G'_\alpha(s_\star) = \frac{u(1-u)^2 b\big(\alpha b^2 + 2 - \alpha\big)}{(1+u)^2(1+b)}. \tag{114}$$

If $1 < \alpha < 2$, then $\alpha b^2 + 2 - \alpha > 0$. If $\alpha = 2$, then $\alpha b^2 + 2 - \alpha = 2b^2 > 0$. Together with $0 < u < 1, b > 0$, this implies $G'_\alpha(s_\star) > 0$. On the other hand, $s_\star$ is the first zero reached from a region where $G_\alpha(s) > 0$. Since G_α is differentiable, this requires $G'_\alpha(s_\star) \le 0$, which is a contradiction.

Therefore G_α has no positive zero. Since it is positive for all sufficiently small $s > 0$, we conclude that

$$G_\alpha(s) > 0 \qquad \forall\, s > 0.$$

□

Theorem C.13 (Global proportional positivity for $1 < \alpha \le 2$)**.** *Let* $1 < \alpha \le 2, \phi_T > 1, \phi_S > 1, \sigma^2 > 0$. *Then*

$$\mathsf{W2SG}^{\mathrm{DE}} > 0. \tag{115}$$

for every nontrivial proportional Stage-II sampling fraction $0 < \phi_S^{-1} < 1$.

Proof. We distinguish the two possible signs of the Stage-II coefficient.

If $\gamma_S \le 1/2$, then $\frac{1-2\gamma_S}{1-\gamma_S} \ge 0$. Since $A_1 > 0$ and $A_2 > 0$, equation (90) immediately gives $\mathsf{W2SG}^{\mathrm{DE}} > 0$

It remains to consider the case $\gamma_S > 1/2$. In this case, (90) shows that $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) > 0$ is equivalent to

$$\frac{2\gamma_S - 1}{1-\gamma_S} s_S \frac{A_2}{A_1} < 2. \tag{116}$$

By Lemma C.11,

$$\frac{A_2}{A_1} \le \frac{B_2(s_S)}{B_1(s_S)}.$$

Recall that $\phi_S^{-1} = \int_0^1 \frac{dx}{1+s_S x^\alpha}$, and $\gamma_S = \phi_S \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}$. Therefore,

$$\gamma_S = \frac{\displaystyle\int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}}{\displaystyle\int_0^1 \frac{dx}{1+s_S x^\alpha}}. \tag{117}$$

It follows that

$$\frac{2\gamma_S - 1}{1-\gamma_S} = \frac{2\displaystyle\int_0^1 \frac{dx}{(1+s_S x^\alpha)^2} - \int_0^1 \frac{dx}{1+s_S x^\alpha}}{\displaystyle\int_0^1 \frac{dx}{1+s_S x^\alpha} - \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}}. \tag{118}$$

Next, directly from the definitions of B_1 and B_2,

$$\int_0^1 \frac{dx}{1+s_S x^\alpha} - \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2} = s_S B_1(s_S), \tag{119}$$

$$1 - 2\int_0^1 \frac{dx}{1+s_S x^\alpha} + \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2} = s_S^2 B_2(s_S). \tag{120}$$

Hence

$$s_S \frac{B_2(s_S)}{B_1(s_S)} = \frac{1 - 2\displaystyle\int_0^1 \frac{dx}{1+s_S x^\alpha} + \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}}{\displaystyle\int_0^1 \frac{dx}{1+s_S x^\alpha} - \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}}. \tag{121}$$

Combining (118), (121), and Lemma C.11, we obtain

$$\begin{aligned} &\frac{2\gamma_S - 1}{1-\gamma_S} s_S \frac{A_2}{A_1} \\ &\quad \le \frac{\left(2\displaystyle\int_0^1 \frac{dx}{(1+s_S x^\alpha)^2} - \int_0^1 \frac{dx}{1+s_S x^\alpha}\right)\left(1 - 2\displaystyle\int_0^1 \frac{dx}{1+s_S x^\alpha} + \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}\right)}{\left(\displaystyle\int_0^1 \frac{dx}{1+s_S x^\alpha} - \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}\right)^2}. \end{aligned} \tag{122}$$

We now apply Lemma C.12. A direct algebraic expansion gives

$$\begin{aligned} &2\left(\int_0^1 \frac{dx}{1+s_S x^\alpha} - \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}\right)^2 \\ &\quad - \left(2\int_0^1 \frac{dx}{(1+s_S x^\alpha)^2} - \int_0^1 \frac{dx}{1+s_S x^\alpha}\right)\left(1 - 2\int_0^1 \frac{dx}{1+s_S x^\alpha} + \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}\right) \end{aligned}$$

$$= \int_0^1 \frac{dx}{1+s_S x^\alpha} - \left(\int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}\right)\left(2 - \int_0^1 \frac{dx}{1+s_S x^\alpha}\right). \tag{123}$$

By Lemma C.12, applied at $s = s_S$, the right-hand side is strictly positive. Hence

$$\frac{\left(2\int_0^1 \frac{dx}{(1+s_S x^\alpha)^2} - \int_0^1 \frac{dx}{1+s_S x^\alpha}\right)\left(1 - 2\int_0^1 \frac{dx}{1+s_S x^\alpha} + \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}\right)}{\left(\int_0^1 \frac{dx}{1+s_S x^\alpha} - \int_0^1 \frac{dx}{(1+s_S x^\alpha)^2}\right)^2} < 2. \tag{124}$$

Together with (122), this proves (116). Therefore, $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) > 0$ also in the case $\gamma_S > 1/2$. Thus,

$$\mathsf{W2SG}^{\mathrm{DE}} > 0 \qquad \text{for all } 1 < \alpha \le 2, \quad \phi_T > 1, \quad \phi_S > 1.$$

□

The situation changes qualitatively when $\alpha > 2$. By Lemma C.10,

$$\gamma_S > \frac{1}{2} \qquad \forall\, \phi_S^{-1} \in (0,1),$$

and hence

$$\mathsf{W2SG}^{\mathrm{DE}} > 0 \iff 2A_1 > \frac{2\gamma_S - 1}{1-\gamma_S} s_S A_2. \tag{125}$$

Unlike the case $1 < \alpha \le 2$, this inequality is not automatically satisfied.

We first prove that it necessarily fails near the sparse edge.

Proposition C.14 (Sparse proportional negativity for $\alpha > 2$)**.** *Let $\alpha > 2, \phi_T > 1, \sigma^2 > 0$. Then there exists $\varepsilon_0 = \varepsilon_0(\alpha, \phi_T) > 0$ such that*

$$0 < \phi_S^{-1} < \varepsilon_0 \quad \Longrightarrow \quad \mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) < 0. \tag{126}$$

Proof. The sparse proportional limit corresponds to

$$\phi_S^{-1} \downarrow 0 \qquad \Longleftrightarrow \qquad s_S \to \infty.$$

From Lemma C.10, $\gamma_S \longrightarrow \frac{\alpha-1}{\alpha}$, and hence

$$\frac{1-2\gamma_S}{1-\gamma_S} \longrightarrow 2 - \alpha < 0. \tag{127}$$

Write $h_1(x) := \frac{1}{(1+s_T x^\alpha)^2}$. Then

$$A_1 = \int_0^1 \frac{x^\alpha}{(1+s_S x^\alpha)^2} h_1(x)\, dx.$$

Multiplying by s_S,

$$s_S A_1 = \int_0^1 \frac{s_S x^\alpha}{(1+s_S x^\alpha)^2} h_1(x)\, dx.$$

For every $x > 0$, $\frac{s_S x^\alpha}{(1+s_S x^\alpha)^2} \longrightarrow 0$, and $0 \le \frac{s_S x^\alpha}{(1+s_S x^\alpha)^2} \le \frac{1}{4}$. Thus dominated convergence yields

$$s_S A_1 \longrightarrow 0. \tag{128}$$

Similarly,

$$s_S^2 A_2 = \int_0^1 \frac{(s_S x^\alpha)^2}{(1+s_S x^\alpha)^2} h_1(x)\, dx.$$

For every $x > 0$, $\frac{(s_S x^\alpha)^2}{(1+s_S x^\alpha)^2} \longrightarrow 1$, while the ratio is bounded by 1. Therefore

$$s_S^2 A_2 \longrightarrow C_1(s_T) := \int_0^1 \frac{dx}{(1+s_T x^\alpha)^2} > 0. \tag{129}$$

Multiplying by s_S,

$$s_S J_\alpha = 2s_S A_1 + \frac{1-2\gamma_S}{1-\gamma_S} s_S^2 A_2.$$

Using (127), (128), and (129),

$$s_S J_\alpha \longrightarrow (2-\alpha) C_1(s_T) < 0.$$

Therefore $J_\alpha < 0$ for all sufficiently large s_S. Since ϕ_S^{-1} is a strictly decreasing function of s_S, there exists $\varepsilon_0 = \varepsilon_0(\alpha, \phi_T) > 0$ such that

$$0 < \phi_S^{-1} < \varepsilon_0 \quad \Longrightarrow \quad J_\alpha < 0.$$

Because $\sigma^2 > 0$, the sign relation (90) then gives $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) < 0$. □

Thus for every $\alpha > 2$, there always exist proportional ratios p/m sufficiently large for which the W2SG condition fails.

We next determine when a positive region necessarily occurs near the opposite boundary $\phi_S^{-1} \uparrow 1$.

Define the near-square quantities

$$B_y(s_T) := \int_0^1 \frac{x^{y\alpha}}{(1+s_T x^\alpha)^2}\, dx, \tag{130}$$

and

$$L_\alpha(\phi_T) := 2B_1(s_T) - (\alpha+1) B_2(s_T). \tag{131}$$

Proposition C.15 (Near-square sign for $\alpha > 2$)**.** *Let $\alpha > 2, \phi_T > 1$. Then*

$$\lim_{\phi_S^{-1} \uparrow 1} J_\alpha(\phi_T, \phi_S) = L_\alpha(\phi_T). \tag{132}$$

Consequently:

1. *if $L_\alpha(\phi_T) > 0$, then $J_\alpha > 0$ for all $\phi_S^{-1} < 1$ sufficiently close to 1;*
2. *if $L_\alpha(\phi_T) < 0$, then $J_\alpha < 0$ for all $\phi_S^{-1} < 1$ sufficiently close to 1;*
3. *if $L_\alpha(\phi_T) = 0$, the leading near-square limit is degenerate and a higher-order expansion is required.*

Proof. As $\phi_S^{-1} \uparrow 1$, the Stage-II fixed point gives $s_S \downarrow 0$. From

$$F_\alpha(s) = 1 - \frac{s}{\alpha+1} + O(s^2), \qquad H_\alpha(s) = 1 - \frac{2s}{\alpha+1} + O(s^2),$$

we obtain

$$\gamma_S = \frac{H_\alpha(s_S)}{F_\alpha(s_S)} = 1 - \frac{s_S}{\alpha+1} + O(s_S^2).$$

Thus

$$s_S \frac{2\gamma_S - 1}{1-\gamma_S} \longrightarrow \alpha + 1. \tag{133}$$

Moreover, dominated convergence gives $A_1 \longrightarrow B_1(s_T), A_2 \longrightarrow B_2(s_T)$. Since $\gamma_S > 1/2$ for $\alpha > 2$, $J_\alpha = 2A_1 - \frac{2\gamma_S - 1}{1-\gamma_S} s_S A_2$. Hence

$$J_\alpha \longrightarrow 2B_1(s_T) - (\alpha+1) B_2(s_T) = L_\alpha(\phi_T).$$

The sign conclusions follow by continuity. □

Combining the sparse and near-square results immediately gives a crossing criterion.

Theorem C.16 (Existence of both W2SG and non-W2SG proportional ratios for $\alpha > 2$)**.** *Assume $\alpha > 2, \phi_T > 1, \sigma^2 > 0$, and suppose $L_\alpha(\phi_T) > 0$. Then there exist $0 < \phi_{S,-}^{-1} < \phi_{S,+}^{-1} < 1$ such that*

$$\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_{S,-}) < 0, \qquad \mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_{S,+}) > 0.$$

Equivalently, there exists a proportional ratio p/m for which the W2SG condition fails and another proportional ratio p/m for which it holds. Moreover, by continuity, there exists at least one $\phi_{S,\star}^{-1} \in (0,1)$ such that $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_{S,\star}) = 0$.

Proof. By Proposition C.14, $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) < 0$ for all sufficiently small $\phi_S^{-1} > 0$. By Proposition C.15 and $L_\alpha(\phi_T) > 0$, $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) > 0$ for all $\phi_S^{-1} < 1$ sufficiently close to 1.

Choose one point from each region. Since $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S)$ is continuous as a function of $\phi_S^{-1} \in (0,1)$, the intermediate value theorem gives at least one zero between them. □

The condition $L_\alpha(\phi_T) > 0$ is automatically satisfied for a nontrivial range above the critical exponent.

Corollary C.17 (An unconditional crossing range above $\alpha = 2$)**.** *If $2 < \alpha \le 1 + \sqrt{2}$, then $L_\alpha(\phi_T) > 0$ for every $\phi_T > 1$. Consequently, if in addition $\sigma^2 > 0$, then for every $2 < \alpha \le 1 + \sqrt{2}$ and every $\phi_T > 1$, there exist both positive-W2SG and negative-W2SG proportional Stage-II ratios.*

Proof. Consider the probability measure

$$d\nu(x) = \frac{x^\alpha \, dx}{\int_0^1 t^\alpha \, dt}.$$

Then

$$\frac{B_2(s_T)}{B_1(s_T)} = \frac{\mathbb{E}_\nu\left[x^\alpha (1 + s_T x^\alpha)^{-2}\right]}{\mathbb{E}_\nu\left[(1 + s_T x^\alpha)^{-2}\right]}.$$

Since x^α is increasing and $(1 + s_T x^\alpha)^{-2}$ is strictly decreasing for $s_T > 0$, the same covariance argument as in Lemma C.11 gives the strict inequality

$$\frac{B_2(s_T)}{B_1(s_T)} < \frac{\int_0^1 x^{2\alpha} \, dx}{\int_0^1 x^\alpha \, dx} = \frac{\alpha + 1}{2\alpha + 1}. \tag{134}$$

Therefore

$$\frac{L_\alpha(\phi_T)}{B_1(s_T)} = 2 - (\alpha + 1)\frac{B_2(s_T)}{B_1(s_T)} > 2 - \frac{(\alpha+1)^2}{2\alpha + 1} = \frac{2 - (\alpha - 1)^2}{2\alpha + 1}. \tag{135}$$

For $2 < \alpha < 1 + \sqrt{2}$, the right-hand side is strictly positive. At $\alpha = 1 + \sqrt{2}$, the final non-strict lower bound equals zero, but (134) is strict because $s_T > 0$. Hence $L_\alpha(\phi_T) > 0$ also at the endpoint. □

For $\alpha > 1 + \sqrt{2}$, the condition $L_\alpha(\phi_T) > 0$ may still hold, depending on the Stage-I aspect ratio ϕ_T. The preceding theorem therefore remains applicable, but positivity near the square boundary is no longer implied by α alone. If $L_\alpha(\phi_T) \le 0$, the sparse region remains rigorously negative, but the present argument does not rule out the possibility of a positive component at an intermediate proportional ratio.

We next strengthen the preceding existence result by proving uniqueness of the proportional W2SG transition whenever $L_\alpha(\phi_T) > 0$. Specifically, we show that the gain crosses zero exactly once as the Stage-II sampling fraction varies from the sparse boundary to the near-square boundary.

The argument relies on two auxiliary results: a Stieltjes representation for the squared Stage-II fixed-point map and a strict curvature property for a corresponding Stieltjes quotient. These are established in Lemmas C.18 and C.19, respectively, and are then used to prove Theorem C.20.

Lemma C.18 (Stieltjes representation of the squared Stage-II fixed-point map)**.** *Let* $0 < \ell < \frac{1}{2}$*, and define*

$$F_\ell(s) := \ell \int_0^1 \frac{y^{\ell-1}}{1+sy}\, dy, \qquad s > 0.$$

Set

$$C_\ell := \int_0^1 \frac{u^{\ell-1} - u^{-\ell}}{1-u}\, du, \tag{136}$$

$$q_\ell(y) := C_\ell y^\ell + \sum_{k=0}^{\infty} \frac{y^{k+1}}{k+1-\ell}, \qquad 0 < y < 1, \tag{137}$$

$$w_\ell(y) := 2\ell^2 y^{\ell-1} q_\ell(y). \tag{138}$$

Then

$$C_\ell > 0, \qquad w_\ell(y) > 0 \quad (0 < y < 1),$$

and

$$\int_0^1 w_\ell(y)\, dy = 1.$$

Moreover,

$$F_\ell(s)^2 = \int_0^1 \frac{w_\ell(y)}{1+sy}\, dy, \qquad s > 0. \tag{139}$$

Thus F_ℓ^2 *is itself a Stieltjes transform of a probability density on* $(0, 1)$*.*

Proof. The proof has three steps: first we establish positivity of the density, then we compute all moments of w_ℓ, and finally we compare those moments with the power-series coefficients of F_ℓ^2.

Step 1: Positivity of C_ℓ and w_ℓ. Since $0 < \ell < 1/2$, we have $1 - 2\ell > 0$. For $0 < u < 1$,

$$u^{\ell-1} - u^{-\ell} = u^{\ell-1}\big(1 - u^{1-2\ell}\big) > 0.$$

Hence the integrand in (136) is positive on $(0, 1)$.

It is also integrable at both endpoints. Therefore $C_\ell > 0$. Using $\frac{1}{1-u} = \sum_{k=0}^{\infty} u^k, \qquad 0 < u < 1,$ and monotone convergence, we obtain the useful series representation

$$C_\ell = \sum_{k=0}^{\infty} \left(\frac{1}{k+\ell} - \frac{1}{k+1-\ell} \right) > 0. \tag{140}$$

Every term in the definition of $q_\ell(y)$ is therefore positive, so

$$q_\ell(y) > 0, \qquad 0 < y < 1.$$

Since $2\ell^2 y^{\ell-1} > 0$, it follows that

$$w_\ell(y) > 0, \qquad 0 < y < 1.$$

Step 2: Moments of w_ℓ. For $n \geq 0$, define $M_n := \int_0^1 y^n w_\ell(y)\, dy$. By (138) and Tonelli's theorem,

$$M_n = 2\ell^2 \int_0^1 y^{n+\ell-1} \left(C_\ell y^\ell + \sum_{k=0}^{\infty} \frac{y^{k+1}}{k+1-\ell} \right) dy = 2\ell^2 \left[\frac{C_\ell}{n+2\ell} + \sum_{k=0}^{\infty} \frac{1}{(k+1-\ell)(n+k+\ell+1)} \right]. \tag{141}$$

Now use the partial-fraction identity

$$\frac{1}{(k+1-\ell)(n+k+\ell+1)} = \frac{1}{n+2\ell} \left(\frac{1}{k+1-\ell} - \frac{1}{n+k+\ell+1} \right). \tag{142}$$

Substituting (142) into (141) gives

$$M_n = \frac{2\ell^2}{n+2\ell} \left[C_\ell + \sum_{k=0}^{\infty} \left(\frac{1}{k+1-\ell} - \frac{1}{n+k+\ell+1} \right) \right]. \tag{143}$$

Using (140), the terms involving $(k+1-\ell)^{-1}$ cancel, so

$$M_n = \frac{2\ell^2}{n+2\ell}\sum_{k=0}^{\infty}\left(\frac{1}{k+\ell}-\frac{1}{n+k+\ell+1}\right) = \frac{2\ell^2}{n+2\ell}\sum_{k=0}^{n}\frac{1}{k+\ell}. \tag{144}$$

The last expression can be symmetrized by observing that

$$\frac{1}{(k+\ell)(n-k+\ell)} = \frac{1}{n+2\ell}\left(\frac{1}{k+\ell}+\frac{1}{n-k+\ell}\right).$$

Summing over $k=0,\dots,n$ yields

$$\sum_{k=0}^{n}\frac{1}{(k+\ell)(n-k+\ell)} = \frac{2}{n+2\ell}\sum_{k=0}^{n}\frac{1}{k+\ell}.$$

Hence

$$M_n = \ell^2\sum_{k=0}^{n}\frac{1}{(k+\ell)(n-k+\ell)}. \tag{145}$$

In particular, setting $n=0$ gives $\int_0^1 w_\ell(y)\,dy = M_0 = \ell^2\frac{1}{\ell^2} = 1$. Thus w_ℓ is a probability density on $(0,1)$.

Step 3: Matching the power series. For $|s|<1$, $\frac{1}{1+sy} = \sum_{n=0}^{\infty}(-s)^n y^n$. Therefore

$$F_\ell(s) = \ell\int_0^1 y^{\ell-1}\sum_{n=0}^{\infty}(-s)^n y^n\,dy = \sum_{n=0}^{\infty}\frac{\ell(-s)^n}{n+\ell}. \tag{146}$$

Squaring (146) and using the Cauchy product gives

$$F_\ell(s)^2 = \sum_{n=0}^{\infty}(-s)^n\left[\ell^2\sum_{k=0}^{n}\frac{1}{(k+\ell)(n-k+\ell)}\right] = \sum_{n=0}^{\infty}(-s)^n M_n, \tag{147}$$

where the last equality follows from (145). Since $M_n = \int_0^1 y^n w_\ell(y)\,dy$, we obtain

$$F_\ell(s)^2 = \int_0^1 w_\ell(y)\sum_{n=0}^{\infty}(-sy)^n\,dy = \int_0^1\frac{w_\ell(y)}{1+sy}\,dy, \qquad |s|<1. \tag{148}$$

Finally, both sides of (139) are real analytic for $s>-1$. Since they agree on the nonempty interval $(-1,1)$, the identity theorem extends the equality to the whole connected domain $s>-1$, and hence in particular to every $s>0$. □

Lemma C.19 (Strict curvature at stationary points of a Stieltjes quotient)**.** *Let w be integrable and positive almost everywhere on $(0,1)$. Suppose r is continuous on $[0,1]$, satisfies*

$$r(0)=r(1)=0, \qquad r(y)>0 \quad (0<y<1),$$

and is strictly increasing up to a unique point in $(0,1)$ and strictly decreasing afterward. Define

$$M(s) := \int_0^1\frac{w(y)}{1+sy}\,dy, \qquad N(s) := \int_0^1\frac{r(y)w(y)}{1+sy}\,dy,$$

and $Q(s) := \frac{N(s)}{M(s)}$. Then every stationary point $s_0>0$ of Q satisfies $Q''(s_0)<0$. Consequently, Q has at most one stationary point on $(0,\infty)$.

Proof. Suppose $Q'(s_0)=0$, and set $c := Q(s_0)$. Since c is a weighted average of r with a positive weight, $0<c<\max_{[0,1]} r$. By strict unimodality of r, there are exactly two points $0<u<v<1$ such that $r(u)=r(v)=c$, and $r-c$ has sign pattern

$$-, \quad +, \quad -$$

on the three intervals determined by u and v.

Put

$$g(y) := (r(y) - c)w(y), \qquad T(s) := N(s) - cM(s).$$

Since $T = M(Q - c)$ and $Q'(s_0) = 0$, $T(s_0) = T'(s_0) = 0$. Hence

$$\int_0^1 \frac{g(y)}{1 + s_0 y}\, dy = 0, \qquad \int_0^1 \frac{yg(y)}{(1 + s_0 y)^2}\, dy = 0. \tag{149}$$

The sign pattern of g implies

$$\int_0^1 \frac{g(y)(y - u)(y - v)}{(1 + s_0 y)^3}\, dy < 0. \tag{150}$$

Using the identity

$$\frac{(y - u)(y - v)}{(1 + s_0 y)^3} = \frac{uv}{1 + s_0 y} - \frac{(u + v + 2s_0 uv)y}{(1 + s_0 y)^2} + \frac{(1 + s_0 u)(1 + s_0 v)y^2}{(1 + s_0 y)^3},$$

the first two terms vanish after integration by (149). Since $(1 + s_0 u)(1 + s_0 v) > 0$, (150) therefore gives

$$T''(s_0) = 2 \int_0^1 \frac{y^2 g(y)}{(1 + s_0 y)^3}\, dy < 0.$$

Differentiating $T = M(Q - c)$ twice at s_0 yields $M(s_0)Q''(s_0) = T''(s_0) < 0$. Thus $Q''(s_0) < 0$.

If Q had two stationary points, both would be strict local maxima. The minimum of Q on the closed interval joining them would then be attained at an interior point, producing a stationary point with nonnegative second derivative, a contradiction. Hence Q has at most one stationary point. □

Theorem C.20 (Uniqueness of the proportional transition for $\alpha > 2$)**.** *Assume $\alpha > 2, \phi_T > 1, \sigma^2 > 0, L_\alpha(\phi_T) > 0$. Let $s_T > 0$ be the Stage-I proportional fixed-point parameter. Then there exists a unique $s_{S,\mathrm{crit}} > 0$ such that*

$$J_\alpha(\phi_T, \phi_S) = 0 \qquad \Longleftrightarrow \qquad s_S = s_{S,\mathrm{crit}}. \tag{151}$$

Equivalently, there exists a unique proportional critical ratio

$$\phi_{m,\mathrm{crit}} := \left[\int_0^1 \frac{dx}{1 + s_{S,\mathrm{crit}} x^\alpha}\right]^{-1} > 1 \tag{152}$$

such that

$$\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) \begin{cases} > 0, & 1 < \phi_S < \phi_{m,\mathrm{crit}}, \\ = 0, & \phi_S = \phi_{m,\mathrm{crit}}, \\ < 0, & \phi_S > \phi_{m,\mathrm{crit}}. \end{cases} \tag{153}$$

Equivalently, in terms of the Stage-II sampling fraction,

$$\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) > 0 \quad \Longleftrightarrow \quad \phi_S^{-1} > \phi_{m,\mathrm{crit}}^{-1}. \tag{154}$$

Proof. Fix α and ϕ_T, and write $t := s_T, s := s_S$. Define $F(s) := \int_0^1 \frac{dx}{1+sx^\alpha}$ and

$$B_t(s) := \int_0^1 \frac{x^\alpha}{(1 + sx^\alpha)(1 + tx^\alpha)^2}\, dx. \tag{155}$$

By the Stage-II fixed-point equation,

$$F(s) = \phi_S^{-1}. \tag{156}$$

Moreover,

$$F(s) + sF'(s) = \int_0^1 \frac{dx}{(1 + sx^\alpha)^2},$$

and therefore

$$\gamma_S = \frac{F + sF'}{F}, \qquad 1 - \gamma_S = -\frac{sF'}{F}. \tag{157}$$

In particular, $F'(s) < 0$. Differentiating (155) under the integral sign gives

$$B_t'(s) = -A_2, \qquad B_t(s) + sB_t'(s) = A_1. \tag{158}$$

Using (90), (157), and (158), we obtain

$$J_\alpha = 2B_t + \frac{s}{1-\gamma_S} B_t' = 2B_t - \frac{F}{F'} B_t'. \tag{159}$$

Now define

$$R_t(s) := \frac{B_t(s)}{F(s)^2}. \tag{160}$$

A direct differentiation of (160) and comparison with (159) gives

$$J_\alpha = -\frac{F(s)^3}{F'(s)} R_t'(s). \tag{161}$$

Since $F > 0$ and $F' < 0$, the prefactor in (161) is strictly positive. Thus

$$\operatorname{sign} J_\alpha = \operatorname{sign} R_t'. \tag{162}$$

It remains to prove that R_t' has at most one zero. Put $\ell := \frac{1}{\alpha} \in (0, 1/2)$. The change of variables $y = x^\alpha$ gives

$$F(s) = \ell \int_0^1 \frac{y^{\ell-1}}{1+sy}\, dy \tag{163}$$

and

$$B_t(s) = \ell \int_0^1 \frac{y^\ell}{(1+ty)^2(1+sy)}\, dy. \tag{164}$$

Let q_ℓ and w_ℓ be defined by (137)–(138). By Lemma C.18,

$$F(s)^2 = \int_0^1 \frac{w_\ell(y)}{1+sy}\, dy.$$

Define

$$r_t(y) := \frac{y}{2\ell(1+ty)^2 q_\ell(y)}, \qquad 0 < y < 1. \tag{165}$$

Then (164) becomes

$$B_t(s) = \int_0^1 \frac{r_t(y) w_\ell(y)}{1+sy}\, dy,$$

and hence

$$R_t(s) = \frac{\displaystyle\int_0^1 \frac{r_t(y) w_\ell(y)}{1+sy}\, dy}{\displaystyle\int_0^1 \frac{w_\ell(y)}{1+sy}\, dy}. \tag{166}$$

We next show that r_t is strictly unimodal. From (137),

$$q_\ell(y) \sim C_\ell y^\ell \qquad (y \downarrow 0),$$

while

$$q_\ell(y) \longrightarrow \infty \qquad (y \uparrow 1),$$

because the series in (137) diverges at $y = 1$. Thus r_t extends continuously to $[0, 1]$ with $r_t(0) = r_t(1) = 0, r_t(y) > 0 (0 < y < 1)$. Write $y = e^v$, $v < 0$. Then

$$q_\ell(e^v) = C_\ell e^{\ell v} + \sum_{k=0}^{\infty} \frac{e^{(k+1)v}}{k+1-\ell}.$$

The series and its derivatives converge locally uniformly for $v < 0$. Writing its positive terms as $a_j e^{\lambda_j v}$ and normalizing them to probabilities $\pi_j(v) = \frac{a_j e^{\lambda_j v}}{q_\ell(e^v)}$, we obtain

$$\frac{d^2}{dv^2} \log q_\ell(e^v) = \sum_j \pi_j(v) \lambda_j^2 - \left(\sum_j \pi_j(v) \lambda_j \right)^2 > 0,$$

because the exponents $\ell, 1, 2, \ldots$ are distinct. Consequently,

$$\frac{d^2}{dv^2}\log r_t(e^v) = -\frac{2te^v}{(1+te^v)^2} - \frac{d^2}{dv^2}\log q_\ell(e^v) < 0. \tag{167}$$

Hence $\log r_t(e^v)$ is strictly concave. Since r_t is positive on $(0,1)$ and vanishes at both endpoints, it has a unique interior maximum and is strictly increasing before that point and strictly decreasing afterward.

Lemma C.19, applied to (166), now implies

$$R_t'(s_0) = 0 \quad \Longrightarrow \quad R_t''(s_0) < 0. \tag{168}$$

In particular, R_t has at most one stationary point, and by (162), J_α has at most one zero as a function of s_S.

It remains only to establish existence. By the assumption $L_\alpha(\phi_T) > 0$, Proposition C.15 gives $J_\alpha > 0$ for $s_S > 0$ sufficiently small. On the other hand, Proposition C.14 gives $J_\alpha < 0$ for s_S sufficiently large. Continuity therefore gives at least one zero $s_{S,\mathrm{crit}}$, and the preceding argument shows that this zero is unique. Moreover,

$$J_\alpha \begin{cases} > 0, & 0 < s_S < s_{S,\mathrm{crit}}, \\ = 0, & s_S = s_{S,\mathrm{crit}}, \\ < 0, & s_S > s_{S,\mathrm{crit}}. \end{cases} \tag{169}$$

Finally, by (156) and $F'(s) < 0$, ϕ_S^{-1} is strictly decreasing in s_S, equivalently ϕ_S is strictly increasing in s_S. Thus (169) is exactly the sign pattern in (153). The prefactor relating $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S)$ and J_α in (90) is strictly positive for $\sigma^2 > 0$, so the same sign pattern holds for the proportional deterministic-equivalent W2SG improvement. □

Theorem C.21 (Proportional W2SG phase diagram)**.** *Assume $\alpha > 1, \phi_T > 1, \sigma^2 > 0, 0 < \phi_S^{-1} < 1$. Then the proportional deterministic-equivalent W2SG limit has the following behavior.*

(i) ***Subcritical and critical decay regime:*** $1 < \alpha \le 2$**.**

If $1 < \alpha < 2$, there exists a unique $\phi_{S,\gamma}^{-1}(\alpha) \in (0,1)$ such that $\gamma_S = \frac{1}{2}$. For $0 < \phi_S^{-1} \le \phi_{S,\gamma}^{-1}(\alpha)$, the coefficient

$$\frac{1-2\gamma_S}{1-\gamma_S}$$

is nonnegative, so positivity follows directly from (90). For $\phi_{S,\gamma}^{-1}(\alpha) < \phi_S^{-1} < 1$, one has $\gamma_S > 1/2$, but Theorem C.13 still gives

$$2A_1 > \frac{2\gamma_S - 1}{1-\gamma_S} s_S A_2.$$

At the critical exponent $\alpha = 2$, there is no interior $\gamma_S = 1/2$ crossing. Instead,

$$\gamma_S > \frac{1}{2} \qquad \forall\, \phi_S^{-1} \in (0,1),$$

with

$$\gamma_S \downarrow \frac{1}{2} \qquad \text{as } \phi_S^{-1} \downarrow 0.$$

Nevertheless, Theorem C.13 again yields

$$2A_1 > \frac{2\gamma_S - 1}{1-\gamma_S} s_S A_2.$$

Consequently,

$$1 < \alpha \le 2 \quad \Longrightarrow \quad \mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) > 0 \qquad \forall\, \phi_S^{-1} \in (0,1). \tag{170}$$

(ii) ***Supercritical unconditional transition regime:*** $2 < \alpha \le 1 + \sqrt{2}$.

In this regime,

$$\gamma_S > \frac{1}{2} \qquad \forall\, \phi_S^{-1} \in (0, 1),$$

and W2SG is governed by

$$2A_1 > \frac{2\gamma_S - 1}{1 - \gamma_S} s_S A_2.$$

By Proposition C.14, $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) < 0$ *for all sufficiently small* $\phi_S^{-1} > 0$. *By Corollary C.17,*

$$L_\alpha(\phi_T) > 0 \qquad \text{for every } \phi_T > 1,$$

including the endpoint $\alpha = 1 + \sqrt{2}$. *Hence Proposition C.15 gives* $\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) > 0$ *for* ϕ_S^{-1} *sufficiently close to* 1.

Finally, Theorem C.20 implies that there is a unique proportional critical ratio $\phi_{m,\mathrm{crit}} > 1$ *such that*

$$\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) \begin{cases} > 0, & 1 < \phi_S < \phi_{m,\mathrm{crit}}, \\ = 0, & \phi_S = \phi_{m,\mathrm{crit}}, \\ < 0, & \phi_S > \phi_{m,\mathrm{crit}}. \end{cases}$$

Equivalently,

$$\mathsf{W2SG}^{\mathrm{DE}}(\phi_T, \phi_S) > 0 \iff \phi_S^{-1} > \phi_{m,\mathrm{crit}}^{-1}.$$

(iii) ***Conditional supercritical regime:*** $\alpha > 1 + \sqrt{2}$.

The sparse proportional region remains negative. If $L_\alpha(\phi_T) > 0$, *then Theorem C.20 again yields a unique negative-to-positive proportional transition. If* $L_\alpha(\phi_T) \le 0$, *the present analysis proves negativity near the sparse proportional boundary but does not determine whether a positive W2SG region exists at an intermediate proportional sampling fraction.*

C.6 Sparse-growing Stage-II sample size

Assumption C.22 (Sparse-growing Stage-II DE continuation)**.** *Assume the power-law covariance model of Assumption C.7. Stage I remains in the proportional overparameterized regime,* $\frac{p}{n} \longrightarrow \phi_T > 1$, $\gamma_{T,p} \longrightarrow \gamma_T \in (0, 1)$. *Stage II follows the sparse-growing regime* $p \to \infty$, $m = m_p \to \infty$, $\frac{m_p}{p} \to 0$. *Along this sequence, we assume that the finite-p deterministic-equivalent expression (35) continues to apply, with* $\tau_{S,p} > 0$ *determined by*

$$\sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \tau_{S,p}} = m_p, \tag{171}$$

and

$$\gamma_{S,p} = \frac{1}{m_p} \sum_{i=1}^{p} \left(\frac{\lambda_i}{\lambda_i + \tau_{S,p}} \right)^2.$$

This continuation assumption is not implied by the standard proportional deterministic-equivalence theorem, because $m_p/p \to 0$.

Theorem C.23 (Sparse-growing DE W2SG limit)**.** *Under Assumption C.22,*

$$\mathsf{W2SG}_p^{\mathrm{DE}} \longrightarrow (2 - \alpha)\sigma^2 \frac{\gamma_T}{1 - \gamma_T}. \tag{172}$$

Consequently, if $\sigma^2 > 0$, *the limiting gain is positive for* $1 < \alpha < 2$, *zero for* $\alpha = 2$, *and negative for* $\alpha > 2$. *If* $\sigma^2 = 0$, *the limiting gain is zero for every* $\alpha > 1$.

Proof. Let

$$C_\alpha := \int_0^\infty \frac{du}{1+u^\alpha}, \qquad c_\alpha := C_\alpha^\alpha.$$

Rescaling (171) by $i = m_p x$ gives

$$m_p^\alpha \tau_{S,p} \longrightarrow c_\alpha. \tag{173}$$

In particular, $\tau_{S,p} \sim c_\alpha m_p^{-\alpha} \longrightarrow 0$, not ∞. The same Riemann-sum argument gives

$$\gamma_{S,p} \longrightarrow \int_0^\infty \frac{dx}{(1+c_\alpha x^\alpha)^2} = \frac{\alpha-1}{\alpha}. \tag{174}$$

The last equality follows from the identity used in Lemma C.10, after rescaling the integral to $[0,\infty)$.

For the remainder of the proof, write

$$r_{i,p} := 2\frac{\tau_{S,p}}{\lambda_i+\tau_{S,p}} - \frac{1}{1-\gamma_{S,p}}\left(\frac{\tau_{S,p}}{\lambda_i+\tau_{S,p}}\right)^2$$

and define the Stage-I noise weights

$$w_{i,p} := \frac{\sigma^2}{n(1-\gamma_{T,p})}\left(\frac{\lambda_i}{\lambda_i+\tau_{T,p}}\right)^2. \tag{175}$$

Define the limiting Stage-I noise mass by $V_T^{\mathrm{noise}} := \sigma^2 \frac{\gamma_T}{1-\gamma_T}$. Then

$$\sum_{i=1}^p w_{i,p} = \sigma^2 \frac{\gamma_{T,p}}{1-\gamma_{T,p}} \longrightarrow V_T^{\mathrm{noise}}. \tag{176}$$

Using (35), the gain can be written as

$$\mathsf{W2SG}_p^{\mathrm{DE}} = T_p^{\mathrm{teach}} + \sum_{i=1}^p r_{i,p} w_{i,p}, \tag{177}$$

where

$$T_p^{\mathrm{teach}} = -\frac{\tau_{S,p}^2}{1-\gamma_{S,p}} \sum_{i=1}^p \frac{\lambda_i^2}{(\lambda_i+\tau_{S,p})^2(\lambda_i+\tau_{T,p})}.$$

The teacher term vanishes. Indeed,

$$0 \le -T_p^{\mathrm{teach}} \le \frac{\tau_{S,p}^2}{1-\gamma_{S,p}} \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i+\tau_{S,p})^2}.$$

If $a_{i,p} = \lambda_i/(\lambda_i+\tau_{S,p})$, then the fixed-point equation implies

$$m_p(1-\gamma_{S,p}) = \tau_{S,p} \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i+\tau_{S,p})^2}.$$

Consequently,

$$0 \le -T_p^{\mathrm{teach}} \le m_p \tau_{S,p} = O(m_p^{1-\alpha}) \longrightarrow 0. \tag{178}$$

It remains to treat the noise term. The pointwise limit at a fixed index is not useful here: because $\tau_{S,p} \to 0$, it would give $\tau_{S,p}/(\lambda_i+\tau_{S,p}) \to 0$. Instead, the weights $w_{i,p}$ are asymptotically carried by indices of order p. More precisely,

$$\lim_{\delta \downarrow 0} \limsup_{p\to\infty} \sum_{i \le \delta p} w_{i,p} = 0,$$

by the Stage-I Riemann-sum limit. On the complementary set $i \ge \delta p$, (173) gives

$$\sup_{i \ge \delta p} \frac{\lambda_i}{\tau_{S,p}} \le O\left(\left(\frac{m_p}{p}\right)^\alpha\right) \longrightarrow 0.$$

Hence, uniformly for $i \geq \delta p$,

$$r_{i,p} \longrightarrow 2 - \frac{1}{1-(\alpha-1)/\alpha} = 2 - \alpha.$$

The coefficients $r_{i,p}$ are uniformly bounded for all sufficiently large p because (174) keeps $1 - \gamma_{S,p}$ bounded away from zero. Splitting the weighted sum at δp, then sending $p \to \infty$ and $\delta \downarrow 0$, yields

$$\sum_{i=1}^{p} r_{i,p} w_{i,p} \longrightarrow (2-\alpha) V_T^{\text{noise}}. \tag{179}$$

Combining (177), (178), and (179) proves (172). □

C.7 Ultra-sparse Stage-II extensions

We next consider the fixed-Stage-II-sample regime, which lies outside the standard proportional asymptotic setting.

Assumption C.24 (Fixed-m deterministic-equivalent continuation)**.** *Assume the power-law covariance model of Assumption C.7. Stage I remains in the proportional overparameterized regime,*

$$\frac{p}{n} \longrightarrow \phi_T > 1, \qquad \gamma_{T,p} \longrightarrow \gamma_T \in (0,1),$$

while the Stage II sample size $m \geq 1$ is fixed as $p \to \infty$.

Along this sequence, we assume that the finite-p deterministic-equivalent expression (35) continues to apply, with $\tau_{m,S,p} > 0$ determined by

$$\sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \tau_{m,S,p}} = m, \tag{180}$$

and

$$\gamma_{m,S,p} = \frac{1}{m} \sum_{i=1}^{p} \left(\frac{\lambda_i}{\lambda_i + \tau_{m,S,p}} \right)^2.$$

This continuation assumption is not implied by the proportional deterministic-equivalence theorem, which requires a nonvanishing Stage II sample-to-dimension ratio.

C.7.1 Fixed Stage-II sample count

Lemma C.25 (Fixed-m spectral parameters)**.** *For every fixed $m \geq 1$, there is a unique $\tau_{m,S} > 0$ satisfying*

$$\sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \tau_{m,S}} = m. \tag{181}$$

Moreover,

$$\tau_{m,S,p} \longrightarrow \tau_{m,S}, \qquad \gamma_{m,S,p} \longrightarrow \gamma_{m,S} := \frac{1}{m} \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \tau_{m,S}} \right)^2,$$

where $0 < \gamma_{m,S} < 1$. Define

$$q_{m,S} := 2 - \frac{1}{1-\gamma_{m,S}} = \frac{1-2\gamma_{m,S}}{1-\gamma_{m,S}}. \tag{182}$$

Proof. Let

$$f(\tau) := \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \tau}.$$

Because $\alpha > 1$, this series is finite for every $\tau > 0$. It is continuous and strictly decreasing, with $f(\tau) \to \infty$ as $\tau \downarrow 0$ and $f(\tau) \to 0$ as $\tau \to \infty$. Thus (181) has a unique solution. Monotone

convergence of the finite partial sums, together with uniqueness of the limiting root, gives $\tau_{m,S,p} \to \tau_{m,S}$. The convergence of $\gamma_{m,S,p}$ follows by dominated convergence.

Finally, with $a_i = \lambda_i/(\lambda_i + \tau_{m,S})$, one has $0 < a_i < 1$ and $\sum_i a_i = m$. Therefore

$$0 < \sum_i a_i^2 < \sum_i a_i = m,$$

which proves $0 < \gamma_{m,S} < 1$. □

Theorem C.26 (Fixed-m deterministic-equivalent W2SG gain)**.** *Under Assumption C.24, for every fixed integer $m \geq 1$,*

$$\mathsf{W2SG}_p^{\mathrm{DE}} \longrightarrow G_m := -m\tau_{m,S} + q_{m,S}\sigma^2 \frac{\gamma_T}{1-\gamma_T}. \tag{183}$$

Proof. Write

$$\mathsf{W2SG}_p^{\mathrm{DE}} = T_{m,p}^{\mathrm{teach}} + \sum_{i=1}^{p} r_{m,i,p} w_{i,p},$$

where

$$\begin{aligned} w_{i,p} :=& \frac{\sigma^2}{n\,(1-\gamma_{T,p})} \left(\frac{\lambda_i}{\lambda_i + \tau_{T,p}}\right)^2, \\ T_{m,p}^{\mathrm{teach}} =& -\frac{\tau_{m,S,p}^2}{1-\gamma_{m,S,p}} \sum_{i=1}^{p} \frac{\lambda_i^2}{(\lambda_i + \tau_{m,S,p})^2(\lambda_i + \tau_{T,p})}, \\ r_{m,i,p} =& 2\frac{\tau_{m,S,p}}{\lambda_i + \tau_{m,S,p}} - \frac{1}{1-\gamma_{m,S,p}} \left(\frac{\tau_{m,S,p}}{\lambda_i + \tau_{m,S,p}}\right)^2. \end{aligned}$$

Since $\tau_{T,p} \to 0$ under the proportional Stage-I scaling (68), Lemma C.25 and dominated convergence give

$$T_{m,p}^{\mathrm{teach}} \longrightarrow -\frac{\tau_{m,S}^2}{1-\gamma_{m,S}} \sum_{i=1}^{\infty} \frac{\lambda_i}{(\lambda_i + \tau_{m,S})^2}. \tag{184}$$

The fixed-point identity implies

$$m(1-\gamma_{m,S}) = \sum_{i=1}^{\infty} \left[\frac{\lambda_i}{\lambda_i + \tau_{m,S}} - \left(\frac{\lambda_i}{\lambda_i + \tau_{m,S}}\right)^2\right] = \tau_{m,S} \sum_{i=1}^{\infty} \frac{\lambda_i}{(\lambda_i + \tau_{m,S})^2}. \tag{185}$$

Substitution into (184) yields

$$T_{m,p}^{\mathrm{teach}} \longrightarrow -m\tau_{m,S}. \tag{186}$$

For the noise term, $r_{m,i,p}$ is uniformly bounded for large p, and Lemma C.25 gives

$$\lim_{K\to\infty} \limsup_{p\to\infty} \sup_{K<i\leq p} |r_{m,i,p} - q_{m,S}| = 0,$$

because $\lambda_i \to 0$. For each fixed K,

$$\sum_{i=1}^{K} w_{i,p} \longrightarrow 0, \qquad \sum_{i=1}^{p} w_{i,p} = \sigma^2 \frac{\gamma_{T,p}}{1-\gamma_{T,p}} \longrightarrow V_T^{\mathrm{noise}}.$$

Splitting the sum at K therefore yields

$$\sum_{i=1}^{p} r_{m,i,p} w_{i,p} \longrightarrow q_{m,S} V_T^{\mathrm{noise}}. \tag{187}$$

Combining (186) and (187) proves (183). □

C.7.2 Analytic sign results for the fixed-m gain

We now analyze the sign of the fixed-m limiting gain as a continuous function of the real variable $m > 0$. Throughout this subsection, assume

$$\alpha > 1, \qquad \lambda_i = i^{-\alpha}, \qquad i \in \mathbb{N} := \{1, 2, \ldots\}.$$

Let

$$0 < \gamma_T < 1, \qquad c_T := \frac{\gamma_T}{1-\gamma_T} > 0, \qquad V := c_T \sigma^2, \qquad \sigma^2 \geq 0. \tag{188}$$

For every real $m > 0$, consider the fixed-point equation

$$f_\alpha(\tau) := \sum_{i=1}^{\infty} \frac{1}{1+\tau i^\alpha} = m. \tag{189}$$

Lemma C.27 below shows that this equation has a unique solution $\tau > 0$; we denote that solution by τ_m. Set

$$a_i(m) := \frac{1}{1+\tau_m i^\alpha},$$

$$\gamma_m := \frac{1}{m}\sum_{i=1}^{\infty} a_i(m)^2, \qquad \eta_m := \frac{1}{m}\sum_{i=1}^{\infty} a_i(m)^3, \tag{190}$$

$$D_m := m(1-\gamma_m), \qquad B_m := m\tau_m, \tag{191}$$

$$q_m := \frac{1-2\gamma_m}{1-\gamma_m} = 2 - \frac{m}{D_m}, \qquad G_m := -B_m + V q_m. \tag{192}$$

For integer m, uniqueness of the fixed-point solution implies $\tau_m = \tau_{m,S}$, $\gamma_m = \gamma_{m,S}$, $q_m = q_{m,S}$. Consequently, for integer m, G_m coincides with the limiting deterministic-equivalent gain in Theorem C.26.

Fixed-point regularity and exact identities

Lemma C.27 (Existence, uniqueness, and differentiability)**.** *For every real $m > 0$, (189) has a unique solution $\tau_m > 0$. The functions used in (190)–(192) are continuously differentiable, and*

$$0 < \gamma_m < 1, \qquad D_m > 0, \qquad q_m < 1. \tag{193}$$

Moreover,

$$\tau_m' = -\frac{\tau_m}{D_m}, \tag{194}$$

$$a_i'(m) = \frac{a_i(m)(1-a_i(m))}{D_m}, \tag{195}$$

$$B_m' = -\frac{\tau_m \gamma_m}{1-\gamma_m} < 0, \tag{196}$$

$$m\gamma_m' = \frac{\gamma_m + \gamma_m^2 - 2\eta_m}{1-\gamma_m}. \tag{197}$$

Proof. For each $\tau > 0$, the defining series is bounded above by $\tau^{-1}\sum_i i^{-\alpha} < \infty$. It is continuous and strictly decreasing in τ, tends to zero as $\tau \to \infty$, and tends to infinity as $\tau \downarrow 0$. The last claim follows by retaining an arbitrarily large finite number of terms, each tending to one. These properties prove existence and uniqueness.

On any compact subinterval of $(0, \infty)$, the summands and their τ-derivatives are dominated by summable constant multiples of $i^{-\alpha}$. Termwise differentiation and the inverse function theorem therefore apply. Since $0 < a_i < 1$ and $\sum_i a_i = m$,

$$0 < \sum_i a_i^2 < \sum_i a_i = m,$$

which proves (193). Also,

$$-\tau_m f'_\alpha(\tau_m) = \sum_i a_i(1-a_i) = m - \sum_i a_i^2 = D_m.$$

Differentiating (189) yields (194); then the chain rule gives (195) and

$$B'_m = \tau_m + m\tau'_m = \tau_m - \frac{\tau_m}{1-\gamma_m},$$

which is (196). Finally,

$$\frac{\mathrm{d}}{\mathrm{d}m}\sum_i a_i^2 = \frac{2}{D_m}\sum_i a_i^2(1-a_i) = \frac{2(\gamma_m - \eta_m)}{1-\gamma_m}.$$

Differentiating $m\gamma_m = \sum_i a_i^2$ and rearranging proves (197). □

Definition C.28 (Moment expression and noise thresholds)**.** Define

$$K_m := \gamma_m^2(2-\gamma_m) - \eta_m. \tag{198}$$

On the set where $q_m > 0$, define thresholds in V and in σ^2 by

$$T_\alpha(m) := \frac{B_m}{q_m}, \qquad H_\alpha(m) := \frac{B_m}{c_T q_m} = \frac{T_\alpha(m)}{c_T}. \tag{199}$$

Proposition C.29 (Exact sign criterion and threshold derivative)**.** *For every $m \geq 1$,*

$$q_m \leq 0 \implies G_m < 0 \quad \textit{for every } \sigma^2 \geq 0, \tag{200}$$

$$q_m > 0 \implies \operatorname{sgn} G_m = \operatorname{sgn}\big(\sigma^2 - H_\alpha(m)\big). \tag{201}$$

Where $q_m > 0$, one has

$$\frac{H'_\alpha(m)}{H_\alpha(m)} = \frac{T'_\alpha(m)}{T_\alpha(m)} = \frac{2K_m}{m(1-\gamma_m)^2(1-2\gamma_m)}. \tag{202}$$

Proof. Since $B_m > 0$, (200) follows directly from (192). If $q_m > 0$, factor the gain as $G_m = c_T q_m\big(\sigma^2 - H_\alpha(m)\big)$. For the derivative identity, first rewrite (197) as

$$m\gamma'_m = \gamma_m(1-2\gamma_m) + \frac{2K_m}{1-\gamma_m}. \tag{203}$$

By (196),

$$\frac{B'_m}{B_m} = -\frac{\gamma_m}{m(1-\gamma_m)}, \qquad \frac{q'_m}{q_m} = -\frac{\gamma'_m}{(1-\gamma_m)(1-2\gamma_m)}.$$

Subtracting the latter identity from the former and using (203) proves (202). □

Integral comparisons and asymptotics Define

$$C_\alpha := \int_0^\infty \frac{\mathrm{d}x}{1+x^\alpha} = \frac{\pi}{\alpha \sin(\pi/\alpha)}. \tag{204}$$

The integral is evaluated by $u = x^\alpha$ and Euler's beta integral.

Lemma C.30 (Explicit fixed-point bounds)**.** *For every $m > 0$,*

$$\left(\frac{C_\alpha}{m+1}\right)^\alpha < \tau_m < \left(\frac{C_\alpha}{m}\right)^\alpha, \tag{205}$$

and

$$\frac{m}{\alpha} - \frac{1}{4} < D_m < \frac{m+1}{\alpha} + \frac{1}{4}. \tag{206}$$

Consequently,

$$C_\alpha^\alpha \frac{m}{(m+1)^\alpha} < B_m < C_\alpha^\alpha m^{1-\alpha}. \tag{207}$$

Proof. Let

$$g(x) := \frac{1}{1+\tau_m x^\alpha}.$$

Since g is strictly decreasing on $(0,\infty)$, the standard integral comparison is strict:

$$\sum_{i\geq 1} g(i) < \int_0^\infty g(x)\,\mathrm{d}x < g(0) + \sum_{i\geq 1} g(i).$$

Using

$$\sum_{i\geq 1} g(i) = m, \qquad g(0) = 1, \qquad \int_0^\infty g(x)\,\mathrm{d}x = C_\alpha \tau_m^{-1/\alpha},$$

we obtain

$$m < C_\alpha \tau_m^{-1/\alpha} < m+1. \tag{208}$$

Solving these inequalities for τ_m gives (205). Since $B_m = m\tau_m$, multiplication by $m > 0$ gives (207).

For the bounds on D_m, define

$$h(x) := \frac{\tau_m x^\alpha}{(1+\tau_m x^\alpha)^2}.$$

The function h increases from zero to its maximum $1/4$ and then decreases to zero. For such an absolutely continuous unimodal function,

$$\left| \sum_{i\geq 1} h(i) - \int_0^\infty h(x)\,\mathrm{d}x \right| \leq \max h = \frac{1}{4}. \tag{209}$$

To justify this bound, writing $\{x\} := x - \lfloor x \rfloor$ gives

$$\sum_{i\geq 1} h(i) - \int_0^\infty h(x)\,\mathrm{d}x = \int_0^\infty \{x\} h'(x)\,\mathrm{d}x.$$

On the increasing part, the last integral contributes between zero and $\max h$, while on the decreasing part it contributes between $-\max h$ and zero. This proves (209). Moreover, $h(x) = -\frac{x g'(x)}{\alpha}$. Hence integration by parts gives

$$\int_0^\infty h(x)\,\mathrm{d}x = \frac{1}{\alpha}\int_0^\infty g(x)\,\mathrm{d}x = \frac{C_\alpha}{\alpha}\tau_m^{-1/\alpha}.$$

Since

$$D_m = \sum_{i\geq 1} h(i),$$

equation (209) yields

$$\frac{C_\alpha}{\alpha}\tau_m^{-1/\alpha} - \frac{1}{4} \leq D_m \leq \frac{C_\alpha}{\alpha}\tau_m^{-1/\alpha} + \frac{1}{4}.$$

Combining these inequalities with the strict comparison (208),

$$\frac{m}{\alpha} < \frac{C_\alpha}{\alpha}\tau_m^{-1/\alpha} < \frac{m+1}{\alpha},$$

gives

$$\frac{m}{\alpha} - \frac{1}{4} < D_m < \frac{m+1}{\alpha} + \frac{1}{4},$$

which proves (206). □

Lemma C.31 (Moment and gain asymptotics)**.** *Fix $\alpha > 1$ and write $b = 1 - 1/\alpha$. As $m \to \infty$,*

$$\tau_m \sim C_\alpha^\alpha m^{-\alpha}, \qquad B_m \sim C_\alpha^\alpha m^{1-\alpha}, \tag{210}$$

$$\gamma_m \longrightarrow b, \qquad \eta_m \longrightarrow \frac{b(1+b)}{2}, \tag{211}$$

$$q_m \longrightarrow 2-\alpha, \qquad K_m \longrightarrow \frac{b(1-b)(2b-1)}{2}. \tag{212}$$

In particular,

$$\lim_{m\to\infty} G_m = (2-\alpha)V. \tag{213}$$

Proof. The first asymptotic in (210) follows by squeezing in (205); the second follows by multiplication by m. For the remaining moments, let $z_m = \tau_m^{-1/\alpha} \to \infty$ and

$$S_k(\alpha, z) = \sum_{i\geq 1} \frac{1}{(1+(i/z)^\alpha)^k}, \qquad C_{\alpha,k} = \int_0^\infty \frac{\mathrm{d}x}{(1+x^\alpha)^k}.$$

Integral comparison gives $0 \leq zC_{\alpha,k} - S_k(\alpha, z) \leq 1$. Integration of the derivative of $x(1+x^\alpha)^{-k}$ over $(0,\infty)$ yields

$$C_{\alpha,k+1} = \left(1 - \frac{1}{k\alpha}\right) C_{\alpha,k}.$$

Since $C_{\alpha,1} = C_\alpha$, this gives

$$C_{\alpha,2} = bC_\alpha, \qquad C_{\alpha,3} = \frac{b(1+b)}{2} C_\alpha.$$

At $z = z_m$, $S_1 = m$, $\gamma_m = S_2/S_1$, and $\eta_m = S_3/S_1$. Taking ratios proves (211). Substitution in the definitions of q_m and K_m gives (212). Finally, $B_m \to 0$, proving (213). □

The initial noise threshold

Lemma C.32 (Positivity of q_1 in the exponent range of interest)**.** *For $1 < \alpha \leq 5/2$, one has $q_1 > 0$.*

Proof. If $1 < \alpha \leq 2$, then

$$f_\alpha(1) = \sum_{i\geq 1} \frac{1}{1+i^\alpha} \geq \sum_{i\geq 1} \frac{1}{1+i^2} > \sum_{i\geq 1} \frac{1}{i(i+1)} = 1.$$

The strict inequality holds termwise for $i \geq 2$. Hence $\tau_1 > 1$. It follows that $a_i(1) \leq a_1(1) < 1/2$, so

$$\gamma_1 = \sum_i a_i(1)^2 \leq a_1(1) \sum_i a_i(1) = a_1(1) < 1/2.$$

For $2 < \alpha \leq 5/2$, we first prove $\tau_1 > 2/3$ using a finite rational lower bound. Let

$$(k_1, \ldots, k_{10}) = (100, 142, 174, 200, 224, 245, 265, 283, 300, 317).$$

Each $k_i^2 \geq 10000i$, so $\sqrt{i} \leq k_i/100$. Therefore

$$f_\alpha(2/3) \geq \sum_{i=1}^{10} \frac{1}{1+(2/3)i^{5/2}} \geq \sum_{i=1}^{10} \frac{150}{150+i^2k_i} > \frac{101}{100} > 1.$$

The penultimate inequality is an inequality between the displayed finite sum of rational numbers and $101/100$, verified by bringing the fractions to a common denominator. Thus $\tau_1 > 2/3$. Consequently, $a_1(1) < 3/5$ and $a_i(1) \leq a_2(1) < 3/11$ for $i \geq 2$. Using $\sum_i a_i(1) = 1$ gives

$$\gamma_1 \leq a_1(1)^2 + \frac{3}{11}\big(1 - a_1(1)\big) \leq \frac{129}{275} < \frac{1}{2}.$$

Here the quadratic is bounded by its maximum at the endpoints of $[0, 3/5]$. In both cases $\gamma_1 < 1/2$, hence $q_1 > 0$. □

Definition C.33 (Initial critical noise)**.** For $1 < \alpha \leq 5/2$, define

$$\sigma_c^2 := H_\alpha(1) = \frac{\tau_1}{c_T q_1} > 0. \tag{214}$$

This is the noise level at which $G_1 = 0$. It is not, by definition, a threshold guaranteeing the same sign for every m.

Corollary C.34. *For $1 < \alpha \leq 5/2$,*

$$G_1 \begin{cases} < 0, & 0 \leq \sigma^2 < \sigma_c^2, \\ = 0, & \sigma^2 = \sigma_c^2, \\ > 0, & \sigma^2 > \sigma_c^2. \end{cases}$$

For $\sigma^2 < \sigma_c^2$, continuity gives $G_m < 0$ on some real interval $[1, 1+\delta)$, with $\delta > 0$.

Proof. Use $G_1 = c_T q_1 (\sigma^2 - \sigma_c^2)$ and $c_T q_1 > 0$. □

The regime $1 < \alpha < 2$

Theorem C.35 (Negative initial gain and positive tail)**.** *Suppose* $1 < \alpha < 2$ *and* $0 < \sigma^2 < \sigma_c^2$. *Then* $G_m < 0$ *near* $m = 1$, *and* $G_m > 0$ *for every sufficiently large* m. *In particular, the continuous extension has at least one zero in* $(1, \infty)$.

Proof. The initial sign follows from Corollary C.34. Since $V > 0$ and $2 - \alpha > 0$, Lemma C.31 gives $G_m \to (2-\alpha)V > 0$. The intermediate value theorem proves the existence of a zero. □

Remark C.36. Theorem C.35 does not prove uniqueness of the zero. It does not exclude additional changes of sign at intermediate m. Positive noise is essential: for $\sigma^2 = 0$, $G_m = -B_m < 0$ for every m.

Corollary C.37 (An explicit sufficient positive region)**.** *Suppose* $1 < \alpha < 2$ *and* $V > 0$. *If*

$$m \geq \frac{\alpha(\alpha+2)}{4(2-\alpha)} \quad \textit{and} \quad m > \left[\frac{2C_\alpha^\alpha}{(2-\alpha)V}\right]^{1/(\alpha-1)}, \tag{215}$$

then $G_m > 0$.

Proof. The first inequality implies $m > \alpha/4$. By (206),

$$q_m = 2 - \frac{m}{D_m} \geq 2 - \frac{\alpha m}{m - \alpha/4} \geq \frac{2-\alpha}{2}.$$

The last inequality is algebraically equivalent to the first condition in (215). Using (207),

$$G_m \geq -C_\alpha^\alpha m^{1-\alpha} + \frac{2-\alpha}{2} V > 0$$

under the second condition. □

Theorem C.38 (Unique transition for sufficiently small positive noise)**.** *Fix* $1 < \alpha < 2$ *and* $c_T > 0$. *There exists* $0 < \sigma_0^2 < \sigma_c^2$ *such that, for every* $0 < \sigma^2 \leq \sigma_0^2$, *there is a unique real* $m_* > 1$ *with*

$$G_m \begin{cases} < 0, & 1 \leq m < m_*, \\ = 0, & m = m_*, \\ > 0, & m > m_*. \end{cases} \tag{216}$$

A valid choice is

$$\sigma_0^2 = \frac{B_M}{c_T}, \tag{217}$$

where $M \geq 1$ *is any fixed number such that* $q_m > 0$ *and* $K_m < 0$ *for all* $m \geq M$. *Such an* M *exists and depends only on* α.

Proof. Let $b = 1 - 1/\alpha \in (0, 1/2)$. By Lemma C.31,

$$q_m \to 2 - \alpha > 0, \qquad K_m \to \frac{b(1-b)(2b-1)}{2} < 0.$$

Thus M with the stated properties exists. On $[M, \infty)$, Proposition C.29 implies $T'_\alpha(m) < 0$. Also,

$$T_\alpha(m) \sim \frac{C_\alpha^\alpha}{2-\alpha} m^{1-\alpha} \longrightarrow 0. \tag{218}$$

Now take $0 < V \leq B_M$. Because B_m is decreasing and $q_m < 1$, for $1 \leq m \leq M$ one has

$$G_m = -B_m + V q_m < -B_m + V \leq -B_M + V \leq 0.$$

At M, $0 < q_M < 1$ gives $T_\alpha(M) > B_M \geq V$. The continuous strictly decreasing function T_α therefore takes the value V exactly once on (M, ∞). Factoring $G_m = q_m[V - T_\alpha(m)]$ on this tail proves (216). Finally, since $B_M \leq B_1$ and $0 < q_1 < 1$,

$$\sigma_0^2 = \frac{B_M}{c_T} \leq \frac{B_1}{c_T} < \frac{B_1}{c_T q_1} = \sigma_c^2.$$

□

Corollary C.39 (Small-noise location of the unique transition)**.** *Under Theorem C.38, with α and c_T fixed,*

$$m_* \sim \left[\frac{C_\alpha^\alpha}{(2-\alpha)c_T\sigma^2}\right]^{1/(\alpha-1)} \qquad \text{as } \sigma^2 \downarrow 0. \tag{219}$$

Proof. The equation $T_\alpha(m_*) = V$ and strict positivity of T_α at every finite point on $[M, \infty)$ imply $m_* \to \infty$ as $V \downarrow 0$. Substitute $m = m_*$ into (218) and solve the resulting asymptotic relation for m_*. □

The critical fixed-m regime at $\alpha = 2$ The case $\alpha = 2$ is critical because the leading large-m limit $G_m \longrightarrow (2-\alpha)V$ vanishes at $\alpha = 2$. Consequently, this leading-order limit does not determine the eventual sign of G_m, and a higher-order analysis is required. At $\alpha = 2$, we can carry out this analysis explicitly using closed-form summation identities.

Here, "fixed-m" refers to the order of limits: we first take $p \to \infty$ with m fixed and then analyze the resulting limiting family as $m \to \infty$. At $\alpha = 2$, retain the definitions $G_m = -B_m + Vq_m,\ B_m = m\tau_m,\ q_m = \frac{1-2\gamma_m}{1-\gamma_m}$, where $\tau_m > 0$ is determined by

$$\sum_{i=1}^{\infty} \frac{1}{1+\tau_m i^2} = m. \tag{220}$$

Lemma C.40 (Exact hyperbolic representation at $\alpha = 2$)**.** *Let $\alpha = 2$ and $m > 0$. Define $z_m := \frac{\pi}{\sqrt{\tau_m}}$, $s_m := 2m+1$. Then $z_m > 0$ is the unique solution of*

$$z_m \coth z_m = s_m. \tag{221}$$

Moreover, defining

$$\delta_m := s_m^2 - z_m^2 = z_m^2 \operatorname{csch}^2 z_m, \tag{222}$$

one has $0 < \delta_m < 1$, and

$$D_m = m(1-\gamma_m) = \frac{s_m - \delta_m}{4}, \tag{223}$$

$$q_m = \frac{2(1-\delta_m)}{s_m - \delta_m}, \tag{224}$$

$$B_m = \frac{\pi^2(s_m - 1)}{2(s_m^2 - \delta_m)}. \tag{225}$$

In particular, $q_m > 0$ for every $m > 0$.

Proof. For $a > 0$, the classical identity

$$\sum_{i=1}^{\infty} \frac{1}{i^2 + a^2} = \frac{\pi a \coth(\pi a) - 1}{2a^2} \tag{226}$$

implies, with $a = z_m/\pi$,

$$\sum_{i=1}^{\infty} \frac{1}{1+\tau_m i^2} = \frac{z_m^2}{\pi^2} \sum_{i=1}^{\infty} \frac{1}{i^2 + (z_m/\pi)^2} = \frac{z_m \coth z_m - 1}{2}. \tag{227}$$

Using (220) therefore yields $z_m \coth z_m = 2m + 1 = s_m$, which proves (221).

To establish uniqueness, define $f(z) := z \coth z,\ z > 0$. Then

$$f'(z) = \coth z - z \operatorname{csch}^2 z = \frac{\sinh z \cosh z - z}{\sinh^2 z} > 0, \tag{228}$$

because $\sinh(2z) > 2z$ for every $z > 0$. Furthermore,

$$\lim_{z \downarrow 0} f(z) = 1, \qquad \lim_{z \to \infty} f(z) = \infty.$$

Since $s_m = 2m+1 > 1$, equation (221) therefore has a unique positive solution. Next, since $\coth^2 z - 1 = \operatorname{csch}^2 z$, we obtain $s_m^2 - z_m^2 = z_m^2 \operatorname{csch}^2 z_m = \delta_m$. Moreover, $\sinh z > z$ for every $z > 0$, so

$$0 < z^2 \operatorname{csch}^2 z = \frac{z^2}{\sinh^2 z} < 1.$$

Hence $0 < \delta_m < 1$. To obtain D_m, define $M(\tau) := \sum_{i=1}^{\infty} \frac{1}{1+\tau i^2}$. Then

$$D_m = \sum_{i=1}^{\infty} a_i(1-a_i) = -\tau_m M'(\tau_m), \qquad a_i := \frac{1}{1+\tau_m i^2}.$$

Since $M(\tau) = \dfrac{z\coth z - 1}{2}$, $z = \pi\tau^{-1/2}$, and $\dfrac{dz}{d\tau} = -\dfrac{z}{2\tau}$, we obtain

$$-\tau M'(\tau) = \frac{z}{4}\left(\coth z - z\operatorname{csch}^2 z\right). \tag{229}$$

Evaluating at $z = z_m$ gives

$$D_m = \frac{z_m \coth z_m - z_m^2 \operatorname{csch}^2 z_m}{4} = \frac{s_m - \delta_m}{4}.$$

Since $q_m = 2 - \frac{m}{D_m}$, $m = \frac{s_m - 1}{2}$, we obtain

$$q_m = 2 - \frac{2(s_m-1)}{s_m - \delta_m} = \frac{2(1-\delta_m)}{s_m - \delta_m}.$$

Because $0 < \delta_m < 1$, this also proves $q_m > 0$. Finally,

$$\tau_m = \frac{\pi^2}{z_m^2} = \frac{\pi^2}{s_m^2 - \delta_m},$$

and therefore

$$B_m = m\tau_m = \frac{\pi^2(s_m-1)}{2(s_m^2 - \delta_m)}.$$

□

Lemma C.41 (Second-order large-m expansion at $\alpha = 2$)**.** *Under the assumptions of Lemma C.40, as $m \to \infty$,*

$$q_m = \frac{1}{m} - \frac{1}{2m^2} + O(m^{-3}), \tag{230}$$

$$B_m = \frac{\pi^2}{4m} - \frac{\pi^2}{4m^2} + O(m^{-3}), \tag{231}$$

and consequently

$$G_m = \frac{V - \pi^2/4}{m} + \frac{\pi^2/4 - V/2}{m^2} + O(m^{-3}). \tag{232}$$

Proof. From (221),

$$s_m - z_m = z_m(\coth z_m - 1) = \frac{2z_m}{e^{2z_m} - 1}.$$

Since $s_m \to \infty$ and $z \mapsto z\coth z$ is strictly increasing and unbounded, we have $z_m \to \infty$. Hence $s_m - z_m \to 0$, so $z_m \sim s_m \sim 2m$. It follows that $s_m - z_m = O\big(me^{-4m}\big)$, and therefore $\delta_m = (s_m - z_m)(s_m + z_m) = O\big(m^2 e^{-4m}\big)$. In particular, $\delta_m = o(m^{-k})$ for every fixed $k > 0$. Using (224),

$$q_m = \frac{2(1-\delta_m)}{2m+1-\delta_m} = \frac{2}{2m+1} + o(m^{-3}). \tag{233}$$

Expanding $(2m+1)^{-1}$ gives

$$q_m = \frac{1}{m} - \frac{1}{2m^2} + O(m^{-3}).$$

Similarly, from (225),

$$B_m = \frac{\pi^2(2m)}{2\big((2m+1)^2 - \delta_m\big)} = \frac{\pi^2 m}{(2m+1)^2} + o(m^{-3}) = \frac{\pi^2}{4m} - \frac{\pi^2}{4m^2} + O(m^{-3}). \tag{234}$$

Finally, substituting these expansions into $G_m = -B_m + V q_m$ gives

$$G_m = \frac{V - \pi^2/4}{m} + \frac{\pi^2/4 - V/2}{m^2} + O(m^{-3}),$$

as claimed. □

Corollary C.42 (Critical tail-noise threshold at $\alpha = 2$)**.** *Let $V = \sigma^2 \frac{\gamma_T}{1-\gamma_T}$, and define*

$$V_c := \frac{\pi^2}{4}, \qquad \sigma^2_{\mathrm{tail},c} := \frac{\pi^2}{4}\frac{1-\gamma_T}{\gamma_T}. \tag{235}$$

Then:

1. *if $\sigma^2 < \sigma^2_{\mathrm{tail},c}$, then $G_m < 0$ for all sufficiently large m;*
2. *if $\sigma^2 > \sigma^2_{\mathrm{tail},c}$, then $G_m > 0$ for all sufficiently large m;*
3. *at the critical value $\sigma^2 = \sigma^2_{\mathrm{tail},c}$, the first-order term vanishes and*

$$G_m = \frac{\pi^2}{8m^2} + O(m^{-3}). \tag{236}$$

Consequently, $G_m > 0$ for all sufficiently large m.

Proof. The first two claims follow immediately from the coefficient of m^{-1} in (232). At the critical value $V = \frac{\pi^2}{4}$, the coefficient of m^{-1} vanishes, while the coefficient of m^{-2} becomes

$$\frac{\pi^2}{4} - \frac{1}{2}\frac{\pi^2}{4} = \frac{\pi^2}{8} > 0.$$

Hence

$$G_m = \frac{\pi^2}{8m^2} + O(m^{-3}),$$

and therefore $G_m > 0$ for all sufficiently large m. □

The preceding results show why $\alpha = 2$ is a genuine critical exponent. For every fixed positive proportional Stage-II sampling fraction and $\sigma^2 > 0$, the proportional deterministic-equivalent gain remains strictly positive. By contrast, in the sparse-growing regime the order-one limit vanishes at $\alpha = 2$, while in the fixed-m continuation the large-m behavior is controlled by the critical tail-noise level $\sigma^2_{\mathrm{tail},c} = \frac{\pi^2}{4}\frac{1-\gamma_T}{\gamma_T}$. Thus, at the critical exponent the sign is determined by a higher-order balance that is invisible in the leading sparse-growing limit.

The regime $2 < \alpha \le 5/2$ For $\alpha > 2$, the large-m behavior is qualitatively different from the case $1 < \alpha < 2$. Indeed, since

$$q_m \longrightarrow 2 - \alpha < 0,$$

the gain is eventually negative independently of the noise level. The next lemma gives an explicit finite point beyond which this occurs.

Lemma C.43 (Noise-independent negative tail)**.** *For every $\alpha > 2$, define*

$$M_\alpha := \frac{\alpha + 4}{2(\alpha - 2)} = \frac{2 + \alpha/2}{\alpha - 2}. \tag{237}$$

Then, for every real $m \ge M_\alpha$,

$$q_m \le 0, \qquad G_m < 0 \quad \text{for every } \sigma^2 \ge 0. \tag{238}$$

Proof. From (206),

$$D_m \le \frac{m+1}{\alpha} + \frac{1}{4} = \frac{4m+\alpha+4}{4\alpha}.$$

If $m \ge M_\alpha$, then

$$\frac{4m+\alpha+4}{4\alpha} \le \frac{m}{2},$$

and therefore $D_m \le m/2$. Hence

$$q_m = 2 - \frac{m}{D_m} \le 0.$$

The conclusion $G_m < 0$ now follows from (200). □

We next obtain an explicit lower bound for the threshold $T_\alpha(m)$ at the integer points for which $q_m > 0$.

Lemma C.44 (Explicit integer threshold bound)**.** *Suppose* $2 < \alpha \le 5/2$. *For every integer* $m \ge 2$ *such that* $q_m > 0$,

$$T_\alpha(m) > C_\alpha^\alpha \frac{\alpha+12}{3^{\alpha+1}(4-\alpha)}. \tag{239}$$

Proof. By Lemma C.43, $q_m > 0$ implies $m < M_\alpha$. Moreover, the upper bound on D_m gives

$$q_m = 2 - \frac{m}{D_m} \le 2 - \frac{4\alpha m}{4m+\alpha+4} =: \overline{q}_\alpha(m).$$

Thus

$$0 < q_m \le \overline{q}_\alpha(m), \qquad \overline{q}_\alpha(m) = \frac{2(\alpha+4) - 4(\alpha-2)m}{4m+\alpha+4}. \tag{240}$$

By the strict lower bound in (207), $B_m > C_\alpha^\alpha \frac{m}{(m+1)^\alpha}$. Using also $0 < q_m \le \overline{q}_\alpha(m)$, we obtain

$$T_\alpha(m) = \frac{B_m}{q_m} > C_\alpha^\alpha \frac{m}{(m+1)^\alpha} \frac{1}{\overline{q}_\alpha(m)} =: L_\alpha(m),$$

where

$$L_\alpha(m) = \frac{C_\alpha^\alpha}{4(\alpha-2)} \frac{m(4m+\alpha+4)}{(m+1)^\alpha (M_\alpha - m)}. \tag{241}$$

To locate the smallest value relevant to integer $m \ge 2$, define $m_{0,\alpha} := \frac{\alpha+4}{4(\alpha-1)}$. For $2 < \alpha \le 5/2$, $1 < m_{0,\alpha} < \frac{3}{2} < 2$. Furthermore,

$$\frac{L'_\alpha(m)}{L_\alpha(m)} = \frac{1}{m} + \frac{4}{4m+\alpha+4} + \frac{1}{M_\alpha - m} - \frac{\alpha}{m+1}. \tag{242}$$

Writing $d := \alpha - 2 \in (0, 1/2], t := m - m_{0,\alpha} \ge 0$, and putting (242) over its positive common denominator, its numerator is, up to a positive factor,

$$P_d(t) = 64d(d+1)^3t^3 + 16d(d+1)^2(d+2)(d+7)t^2 + 4(2-d)(d+1)(d+2)(d+6)t + (3-d)(d+2)^2(d+6)^2.$$

Every coefficient is strictly positive for $0 < d \le 1/2$. Hence $L'_\alpha(m) > 0$ for $m_{0,\alpha} < m < M_\alpha$. Since every integer under consideration satisfies $2 \le m < M_\alpha, 2 > m_{0,\alpha}$, it follows that $L_\alpha(m) \ge L_\alpha(2)$.

Finally,

$$\overline{q}_\alpha(2) = \frac{6(4-\alpha)}{\alpha+12},$$

so

$$L_\alpha(2) = C_\alpha^\alpha \frac{2}{3^\alpha} \frac{\alpha+12}{6(4-\alpha)} = C_\alpha^\alpha \frac{\alpha+12}{3^{\alpha+1}(4-\alpha)}.$$

Therefore

$$T_\alpha(m) > L_\alpha(m) \ge L_\alpha(2) = C_\alpha^\alpha \frac{\alpha+12}{3^{\alpha+1}(4-\alpha)},$$

which proves (239). □

Theorem C.45 (Explicit no-W2SG bound for integer sample counts)**.** *Suppose* $2 < \alpha \le 5/2$ *and* $c_T > 0$*. Define*

$$\sigma^2_{\text{safe}} := \min\left\{\sigma_c^2,\ \frac{C_\alpha^\alpha}{c_T}\frac{\alpha+12}{3^{\alpha+1}(4-\alpha)}\right\}. \tag{243}$$

Then

$$0 \le \sigma^2 \le \sigma^2_{\text{safe}} \quad \Longrightarrow \quad G_m \le 0 \quad \textit{for every } m \in \mathbb{N}. \tag{244}$$

More precisely,

$$G_m < 0 \qquad \textit{for every integer } m \ge 2 \tag{245}$$

throughout the whole range $0 \le \sigma^2 \le \sigma^2_{\text{safe}}$*. Furthermore,*

$$0 \le \sigma^2 < \sigma^2_{\text{safe}} \quad \Longrightarrow \quad G_m < 0 \quad \textit{for every } m \in \mathbb{N}. \tag{246}$$

At the endpoint $\sigma^2 = \sigma^2_{\text{safe}}$*, equality* $G_m = 0$ *can occur only at* $m = 1$*, and only when* $\sigma^2_{\text{safe}} = \sigma_c^2$*.*

Proof. First consider $m = 1$. Since $\sigma^2_{\text{safe}} \le \sigma_c^2$, the initial sign characterization gives

$$G_1 \le 0 \qquad \text{whenever } 0 \le \sigma^2 \le \sigma^2_{\text{safe}}.$$

Moreover, $G_1 < 0$ if $\sigma^2 < \sigma^2_{\text{safe}}$, while at the endpoint $\sigma^2 = \sigma^2_{\text{safe}}$ equality can occur only if $\sigma^2_{\text{safe}} = \sigma_c^2$. Now let $m \ge 2$ be an integer. If $q_m \le 0$, then $G_m < 0$ for every $\sigma^2 \ge 0$ by (200). It remains to consider $q_m > 0$. By Lemma C.44,

$$H_\alpha(m) = \frac{T_\alpha(m)}{c_T} > \frac{C_\alpha^\alpha}{c_T}\frac{\alpha+12}{3^{\alpha+1}(4-\alpha)}.$$

Since

$$\sigma^2_{\text{safe}} \le \frac{C_\alpha^\alpha}{c_T}\frac{\alpha+12}{3^{\alpha+1}(4-\alpha)},$$

we obtain the strict inequality $\sigma^2_{\text{safe}} < H_\alpha(m)$. Hence, for every $0 \le \sigma^2 \le \sigma^2_{\text{safe}}$, $\sigma^2 < H_\alpha(m)$. The pointwise sign criterion (201) therefore yields $G_m < 0$. Thus $G_m < 0$ for every integer $m \ge 2$, while $G_1 \le 0$. This proves (244) and (245). For $\sigma^2 < \sigma^2_{\text{safe}}$, one also has $G_1 < 0$, which proves (246). The final endpoint statement follows from the defining property of σ_c^2 at $m = 1$. □

Remark C.46. The bound σ^2_{safe} is an explicit sufficient noise threshold that rules out positive fixed-m W2SG gain for every integer Stage II sample count. At the endpoint $\sigma^2 = \sigma^2_{\text{safe}}$, the gain remains strictly negative for every $m \ge 2$; the only possible zero is at $m = 1$, when $\sigma^2_{\text{safe}} = \sigma_c^2$. No claim is made that σ^2_{safe} is the largest possible threshold with this property.

D Two-Stage Random-Feature Interpolation Theory

This section develops the two-stage random-feature theory for Case II. The teacher and student operate through independent random linear feature maps, so the Stage-II predictor is constrained simultaneously by the pseudo-labeled sample size and by the dimension of the student feature space. The resulting weak-to-strong generalization behavior is therefore governed not only by the quality of the Stage-I teacher, but also by the interaction among the signal dimension, the teacher width, the number of Stage-II samples, and the student width.

The analysis proceeds in two layers. We first derive a master deterministic-equivalent description of the two-stage procedure. After recording the two-point structured-Wishart deterministic equivalent needed for second-order risk calculations, we formulate the random-feature model and analyze Stage I through its exact projection geometry and deterministic-equivalent strong- and weak-block energies. We then condition on the Stage-I teacher and derive the Stage-II random-feature regression problem. Averaging successively over the Stage-II samples and the student features yields the corresponding effective-resolution equations, mimic and cross terms, and ultimately a master deterministic equivalent for the teacher-minus-student prediction-risk gap.

We next specialize this master expression to the positive-signal- fraction regime. The resulting phase diagram separates the sample-bottleneck, feature-bottleneck, and no-bottleneck branches and identifies the parameter regions in which the deterministic-equivalent weak-to-strong gain is positive or negative. In particular, the analysis isolates the role of the Stage-I mismatch through a single

effective quantity and shows how the location of the active Stage-II bottleneck relative to the signal dimension determines the form of the phase boundary.

Finally, we examine the principal regimes excluded from the regular fixed-ratio phase diagram. We give an exact finite-dimensional classification of the Stage-II interpolation boundary, including its integrability threshold and its termination at the ambient no-bottleneck boundary. We then describe the $\sqrt{p}$ critical window around the signal dimension. Within the master deterministic equivalent, this analysis also reveals an asymmetry under joint teacher–student criticality: the sample-bottleneck branch remains unfavorable, whereas a positive weak-to-strong region may survive on the feature-bottleneck branch.

D.1 Deterministic equivalence

In addition to Lemma C.1 from Appendix C, the derivation of the two-random-feature ridgeless-regression results requires the following two-point deterministic-equivalence lemma, introduced in Atanasov et al. (2025).

Lemma D.1 (Two-point deterministic equivalence for structured Wishart matrices)**.** *Let $A_p \succeq 0$ be deterministic with uniformly bounded operator norm, let $Z_p \in \mathbb{R}^{n\times p}$ have independent standard Gaussian entries, and define*

$$q_p := \frac{p}{n}, \qquad W_p := \frac{1}{n} Z_p^\top Z_p, \qquad \widehat{A}_p := A_p^{1/2} W_p A_p^{1/2}.$$

Assume that $q_p \to q \in (0,\infty)$. For $\lambda, \lambda' > 0$, let $\kappa_{p,\lambda}, \kappa_{p,\lambda'} > 0$ denote the renormalized ridges determined by

$$\kappa_{p,\lambda}\left[1 - q_p \operatorname{tr}\left(A_p(A_p + \kappa_{p,\lambda} I_p)^{-1}\right)\right] = \lambda,$$

and the analogous equation with λ'. Define

$$\widehat{T}_{p,\lambda} := \widehat{A}_p(\widehat{A}_p + \lambda I_p)^{-1}, \qquad G_{p,\lambda} := (A_p + \kappa_{p,\lambda} I_p)^{-1}, \qquad T_{p,\lambda} := A_p G_{p,\lambda},$$

with $G_{p,\lambda'}$ and $T_{p,\lambda'}$ defined similarly. Then, for every deterministic matrix M_p satisfying $\sup_p \|M_p\|_{\mathrm{op}} < \infty$, the two-point deterministic equivalent of (Atanasov et al., 2025, Eq. (A.3) and the subsequent free-product specialization) specializes to

$$\widehat{T}_{p,\lambda} M_p \widehat{T}_{p,\lambda'} \simeq T_{p,\lambda} M_p T_{p,\lambda'} + \kappa_{p,\lambda}\kappa_{p,\lambda'} G_{p,\lambda} A_p G_{p,\lambda'} \frac{q_p \operatorname{tr}\left[A_p G_{p,\lambda} M_p G_{p,\lambda'}\right]}{1 - q_p \operatorname{tr}\left[A_p^2 G_{p,\lambda} G_{p,\lambda'}\right]}. \tag{247}$$

Here $\simeq$ has the strong test-matrix meaning of Lemma C.1, and the denominator is assumed to remain bounded away from zero.

In particular, for $\lambda' = \lambda$,

$$\widehat{T}_{p,\lambda} M_p \widehat{T}_{p,\lambda} \simeq T_{p,\lambda} M_p T_{p,\lambda} + \kappa_{p,\lambda}^2 G_{p,\lambda} A_p G_{p,\lambda} \frac{q_p \operatorname{tr}\left[A_p G_{p,\lambda} M_p G_{p,\lambda}\right]}{1 - q_p \operatorname{tr}\left[A_p^2 G_{p,\lambda}^2\right]}. \tag{248}$$

Since $0 \preceq \widehat{T}_{p,\lambda} \preceq I_p$, the same normalized-trace deterministic equivalent remains valid after taking expectation over W_p.

D.2 Defining Random-Feature Model

In the second setting, we consider a linear random-feature regression model in which the teacher and student operate through random feature representations of possibly different dimensions. Unlike Case I, where the teacher and student share the same hypothesis space and the difference arises only from the two-stage training procedure, the weak-to-strong generalization phenomenon in this setting is affected by the interaction between the random feature representations and the Stage-II training procedure.

Let the input distribution in the preliminaries be $x \sim P_x = \mathcal{N}(0, \Sigma_p)$, where

$$\Sigma_p = \operatorname{diag}\left(\underbrace{1,\ldots,1}_{k}, \underbrace{\delta_p,\ldots,\delta_p}_{p-k}\right), \qquad 0 < \delta_p < 1.$$

Let $\beta^\star \in \mathbb{R}^p$ be supported on the first k coordinates, so $\beta_i^\star = 0$ for $i > k$.

The target function and noiseless observation model are $\psi^\star(x) = x^\top \beta^\star$ and $y = \psi^\star(x)$. Thus, for the labeled dataset $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ introduced in the preliminaries, $y_i = x_i^\top \beta^\star$. We normalize the target so that $\mathbb{E}[y^2] = (\beta^\star)^\top \Sigma_p \beta^\star = 1$. The teacher and student feature extractors are random linear maps, which are defined through independent random matrices U_T and U_S, respectively. Case II studies its population-limit analogue: conditional on U_T, the teacher minimizes population risk. Consequently, n does not enter the asymptotic analysis below.

Stage I (teacher). Let

$$U_T := [u_{T,1}| \dots |u_{T,N_T}]^\top \in \mathbb{R}^{N_T \times p},$$

where $u_{T,1}, \dots, u_{T,N_T} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, I_p)$. The matrix U_T is randomly drawn and is not learnable. Conditional on U_T, the teacher feature extractor is $f_T(x) = U_T x \in \mathbb{R}^{N_T}$, and the conditional teacher hypothesis class is

$$\mathcal{H}_T(U_T) := \left\{ h_{T,w} : x \mapsto \langle w, f_T(x) \rangle = x^\top U_T^\top w : w \in \mathbb{R}^{N_T} \right\}.$$

The population version of the teacher learning rule $\mathcal{A}_T$ returns the minimum-Euclidean-norm population-risk minimizer:

$$w_T^\star = \operatorname*{argmin}_{w \in \mathbb{R}^{N_T}} \|w\|_2 \qquad \text{subject to} \qquad w \in \operatorname*{argmin}_{v \in \mathbb{R}^{N_T}} \mathbb{E}_x \left[\left(x^\top U_T^\top v - x^\top \beta^\star \right)^2 \right].$$

Define the induced coefficient in the original input space by $\widehat{\beta}_T = U_T^\top w_T^\star$. The resulting teacher predictor is $\widehat{h}_T = h_{T,w_T^\star}$, with $\widehat{h}_T(x) = x^\top \widehat{\beta}_T$.

Stage II (student). Independently of U_T, generate fresh feature vectors $\widetilde{x}_1, \dots, \widetilde{x}_m \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_p)$, and assign pseudo-labels using the teacher predictor:

$$\widetilde{y}_i \triangleq \widehat{h}_T(\widetilde{x}_i) = \widetilde{x}_i^\top \widehat{\beta}_T, \qquad i \in [m].$$

The resulting Stage-II dataset is the pseudo-labeled sample from the preliminaries, $\widetilde{\mathcal{D}}_m = \{(\widetilde{x}_i, \widetilde{y}_i)\}_{i \in [m]}$. Define the corresponding input matrix by $\widetilde{X} \triangleq [\widetilde{x}_1| \dots |\widetilde{x}_m]^\top \in \mathbb{R}^{m \times p}$ and label vector $\widetilde{y} \triangleq [\widetilde{y}_1, \dots, \widetilde{y}_m]^\top \in \mathbb{R}^m$.

The student uses an independent random feature matrix

$$U_S \triangleq [u_{S,1}| \dots |u_{S,N_S}]^\top \in \mathbb{R}^{N_S \times p},$$

where $u_{S,1}, \dots, u_{S,N_S} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, I_p)$. The matrix U_S is independent of U_T and the Stage-II samples. Conditional on U_S, the student feature extractor is $f_S(x) = U_S x \in \mathbb{R}^{N_S}$. The conditional student hypothesis class is

$$\mathcal{H}_S(U_S) := \left\{ h_{S,w} : x \mapsto \langle w, f_S(x) \rangle = x^\top U_S^\top w : w \in \mathbb{R}^{N_S} \right\}.$$

The Stage-II training rule $\mathcal{A}_S$ corresponds to ridgeless least-squares regression in the random-feature space. We take the minimum-norm least-squares solution $\widehat{w}_S = \left(\widetilde{X} U_S^\top \right)^\dagger \widetilde{y}$. Equivalently,

$$\widehat{w}_S = \operatorname*{argmin}_{w \in \mathbb{R}^{N_S}} \|w\|_2 \qquad \text{subject to} \qquad w \in \operatorname*{argmin}_{v \in \mathbb{R}^{N_S}} \left\| \widetilde{X} U_S^\top v - \widetilde{y} \right\|_2^2.$$

No direct interpolation constraint is imposed here, since $\widetilde{y}$ need not lie in $\mathrm{range}(\widetilde{X} U_S^\top)$. The induced coefficient and student predictor are $\widehat{\beta}_S = U_S^\top \widehat{w}_S$, $\widehat{h}_S = h_{S,\widehat{w}_S}$, and $\widehat{h}_S(x) = x^\top \widehat{\beta}_S$. For a linear predictor $h_\beta(x) := x^\top \beta$, the squared-loss population risk in (2) specializes to

$$\mathcal{R}(h_\beta) = \mathbb{E}_x \left[(x^\top \beta - x^\top \beta^\star)^2 \right] = (\beta - \beta^\star)^\top \Sigma_p (\beta - \beta^\star).$$

Accordingly, in the notation of the preliminaries, $\overline{\mathcal{R}}_{T,p} = \mathbb{E}\left[\mathcal{R}(\widehat{h}_T) \right]$ and $\overline{\mathcal{R}}_{S,p} = \mathbb{E}\left[\mathcal{R}(\widehat{h}_S) \right]$. The expectation in $\overline{\mathcal{R}}_{T,p}$ is over U_T, whereas the expectation in $\overline{\mathcal{R}}_{S,p}$ is over U_T, U_S, and $\widetilde{\mathcal{D}}_m$. The finite-dimensional W2SG gain is

$$\mathsf{W2SG}_p := \overline{\mathcal{R}}_{T,p} - \overline{\mathcal{R}}_{S,p}. \tag{249}$$

Thus, W2SG occurs whenever $\mathsf{W2SG}_p > 0$.

D.3 THEORY OF DE FOR STAGE I

Lemma D.2 (Exact Stage-I estimator and projection geometry)**.** *Let*

$$A_{T,p} := U_T \Sigma_p^{1/2}, \qquad P_{T,p} := A_{T,p}^\top (A_{T,p} A_{T,p}^\top)^\dagger A_{T,p}.$$

The minimum-norm population estimator and its predictor are

$$\widehat{\beta}_T = U_T^\top (U_T \Sigma_p U_T^\top)^\dagger U_T \Sigma_p \beta^\star, \qquad \widehat{h}_T(x) := x^\top \widehat{\beta}_T.$$

In whitened coordinates,

$$\theta_T := \Sigma_p^{1/2} \widehat{\beta}_T = P_{T,p} \theta^\star, \qquad \theta^\star := \Sigma_p^{1/2} \beta^\star,$$

where $P_{T,p}$ is the Euclidean orthogonal projector onto $\mathrm{row}(A_{T,p})$*. Consequently, $\overline{\mathcal{R}}_{T,p} = 1 - \mathbb{E}\|\theta_T\|_2^2 = 1 - S_{T,p} - V_{T,p}$, where $S_{T,p} := \mathbb{E}\|\theta_{T,\mathrm{str}}\|_2^2, V_{T,p} := \mathbb{E}\|\theta_{T,\mathrm{wk}}\|_2^2$. Moreover,*

$$\mathbb{E}[(\beta_{\mathrm{str}}^\star)^\top \theta_{T,\mathrm{str}}] = S_{T,p} + V_{T,p}.$$

Proof. **Exact Stage-I population estimator.**

For a candidate coefficient $\beta = U_T^\top w$, the population squared loss is

$$\mathcal{L}_T(w) = \mathbb{E}_x \left[\left(x^\top U_T^\top w - x^\top \beta^\star \right)^2 \right] = \mathbb{E}_x \left[\left(x^\top (U_T^\top w - \beta^\star) \right)^2 \right] = (U_T^\top w - \beta^\star)^\top \Sigma_p (U_T^\top w - \beta^\star), \tag{250}$$

where we used $\mathbb{E}[xx^\top] = \Sigma_p$. Differentiating (250) with respect to w gives $\nabla_w \mathcal{L}_T(w) = 2 U_T \Sigma_p (U_T^\top w - \beta^\star)$. Hence every minimizer satisfies the normal equation

$$U_T \Sigma_p U_T^\top w = U_T \Sigma_p \beta^\star. \tag{251}$$

Choosing the minimum-Euclidean-norm solution gives $w_T^\star = (U_T \Sigma_p U_T^\top)^\dagger U_T \Sigma_p \beta^\star$. Therefore the exact Stage-I coefficient in the original input coordinates is

$$\widehat{\beta}_T = U_T^\top (U_T \Sigma_p U_T^\top)^\dagger U_T \Sigma_p \beta^\star. \tag{252}$$

Equation (252) is a finite-p identity; no deterministic-equivalent approximation has yet been used. Introduce whitened coordinates $\theta := \Sigma_p^{1/2} \beta, \theta^\star := \Sigma_p^{1/2} \beta^\star$. Since the teacher is supported on the strong block,

$$\theta^\star = \begin{pmatrix} \beta_{\mathrm{str}}^\star \\ 0 \end{pmatrix}, \qquad \|\theta^\star\|_2^2 = 1.$$

The population prediction risk becomes ordinary Euclidean squared distance:

$$\mathcal{R}(h_\beta) := (\beta - \beta^\star)^\top \Sigma_p (\beta - \beta^\star) = \left\| \Sigma_p^{1/2} (\beta - \beta^\star) \right\|_2^2 = \|\theta - \theta^\star\|_2^2. \tag{253}$$

Define the whitened Stage-I feature matrix

$$A_{T,p} := U_T \Sigma_p^{1/2} \in \mathbb{R}^{N_T \times p}. \tag{254}$$

Multiplying (252) by $\Sigma_p^{1/2}$ gives

$$\theta_T := \Sigma_p^{1/2} \widehat{\beta}_T = \Sigma_p^{1/2} U_T^\top (U_T \Sigma_p U_T^\top)^\dagger U_T \Sigma_p \beta^\star = A_{T,p}^\top (A_{T,p} A_{T,p}^\top)^\dagger A_{T,p} \theta^\star. \tag{255}$$

Thus

$$\theta_T = P_{T,p} \theta^\star, \qquad P_{T,p} := A_{T,p}^\top (A_{T,p} A_{T,p}^\top)^\dagger A_{T,p}. \tag{256}$$

The matrix $P_{T,p}$ is exactly the Euclidean orthogonal projector onto $\mathrm{row}(A_{T,p}) \subseteq \mathbb{R}^p$. Indeed, $P_{T,p}^\top = P_{T,p}, P_{T,p}^2 = P_{T,p}$, by the Moore–Penrose identities. Consequently, $\theta^\star - \theta_T \perp \theta_T$, and therefore

$$(\theta^\star - \theta_T)^\top \theta_T = 0. \tag{257}$$

By Pythagoras,

$$\|\theta^\star - \theta_T\|_2^2 = \|\theta^\star\|_2^2 - \|\theta_T\|_2^2 = 1 - \|\theta_T\|_2^2. \tag{258}$$

Taking expectation over U_T, the exact Stage-I risk is therefore

$$\overline{\mathcal{R}}_{T,p} = 1 - \mathbb{E}\|\theta_T\|_2^2. \tag{259}$$

This exact projection identity will later allow us to determine the total Stage-I second moment without a separate random matrix calculation. Split the whitened Stage-I estimator according to the two covariance blocks:

$$\theta_T = \begin{pmatrix} \theta_{T,\mathrm{str}} \\ \theta_{T,\mathrm{wk}} \end{pmatrix}, \theta_{T,\mathrm{str}} \in \mathbb{R}^k, \theta_{T,\mathrm{wk}} \in \mathbb{R}^{p-k}.$$

Define the aggregate strong and weak prediction energies

$$S_{T,p} := \mathbb{E}\|\theta_{T,\mathrm{str}}\|_2^2, \qquad V_{T,p} := \mathbb{E}\|\theta_{T,\mathrm{wk}}\|_2^2. \tag{260}$$

Because $\|\theta_T\|_2^2 = \|\theta_{T,\mathrm{str}}\|_2^2 + \|\theta_{T,\mathrm{wk}}\|_2^2$, equation (259) yields the exact finite-p identity

$$\overline{\mathcal{R}}_{T,p} = 1 - S_{T,p} - V_{T,p}. \tag{261}$$

Moreover, expanding (257) blockwise gives $(\beta^\star_{\mathrm{str}} - \theta_{T,\mathrm{str}})^\top \theta_{T,\mathrm{str}} - \|\theta_{T,\mathrm{wk}}\|_2^2 = 0$. Hence

$$(\beta^\star_{\mathrm{str}} - \theta_{T,\mathrm{str}})^\top \theta_{T,\mathrm{str}} = \|\theta_{T,\mathrm{wk}}\|_2^2. \tag{262}$$

Equivalently, $(\beta^\star_{\mathrm{str}})^\top \theta_{T,\mathrm{str}} = \|\theta_{T,\mathrm{str}}\|_2^2 + \|\theta_{T,\mathrm{wk}}\|_2^2$. Taking expectation gives

$$\mathbb{E}[(\beta^\star_{\mathrm{str}})^\top \theta_{T,\mathrm{str}}] = S_{T,p} + V_{T,p}. \tag{263}$$

The identities (261) and (263) are exact and will be repeatedly used in the Stage-II analysis.

□

Theorem D.3 (Stage-I deterministic equivalents)**.** *Assume the proportional high-dimensional regime specified in the preliminaries and the regularity conditions required by the one-point and two-point structured-Wishart deterministic-equivalent results. Let $\tau_{T,p} = 0$ when $N_T \geq p$; when $N_T < p$, let $\tau_{T,p} > 0$ be the unique solution of*

$$\mathrm{df}_p(\tau_{T,p}) = \frac{N_T}{p}, \qquad \mathrm{df}_p(\tau) := \frac{1}{p}\left[\frac{k}{1+\tau} + \frac{(p-k)\delta_p}{\delta_p + \tau}\right].$$

Define $a_{T,p} := \frac{1}{1+\tau_{T,p}}$, $c_{T,p} := a_{T,p} - a_{T,p}^2$, and

$$\Theta_{T,p} := \left[1 + \frac{k(\delta_p + \tau_{T,p})^2}{(p-k)\delta_p(1+\tau_{T,p})^2}\right]^{-1}.$$

Then

$$\overline{\mathcal{R}}_{T,p} \overset{\mathrm{DE}}{\simeq} \overline{\mathcal{R}}^{\mathrm{DE}}_{T,p} := \frac{\tau_{T,p}}{1+\tau_{T,p}},$$

$$\mathbb{E}\widehat{\beta}_T \overset{\mathrm{DE}}{\simeq} \begin{pmatrix} a_{T,p}\beta^\star_{\mathrm{str}} \\ 0 \end{pmatrix},$$

and

$$V_{T,p} \overset{\mathrm{DE}}{\simeq} V^{\mathrm{DE}}_{T,p} := c_{T,p}\Theta_{T,p}, \qquad S_{T,p} \overset{\mathrm{DE}}{\simeq} S^{\mathrm{DE}}_{T,p} := a_{T,p} - c_{T,p}\Theta_{T,p}.$$

Proof. **Structured-Wishart resolvent and effective ridge.**

We now pass from the exact projector representation to its high-dimensional deterministic equivalent.

For $j = 1, \ldots, N_T$, define the effective teacher-feature vector

$$z_{T,j} := \Sigma_p^{1/2} u_{T,j}.$$

Since $u_{T,j} \sim \mathcal{N}(0, I_p)$, $z_{T,j} \sim \mathcal{N}(0, \Sigma_p)$. Moreover, the j-th row of $A_{T,p} = U_T \Sigma_p^{1/2}$ is $z_{T,j}^\top$. Therefore,

$$\widehat{\Sigma}_{T,p} := \frac{1}{N_T} A_{T,p}^\top A_{T,p} = \frac{1}{N_T} \sum_{j=1}^{N_T} z_{T,j} z_{T,j}^\top = \Sigma_p^{1/2} \left(\frac{1}{N_T} U_T^\top U_T \right) \Sigma_p^{1/2}. \tag{264}$$

Thus $\widehat{\Sigma}_{T,p}$ is the empirical covariance of N_T independent samples from $\mathcal{N}(0, \Sigma_p)$. In particular, $\mathbb{E}\widehat{\Sigma}_{T,p} = \Sigma_p$, and $\widehat{\Sigma}_{T,p}$ is a structured Wishart matrix with population covariance Σ_p. To verify the projector identity algebraically, let $A_{T,p} = Q_{T,p} D_{T,p} R_{T,p}^\top$ be a thin singular-value decomposition, where $D_{T,p} = \mathrm{diag}(d_1, \ldots, d_r)$, $r = \mathrm{rank}(A_{T,p})$, and every $d_j > 0$. Then

$$\begin{aligned} P_{T,p} =& A_{T,p}^\top (A_{T,p} A_{T,p}^\top)^\dagger A_{T,p} = R_{T,p} D_{T,p} Q_{T,p}^\top \left(Q_{T,p} D_{T,p}^2 Q_{T,p}^\top \right)^\dagger Q_{T,p} D_{T,p} R_{T,p}^\top \\ =& R_{T,p} D_{T,p} D_{T,p}^{-2} D_{T,p} R_{T,p}^\top = R_{T,p} R_{T,p}^\top. \end{aligned} \tag{265}$$

Moreover, $\widehat{\Sigma}_{T,p} = R_{T,p} \left(\frac{D_{T,p}^2}{N_T} \right) R_{T,p}^\top$, so for every fixed bare ridge $\lambda > 0$,

$$\widehat{\Sigma}_{T,p} (\widehat{\Sigma}_{T,p} + \lambda I_p)^{-1} = R_{T,p} \, \mathrm{diag} \left(\frac{d_1^2}{d_1^2 + N_T \lambda}, \ldots, \frac{d_r^2}{d_r^2 + N_T \lambda} \right) R_{T,p}^\top. \tag{266}$$

Since

$$\lim_{\lambda \downarrow 0} \frac{d_j^2}{d_j^2 + N_T \lambda} = 1, \qquad j = 1, \ldots, r,$$

equations (265) and (266) give

$$P_{T,p} = \lim_{\lambda \downarrow 0} \widehat{\Sigma}_{T,p} (\widehat{\Sigma}_{T,p} + \lambda I_p)^{-1}. \tag{267}$$

We now apply Lemma C.1 (equation DE2). The population covariance in this application is Σ_p. Consequently, the empirical teacher-feature covariance resolvent has the deterministic equivalent

$$\widehat{\Sigma}_{T,p} (\widehat{\Sigma}_{T,p} + \lambda I_p)^{-1} \overset{\mathrm{DE}}{\simeq} \Sigma_p (\Sigma_p + \kappa_{T,\lambda} I_p)^{-1}. \tag{268}$$

where the renormalized ridge $\kappa_{T,\lambda} > 0$ is determined by

$$\kappa_{T,\lambda} \left[1 - \frac{p}{N_T} \, \mathrm{df}_p(\kappa_{T,\lambda}) \right] = \lambda. \tag{269}$$

Here

$$\mathrm{df}_p(\tau) := \mathrm{tr} \left[\Sigma_p (\Sigma_p + \tau I_p)^{-1} \right], \tag{270}$$

For the two-block covariance, $\Sigma_p = \mathrm{diag}\left(I_k, \delta_p I_{p-k} \right)$, so

$$\mathrm{df}_p(\tau) = \frac{1}{p} \left[\frac{k}{1+\tau} + \frac{(p-k)\delta_p}{\delta_p + \tau} \right]. \tag{271}$$

We will also need the squared shrinkage trace

$$\mathrm{df}_p^{(2)}(\tau) := \mathrm{tr} \left[\Sigma_p^2 (\Sigma_p + \tau I_p)^{-2} \right], \tag{272}$$

which in the present two-block model becomes

$$\mathrm{df}_p^{(2)}(\tau) = \frac{1}{p} \left[\frac{k}{(1+\tau)^2} + \frac{(p-k)\delta_p^2}{(\delta_p + \tau)^2} \right]. \tag{273}$$

Suppose first that $N_T < p$. Then the ridgeless limit remains on the positive effective-ridge branch. Sending $\lambda \downarrow 0$ in (269) gives

$$\kappa_{T,\lambda} \left[1 - \frac{p}{N_T} \, \mathrm{df}_p(\kappa_{T,\lambda}) \right] \longrightarrow 0.$$

Since the limiting effective ridge is positive in the underparameterized branch, its limit $\tau_{T,p} > 0$ must satisfy

$$1 - \frac{p}{N_T} \mathrm{df}_p(\tau_{T,p}) = 0.$$

Therefore,

$$\frac{k}{1+\tau_{T,p}} + \frac{(p-k)\delta_p}{\delta_p + \tau_{T,p}} = N_T, \qquad N_T < p \tag{274}$$

Because $\mathrm{df}_p(\tau)$ is continuous and strictly decreasing from 1 to 0, equation (274) has a unique positive solution.

If instead $N_T \geq p$, then $A_{T,p}$ has full column rank almost surely, so $P_{T,p} = I_p$ and Stage I recovers the target coefficient $\beta^\star$ exactly. In this branch we set $\tau_{T,p} = 0$. From the exact projection representation,

$$\overline{\mathcal{R}}_{T,p} = 1 - \mathbb{E}\left[(\theta^\star)^\top P_{T,p}\theta^\star\right].$$

Using (267) and the structured-Wishart deterministic equivalent, the ridgeless projector is replaced at leading order by $\Sigma_p(\Sigma_p + \tau_{T,p}I_p)^{-1}$. Therefore

$$\overline{\mathcal{R}}_{T,p} \overset{\mathrm{DE}}{\simeq} 1 - (\theta^\star)^\top \Sigma_p(\Sigma_p + \tau_{T,p}I_p)^{-1}\theta^\star. \tag{275}$$

The target coefficient lies entirely in the unit-eigenvalue block, hence $(\theta^\star)^\top \Sigma_p(\Sigma_p + \tau I_p)^{-1}\theta^\star = \frac{1}{1+\tau}\|\theta^\star\|_2^2 = \frac{1}{1+\tau}$.

So we have

$$\overline{\mathcal{R}}_{T,p} \overset{\mathrm{DE}}{\simeq} \overline{\mathcal{R}}^{\mathrm{DE}}_{T,p} := \frac{\tau_{T,p}}{1+\tau_{T,p}}. \tag{276}$$

It is convenient to define $a_{T,p} := \frac{1}{1+\tau_{T,p}}$. Then

$$\overline{\mathcal{R}}^{\mathrm{DE}}_{T,p} = 1 - a_{T,p}. \tag{277}$$

Combining this with the exact identity $S_{T,p} + V_{T,p} = 1 - \overline{\mathcal{R}}_{T,p}$ already shows that the total Stage-I prediction energy satisfies

$$S_{T,p} + V_{T,p} \overset{\mathrm{DE}}{\simeq} a_{T,p}. \tag{278}$$

We first determine the form of the mean exactly from the rotational symmetries of the model. The deterministic equivalent will only be needed afterward to determine the remaining scalar coefficient. For clarity, write the Stage-I estimator as a function of the random feature matrix:

$$\widehat{\beta}_T(U_T) = U_T^\top (U_T\Sigma_p U_T^\top)^\dagger U_T\Sigma_p\beta^\star.$$

Let $D \in O(p)$ be any orthogonal matrix satisfying $D\Sigma_p D^\top = \Sigma_p, D\beta^\star = \beta^\star$. The first identity implies that $D\Sigma_p = \Sigma_p D$. Evaluating the estimator at $U_T D$ gives

$$\begin{aligned}
\widehat{\beta}_T(U_T D) =& D^\top U_T^\top \left(U_T D\Sigma_p D^\top U_T^\top\right)^\dagger U_T D\Sigma_p\beta^\star \\
=& D^\top U_T^\top (U_T\Sigma_p U_T^\top)^\dagger U_T\Sigma_p D\beta^\star \\
=& D^\top U_T^\top (U_T\Sigma_p U_T^\top)^\dagger U_T\Sigma_p\beta^\star \\
=& D^\top \widehat{\beta}_T(U_T).
\end{aligned} \tag{279}$$

Because U_T has independent isotropic Gaussian rows, $U_T D \overset{d}{=} U_T$. Combining this distributional invariance with (279) yields

$$\widehat{\beta}_T(U_T) \overset{d}{=} D^\top \widehat{\beta}_T(U_T). \tag{280}$$

We now apply this identity to rotations of the weak block. For any $O \in O(p-k)$, let $D_O := \begin{pmatrix} I_k & 0 \\ 0 & O \end{pmatrix}$. Since $\Sigma_p = \begin{pmatrix} I_k & 0 \\ 0 & \delta_p I_{p-k} \end{pmatrix}, \beta^\star = \begin{pmatrix} \beta^\star_{\mathrm{str}} \\ 0 \end{pmatrix}$, we have $D_O\Sigma_p D_O^\top = \Sigma_p, D_O\beta^\star = \beta^\star$.

Define $\mu_{T,p} := \mathbb{E}\widehat{\beta}_T = \begin{pmatrix} \mu_{T,\mathrm{str},p} \\ \mu_{T,\mathrm{wk},p} \end{pmatrix}$. Taking expectations in (280) gives $\mu_{T,p} = D_O^\top \mu_{T,p}$, and therefore

$$\mu_{T,\mathrm{wk},p} = O^\top \mu_{T,\mathrm{wk},p} \qquad \text{for every } O \in O(p-k).$$

Choosing $O = -I_{p-k}$ gives $\mu_{T,\mathrm{wk},p} = -\mu_{T,\mathrm{wk},p}$, so the weak component of the mean vanishes exactly:

$$\mathbb{E}[\widehat{\beta}_{T,\mathrm{wk}}] = \mu_{T,\mathrm{wk},p} = 0. \tag{281}$$

To determine the direction of the strong mean, consider $D_R := \begin{pmatrix} R & 0 \\ 0 & I_{p-k} \end{pmatrix}$, where $R \in O(k)$ satisfies $R\beta^\star_{\mathrm{str}} = \beta^\star_{\mathrm{str}}$. Because the strong covariance block equals I_k, we again have

$$D_R \Sigma_p D_R^\top = \Sigma_p, \qquad D_R \beta^\star = \beta^\star.$$

Thus $\mu_{T,\mathrm{str},p} = R^\top \mu_{T,\mathrm{str},p}$ for every strong-block rotation that fixes $\beta^\star_{\mathrm{str}}$. Such rotations act arbitrarily on the orthogonal complement of $\beta^\star_{\mathrm{str}}$. Hence their common fixed subspace is $\mathrm{span}\{\beta^\star_{\mathrm{str}}\}$. It follows that, for some scalar $b_{T,p}$,

$$\mathbb{E}\widehat{\beta}_T = \begin{pmatrix} b_{T,p}\beta^\star_{\mathrm{str}} \\ 0 \end{pmatrix}. \tag{282}$$

This conclusion is exact for every finite p: symmetry determines the direction of the mean, but not yet the scalar $b_{T,p}$. We next determine $b_{T,p}$. Since $\|\beta^\star_{\mathrm{str}}\|_2^2 = (\beta^\star)^\top \Sigma_p \beta^\star = 1$, equation (282) implies

$$b_{T,p} = (\beta^\star_{\mathrm{str}})^\top \mathbb{E}[\widehat{\beta}_{T,\mathrm{str}}] = \mathbb{E}\left[(\theta^\star)^\top \theta_T\right] = \mathbb{E}\left[(\theta^\star)^\top P_{T,p}\theta^\star\right]. \tag{283}$$

Following the ridgeless branch in (267) and (268) gives

$$P_{T,p} \overset{\mathrm{DE}}{\simeq} \Sigma_p(\Sigma_p + \tau_{T,p} I_p)^{-1}.$$

In the strong deterministic-equivalence definition, choose the deterministic test matrix $M_p := \theta^\star(\theta^\star)^\top$. Then $\mathrm{Tr}(P_{T,p}M_p) = (\theta^\star)^\top P_{T,p}\theta^\star$. Because $P_{T,p}$ is an orthogonal projector, $0 \le (\theta^\star)^\top P_{T,p}\theta^\star \le \|\theta^\star\|_2^2 = 1$, so the corresponding convergence in probability also passes to the expectation. Therefore

$$b_{T,p} \overset{\mathrm{DE}}{\simeq} (\theta^\star)^\top \Sigma_p(\Sigma_p + \tau_{T,p} I_p)^{-1}\theta^\star = \frac{1}{1+\tau_{T,p}}\|\theta^\star\|_2^2 = \frac{1}{1+\tau_{T,p}} = a_{T,p}. \tag{284}$$

Equivalently, the same conclusion follows entirely from the exact cross identity:

$$b_{T,p} = \mathbb{E}[(\beta^\star_{\mathrm{str}})^\top \theta_{T,\mathrm{str}}] = S_{T,p} + V_{T,p} \overset{\mathrm{DE}}{\simeq} a_{T,p}.$$

Consequently,

$$\mathbb{E}\widehat{\beta}_T \overset{\mathrm{DE}}{\simeq} \begin{pmatrix} a_{T,p}\beta^\star_{\mathrm{str}} \\ 0 \end{pmatrix}. \tag{285}$$

Finally, using the exact mean form,

$$(\mathbb{E}\widehat{\beta}_T)^\top \Sigma_p (\mathbb{E}\widehat{\beta}_T) = b_{T,p}^2 (\beta^\star_{\mathrm{str}})^\top I_k \beta^\star_{\mathrm{str}} = b_{T,p}^2. \tag{286}$$

Together with $b_{T,p} \overset{\mathrm{DE}}{\simeq} a_{T,p}$, this gives

$$(\mathbb{E}\widehat{\beta}_T)^\top \Sigma_p (\mathbb{E}\widehat{\beta}_T) \overset{\mathrm{DE}}{\simeq} a_{T,p}^2. \tag{287}$$

Define the exact second-moment matrix $\mathsf{M}_{T,p} := \mathbb{E}\left[\widehat{\beta}_T \widehat{\beta}_T^\top\right]$. Partition it according to the strong and weak covariance blocks:

$$\mathsf{M}_{T,p} = \begin{pmatrix} \mathsf{M}_{\mathrm{ss},p} & \mathsf{M}_{\mathrm{sw},p} \\ \mathsf{M}_{\mathrm{ws},p} & \mathsf{M}_{\mathrm{ww},p} \end{pmatrix}.$$

For every $O \in O(p-k)$, the matrix D_O defined above satisfies $\widehat{\beta}_T(U_T D_O) = D_O^\top \widehat{\beta}_T(U_T), U_T D_O \overset{d}{=} U_T$. Therefore

$$\mathsf{M}_{T,p} = \mathbb{E}\left[\widehat{\beta}_T(U_T D_O)\widehat{\beta}_T(U_T D_O)^\top\right] = D_O^\top \mathbb{E}\left[\widehat{\beta}_T(U_T)\widehat{\beta}_T(U_T)^\top\right] D_O = D_O^\top \mathsf{M}_{T,p} D_O. \quad (288)$$

Expanding the right-hand side blockwise gives

$$D_O^\top \mathsf{M}_{T,p} D_O = \begin{pmatrix} \mathsf{M}_{\mathrm{ss},p} & \mathsf{M}_{\mathrm{sw},p} O \\ O^\top \mathsf{M}_{\mathrm{ws},p} & O^\top \mathsf{M}_{\mathrm{ww},p} O \end{pmatrix}.$$

Comparison with $\mathsf{M}_{T,p}$ yields

$$\mathsf{M}_{\mathrm{sw},p} O = \mathsf{M}_{\mathrm{sw},p}, \qquad O^\top \mathsf{M}_{\mathrm{ws},p} = \mathsf{M}_{\mathrm{ws},p},$$

and

$$O^\top \mathsf{M}_{\mathrm{ww},p} O = \mathsf{M}_{\mathrm{ww},p} \qquad \text{for every } O \in O(p-k).$$

Taking $O = -I_{p-k}$ in the first two identities gives $\mathsf{M}_{\mathrm{sw},p} = -\mathsf{M}_{\mathrm{sw},p}$, $\mathsf{M}_{\mathrm{ws},p} = -\mathsf{M}_{\mathrm{ws},p}$, so $\mathsf{M}_{\mathrm{sw},p} = \mathsf{M}_{\mathrm{ws},p} = 0$.

The last invariance identity states that the weak–weak block is unchanged by conjugation with every orthogonal matrix. Sign-flip matrices force all off-diagonal entries to vanish, and permutation matrices force all diagonal entries to be equal. Therefore there is a scalar $\nu_{T,p} \geq 0$ such that $\mathsf{M}_{\mathrm{ww},p} = \nu_{T,p} I_{p-k}$. It remains to determine $\nu_{T,p}$. On the weak block, $\theta_{T,\mathrm{wk}} = \delta_p^{1/2} \widehat{\beta}_{T,\mathrm{wk}}$, and hence

$$V_{T,p} = \mathbb{E}\|\theta_{T,\mathrm{wk}}\|_2^2 = \delta_p \mathbb{E}\|\widehat{\beta}_{T,\mathrm{wk}}\|_2^2 = \delta_p \operatorname{Tr}(\mathsf{M}_{\mathrm{ww},p}) = \delta_p \nu_{T,p}(p-k). \quad (289)$$

Thus $\nu_{T,p} = \frac{V_{T,p}}{(p-k)\delta_p}$. On the strong block, whitening has no effect because the corresponding eigenvalue of Σ_p is 1:

$$\theta_{T,\mathrm{str}} = \widehat{\beta}_{T,\mathrm{str}}.$$

Therefore $S_{T,p} = \mathbb{E}\|\widehat{\beta}_{T,\mathrm{str}}\|_2^2 = \operatorname{Tr}(\mathsf{M}_{\mathrm{ss},p})$. Writing $\mathsf{M}_{T,\mathrm{str},p} := \mathsf{M}_{\mathrm{ss},p}$, we arrive at the exact finite-p block structure

$$\mathsf{M}_{T,p} = \begin{pmatrix} \mathsf{M}_{T,\mathrm{str},p} & 0 \\ 0 & \dfrac{V_{T,p}}{(p-k)\delta_p} I_{p-k} \end{pmatrix}, \qquad \operatorname{Tr} \mathsf{M}_{T,\mathrm{str},p} = S_{T,p}. \quad (290)$$

No assumption such as $k = o(p)$ has been used. The equivariance, vanishing cross blocks, weak-block isotropy, and normalization above are exact for every finite p.

Let $\mu_{T,p} := \mathbb{E}\widehat{\beta}_T$ and define the covariance of the Stage-I estimator by $\Gamma_{T,p} := \mathsf{M}_{T,p} - \mu_{T,p}\mu_{T,p}^\top$. Its total prediction-norm trace is

$$\operatorname{Tr}(\Sigma_p \Gamma_{T,p}) = \mathbb{E}[\widehat{\beta}_T^\top \Sigma_p \widehat{\beta}_T] - \mu_{T,p}^\top \Sigma_p \mu_{T,p}. \quad (291)$$

The first term equals

$$S_{T,p} + V_{T,p} = 1 - \overline{\mathcal{R}}_{T,p}$$

exactly. Using (278) and (287), we obtain

$$\operatorname{Tr}(\Sigma_p \Gamma_{T,p}) \overset{\mathrm{DE}}{\simeq} a_{T,p} - a_{T,p}^2.$$

Define $c_{T,p} := a_{T,p} - a_{T,p}^2 = \frac{\tau_{T,p}}{(1+\tau_{T,p})^2}$. Thus $c_{T,p}$ is the deterministic-equivalent total projection-induced fluctuation energy.

Knowing the scalar total fluctuation energy $c_{T,p}$ does not yet identify its location. To separate the strong and weak contributions, we first write down the precise random quantity that must be evaluated; only after that do we invoke the two-point deterministic-equivalent lemma.

Let $D_p \succeq 0$ be a deterministic prediction-energy weight and define

$$\mathcal{F}_{D,p} := \mathbb{E}\left[(\widehat{\beta}_T - \mu_{T,p})^\top D_p (\widehat{\beta}_T - \mu_{T,p})\right], \qquad \mu_{T,p} := \mathbb{E}\widehat{\beta}_T. \quad (292)$$

Thus $\mathcal{F}_{D,p}$ is the amount of Stage-I covariance measured in the quadratic norm generated by D_p. Expanding the centered quadratic form gives the exact identity

$$\mathcal{F}_{D,p} = \mathbb{E}[\widehat{\beta}_T^\top D_p \widehat{\beta}_T] - \mu_{T,p}^\top D_p \mu_{T,p}. \tag{293}$$

We now express the first term through the random projector. Recall that $\widehat{\beta}_T = \Sigma_p^{-1/2} P_{T,p} \theta^\star$. Define $M_{D,p} := \Sigma_p^{-1/2} D_p \Sigma_p^{-1/2}$. Then, using $P_{T,p}^\top = P_{T,p}$,

$$\widehat{\beta}_T^\top D_p \widehat{\beta}_T = (\theta^\star)^\top P_{T,p} \Sigma_p^{-1/2} D_p \Sigma_p^{-1/2} P_{T,p} \theta^\star = (\theta^\star)^\top P_{T,p} M_{D,p} P_{T,p} \theta^\star. \tag{294}$$

Consequently, the random matrix that must be understood is $\mathbb{E}[P_{T,p} M_{D,p} P_{T,p}]$. This is a two-point object: the same random projector appears on both sides of $M_{D,p}$. A one-point equivalent for $P_{T,p}$ is not sufficient, because in general

$$\mathbb{E}[P_{T,p} M_{D,p} P_{T,p}] \neq (\mathbb{E} P_{T,p}) M_{D,p} (\mathbb{E} P_{T,p}).$$

The difference between these two matrices is exactly the contribution of the projector fluctuations. To place this random object in the resolvent setting, regularize the projector. For $\lambda > 0$, define $\widehat{T}_{T,\lambda} := \widehat{\Sigma}_{T,p}(\widehat{\Sigma}_{T,p} + \lambda I_p)^{-1}$. By (267),

$$\widehat{T}_{T,\lambda} \longrightarrow P_{T,p} \qquad \text{as } \lambda \downarrow 0.$$

Define the associated regularized coefficient and its mean by $\widehat{\beta}_{T,\lambda} := \Sigma_p^{-1/2} \widehat{T}_{T,\lambda} \theta^\star, \mu_{T,p,\lambda} := \mathbb{E}\widehat{\beta}_{T,\lambda}$. Then the regularized weighted fluctuation is exactly

$$\mathcal{F}_{D,p}(\lambda) := \mathbb{E}\Big[(\widehat{\beta}_{T,\lambda} - \mu_{T,p,\lambda})^\top D_p (\widehat{\beta}_{T,\lambda} - \mu_{T,p,\lambda})\Big] = (\theta^\star)^\top \mathbb{E}\Big[\widehat{T}_{T,\lambda} M_{D,p} \widehat{T}_{T,\lambda}\Big] \theta^\star - \mu_{T,p,\lambda}^\top D_p \mu_{T,p,\lambda}. \tag{295}$$

We now invoke Lemma D.1. Suppose first that $N_T < p$, and set $q_{T,p} := \frac{p}{N_T}$. For fixed $\lambda > 0$, write $G_{T,\lambda} := (\Sigma_p + \kappa_{T,\lambda} I_p)^{-1}, T_{T,\lambda} := \Sigma_p G_{T,\lambda}$. The equal-ridge specialization of the two-point lemma gives, for every deterministic M_p of uniformly bounded operator norm,

$$\mathbb{E}\Big[\widehat{T}_{T,\lambda} M_{D,p} \widehat{T}_{T,\lambda}\Big] \overset{\mathrm{DE}}{\simeq} T_{T,\lambda} M_{D,p} T_{T,\lambda} + \kappa_{T,\lambda}^2 G_{T,\lambda} \Sigma_p G_{T,\lambda} \frac{q_{T,p} \operatorname{tr}\left[\Sigma_p G_{T,\lambda} M_{D,p} G_{T,\lambda}\right]}{1 - q_{T,p} \operatorname{tr}\left[\Sigma_p^2 G_{T,\lambda}^2\right]}. \tag{296}$$

The first term on the right-hand side is the squared-mean contribution. Indeed, the one-point deterministic equivalent gives

$$\mu_{T,p,\lambda} \overset{\mathrm{DE}}{\simeq} \Sigma_p^{-1/2} T_{T,\lambda} \theta^\star,$$

and hence

$$\mu_{T,p,\lambda}^\top D_p \mu_{T,p,\lambda} \overset{\mathrm{DE}}{\simeq} (\theta^\star)^\top T_{T,\lambda} \Sigma_p^{-1/2} D_p \Sigma_p^{-1/2} T_{T,\lambda} \theta^\star = (\theta^\star)^\top T_{T,\lambda} M_{D,p} T_{T,\lambda} \theta^\star. \tag{297}$$

Therefore the first term in (296) cancels the squared-mean term in (295). The second term is precisely the leading covariance contribution. We apply the two-point deterministic equivalent only to the strong- and weak-block prediction-energy weights defined below. Both commute with Σ_p, and their whitened weights $M_{D,p} = \Sigma_p^{-1/2} D_p \Sigma_p^{-1/2}$ are orthogonal coordinate projectors. In particular,

$$\|M_{D,p}\|_{\mathrm{op}} = 1$$

uniformly in p. We first take the high-dimensional deterministic equivalent at fixed $\lambda > 0$, and then send $\lambda \downarrow 0$. Write $G_{T,p} := (\Sigma_p + \tau_{T,p} I_p)^{-1}, \gamma_{T,p} := q_{T,p} \operatorname{tr}\left[\Sigma_p^2 G_{T,p}^2\right]$. The surviving covariance term gives

$$\mathcal{F}_{D,p} \overset{\mathrm{DE}}{\simeq} \tau_{T,p}^2 (\theta^\star)^\top G_{T,p} \Sigma_p G_{T,p} \theta^\star \frac{q_{T,p} \operatorname{tr}\left[\Sigma_p G_{T,p} M_{D,p} G_{T,p}\right]}{1 - \gamma_{T,p}}. \tag{298}$$

We now simplify every factor. Since $\theta^\star$ is supported on the strong block, on which $\Sigma_p = I_k$, $(\theta^\star)^\top G_{T,p} \Sigma_p G_{T,p} \theta^\star = \frac{1}{(1+\tau_{T,p})^2}$. Since D_p, Σ_p, and $G_{T,p}$ commute, $\operatorname{tr}\left[\Sigma_p G_{T,p} M_{D,p} G_{T,p}\right] =$

$\operatorname{tr}[D_p G_{T,p}^2]$. It remains to simplify the denominator. The ridgeless fixed-point equation is $q_{T,p} \operatorname{tr}[\Sigma_p G_{T,p}] = 1$. Using

$$\Sigma_p G_{T,p} = \Sigma_p^2 G_{T,p}^2 + \tau_{T,p} \Sigma_p G_{T,p}^2,$$

we obtain

$$1 = q_{T,p} \operatorname{tr}[\Sigma_p G_{T,p}] = q_{T,p} \operatorname{tr}[\Sigma_p^2 G_{T,p}^2] + q_{T,p}\tau_{T,p} \operatorname{tr}[\Sigma_p G_{T,p}^2] = \gamma_{T,p} + q_{T,p}\tau_{T,p} \operatorname{tr}[\Sigma_p G_{T,p}^2].$$

Therefore $1 - \gamma_{T,p} = q_{T,p}\tau_{T,p} \operatorname{tr}\left[\Sigma_p G_{T,p}^2\right]$. Substituting these three simplifications into (298) yields

$$\mathcal{F}_{D,p} \overset{\mathrm{DE}}{\simeq} \tau_{T,p}^2 \frac{1}{(1+\tau_{T,p})^2} \frac{q_{T,p} \operatorname{tr}[D_p G_{T,p}^2]}{q_{T,p}\tau_{T,p} \operatorname{tr}[\Sigma_p G_{T,p}^2]} = \frac{\tau_{T,p}}{(1+\tau_{T,p})^2} \frac{\operatorname{tr}[D_p G_{T,p}^2]}{\operatorname{tr}[\Sigma_p G_{T,p}^2]}.$$

Since $c_{T,p} = \frac{\tau_{T,p}}{(1+\tau_{T,p})^2}$, $G_{T,p}^2 = (\Sigma_p + \tau_{T,p} I_p)^{-2}$, the desired weighted covariance formula is

$$\mathcal{F}_{D,p} \overset{\mathrm{DE}}{\simeq} c_{T,p} \frac{\operatorname{tr}\left[D_p (\Sigma_p + \tau_{T,p} I_p)^{-2}\right]}{\operatorname{tr}\left[\Sigma_p (\Sigma_p + \tau_{T,p} I_p)^{-2}\right]}. \tag{299}$$

Equation (299) says that the total fluctuation energy is $c_{T,p}$, while the trace ratio determines the fraction assigned to the directions selected by D_p. If $N_T \geq p$, then $P_{T,p} = I_p$ almost surely and $c_{T,p} = 0$, so every weighted covariance vanishes identically.

We now choose D_p to select the two blocks. For the strong prediction-energy weight,

$$D_{\mathrm{str},p} = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix},$$

the corresponding trace is

$$L_{\mathrm{str},p}(\tau) := \operatorname{tr}\left[D_{\mathrm{str},p}(\Sigma_p + \tau I_p)^{-2}\right] = \frac{k}{p} \frac{1}{(1+\tau)^2}, \tag{300}$$

For the weak prediction-energy weight, $D_{\mathrm{wk},p} = \begin{pmatrix} 0 & 0 \\ 0 & \delta_p I_{p-k} \end{pmatrix}$, we obtain

$$L_{\mathrm{wk},p}(\tau) := \operatorname{tr}\left[D_{\mathrm{wk},p}(\Sigma_p + \tau I_p)^{-2}\right] = \frac{p-k}{p} \frac{\delta_p}{(\delta_p + \tau)^2}. \tag{301}$$

Because $D_{\mathrm{str},p} + D_{\mathrm{wk},p} = \Sigma_p$, the denominator in (299) is exactly $L_{\mathrm{str},p}(\tau) + L_{\mathrm{wk},p}(\tau)$. Define the weak fraction of the Stage-I fluctuation energy by

$$\Theta_{T,p} := \frac{L_{\mathrm{wk},p}(\tau_{T,p})}{L_{\mathrm{str},p}(\tau_{T,p}) + L_{\mathrm{wk},p}(\tau_{T,p})}. \tag{302}$$

Substituting (300)–(301) gives

$$\Theta_{T,p} = \frac{\frac{p-k}{p} \frac{\delta_p}{(\delta_p + \tau_{T,p})^2}}{\frac{k}{p} \frac{1}{(1+\tau_{T,p})^2} + \frac{p-k}{p} \frac{\delta_p}{(\delta_p + \tau_{T,p})^2}} = \left[1 + \frac{k(\delta_p + \tau_{T,p})^2}{(p-k)\delta_p (1+\tau_{T,p})^2}\right]^{-1}. \tag{303}$$

Hence

$$\Theta_{T,p} = \left[1 + \frac{k(\delta_p + \tau_{T,p})^2}{(p-k)\delta_p (1+\tau_{T,p})^2}\right]^{-1}. \tag{304}$$

The weak mean vanishes exactly. Hence

$$\mathcal{F}_{D_{\mathrm{wk}},p} = \mathbb{E}\left[\widehat{\beta}_{T,\mathrm{wk}}^\top (\delta_p I_{p-k}) \widehat{\beta}_{T,\mathrm{wk}}\right] = \delta_p \mathbb{E}\|\widehat{\beta}_{T,\mathrm{wk}}\|_2^2 = V_{T,p}.$$

Therefore (299) and (302) give

$$V_{T,p} \overset{\mathrm{DE}}{\simeq} V_{T,p}^{\mathrm{DE}} := c_{T,p}\Theta_{T,p}. \tag{305}$$

For the strong weight, the same formula gives $\mathcal{F}_{D_{\mathrm{str}},p} \overset{\mathrm{DE}}{\simeq} c_{T,p}(1-\Theta_{T,p})$. The strong block contains both this fluctuation energy and the deterministic mean energy $a_{T,p}^2$, and therefore

$$S_{T,p} \overset{\mathrm{DE}}{\simeq} a_{T,p}^2 + c_{T,p}(1-\Theta_{T,p}) = a_{T,p}^2 + (a_{T,p} - a_{T,p}^2) - c_{T,p}\Theta_{T,p} = a_{T,p} - c_{T,p}\Theta_{T,p}. \tag{306}$$

Therefore

$$S_{T,p} \overset{\mathrm{DE}}{\simeq} S_{T,p}^{\mathrm{DE}} := a_{T,p} - c_{T,p}\Theta_{T,p}. \tag{307}$$

Combining (305) and (307),

$$\begin{aligned} V_{T,p}^{\mathrm{DE}} =& c_{T,p}\Theta_{T,p}, \\ S_{T,p}^{\mathrm{DE}} =& a_{T,p} - c_{T,p}\Theta_{T,p}. \end{aligned} \tag{308}$$

As a consistency check, $S_{T,p}^{\mathrm{DE}} + V_{T,p}^{\mathrm{DE}} = a_{T,p}$, so $1 - S_{T,p}^{\mathrm{DE}} - V_{T,p}^{\mathrm{DE}} = 1 - a_{T,p} = \frac{\tau_{T,p}}{1+\tau_{T,p}} = \overline{\mathcal{R}}_{T,p}^{\mathrm{DE}}$, exactly as required by the projection identity.

□

D.4 Stage II: linear-random-feature derivation conditional on Stage I

We now derive Stage II by specializing the linear-random-feature calculation of (Atanasov et al., 2026, Sec. IV). Conditional on Stage I, set

$$b := \widehat{\beta}_T$$

and regard $b \in \mathbb{R}^p$ as deterministic. Throughout Stage II, $\mathbb{E}_S$ denotes expectation over the Stage-II randomness $(U_S, \widetilde{X})$, conditional on the Stage-I teacher b.

Define $F_S := \frac{1}{\sqrt{N_S}} U_S^\top \in \mathbb{R}^{p\times N_S}$, $\widetilde{\Sigma} := \frac{1}{m}\widetilde{X}^\top\widetilde{X}$. Because the pseudo-labels are $\widetilde{y} = \widetilde{X}b$, the induced input-space coefficient of the minimum-norm ridgeless student is

$$\widehat{\beta}_S = F_S(\widetilde{X}F_S)^\dagger \widetilde{X}b. \tag{309}$$

Introduce a ridge $\lambda > 0$ in the normalized feature coordinates. Define

$$\begin{aligned} \widehat{v}_{S,\lambda} :=& \left(F_S^\top\widetilde{\Sigma}F_S + \lambda I_{N_S}\right)^{-1} F_S^\top\widetilde{\Sigma}b, \\ \widehat{\beta}_{S,\lambda} :=& F_S\widehat{v}_{S,\lambda} = H_{S,\lambda}b, \end{aligned} \tag{310}$$

where

$$H_{S,\lambda} := F_S\left(F_S^\top\widetilde{\Sigma}F_S + \lambda I_{N_S}\right)^{-1} F_S^\top\widetilde{\Sigma}.$$

Let $G_S := F_S F_S^\top$. The push-through identity $A(BA+\lambda I)^{-1} = (AB+\lambda I)^{-1}A$ gives

$$H_{S,\lambda} = G_S\widetilde{\Sigma}(G_S\widetilde{\Sigma} + \lambda I_p)^{-1} = I_p - \lambda(G_S\widetilde{\Sigma} + \lambda I_p)^{-1}. \tag{311}$$

Therefore the regularized displacement from the fixed teacher is

$$d_{S,\lambda} := \widehat{\beta}_{S,\lambda} - b = -Q_{S,\lambda}b, \qquad Q_{S,\lambda} := \lambda(G_S\widetilde{\Sigma} + \lambda I_p)^{-1}. \tag{312}$$

Thus Stage II requires two different random-matrix objects:

$$\mathbb{E}_S[Q_{S,\lambda}] \quad \text{for the cross term,} \tag{313}$$

$$\mathbb{E}_S\left[Q_{S,\lambda}^\top\Sigma_p Q_{S,\lambda}\right] \quad \text{for the mimic term.} \tag{314}$$

The first is a one-point resolvent problem; the second is a two-point or source-differentiated resolvent problem.

For the ridgeless displacement $d_S := \widehat{\beta}_S - b$, we have

$$\widehat{\beta}_S - \beta^\star = (b - \beta^\star) + d_S.$$

Consequently,

$$\mathcal{R}(\widehat{h}_S) = (b-\beta^\star)^\top \Sigma_p (b-\beta^\star) + d_S^\top \Sigma_p d_S + 2(b-\beta^\star)^\top \Sigma_p d_S. \tag{315}$$

Define $\mathcal{R}_S^{\mathrm{mim}}(b) := \mathbb{E}_S\left[d_S^\top \Sigma_p d_S\right], \mathcal{C}_S^{\mathrm{cross}}(b) := 2\mathbb{E}_S\left[(b-\beta^\star)^\top \Sigma_p d_S\right]$. Then the conditional identity is

$$\mathbb{E}_S[\mathcal{R}(\widehat{h}_S)] = \mathcal{R}(h_b) + \mathcal{R}_S^{\mathrm{mim}}(b) + \mathcal{C}_S^{\mathrm{cross}}(b). \tag{316}$$

Define the feature-dressed population covariance $\Sigma_{F,S} := \Sigma_p^{1/2} G_S \Sigma_p^{1/2}$. For a positive semidefinite $p \times p$ matrix A, write

$$\begin{aligned} \mathrm{df}_{A,p}^{(1)}(\kappa) &:= \mathrm{tr}\left[A(A+\kappa I_p)^{-1}\right], \\ \mathrm{df}_{A,p}^{(2)}(\kappa) &:= \mathrm{tr}\left[A^2(A+\kappa I_p)^{-2}\right]. \end{aligned} \tag{317}$$

At fixed F_S, the sample average in (Atanasov et al., 2026, Eq. (35)) introduces the first effective ridge $\widehat{\kappa}_{1,\lambda}(F_S) > 0$, determined by

$$\widehat{\kappa}_{1,\lambda}(F_S)\left[1 - \frac{p}{m}\,\mathrm{df}_{\Sigma_{F,S},p}^{(1)}(\widehat{\kappa}_{1,\lambda}(F_S))\right] = \lambda. \tag{318}$$

The corresponding two-resolvent stability parameter is

$$\widehat{\gamma}_{1,\lambda}(F_S) := \frac{p}{m}\,\mathrm{df}_{\Sigma_{F,S},p}^{(2)}(\widehat{\kappa}_{1,\lambda}(F_S)). \tag{319}$$

The positive-ridge mimic quadratic is exactly

$$d_{S,\lambda}^\top \Sigma_p d_{S,\lambda} = \lambda^2 b^\top (\widetilde{\Sigma} G_S + \lambda I_p)^{-1} \Sigma_p (G_S \widetilde{\Sigma} + \lambda I_p)^{-1} b. \tag{320}$$

To evaluate this product, introduce

$$\widetilde{T}_{b,F}(\kappa_1) := b^\top \Sigma_p^{1/2} (\Sigma_{F,S} + \kappa_1 I_p)^{-1} \Sigma_p^{1/2} b. \tag{321}$$

The source-differentiation calculation of (Atanasov et al., 2026, Eqs. (34)–(36)) gives

$$\mathbb{E}_{\widetilde{X}}\left[d_{S,\lambda}^\top \Sigma_p d_{S,\lambda} \,\middle|\, F_S, b\right] \overset{\mathrm{DE}}{\simeq} -\frac{\widehat{\kappa}_{1,\lambda}(F_S)^2}{1-\widehat{\gamma}_{1,\lambda}(F_S)} \frac{\partial}{\partial \kappa_1} \widetilde{T}_{b,F}(\kappa_1)\Big|_{\kappa_1 = \widehat{\kappa}_{1,\lambda}(F_S)}. \tag{322}$$

The derivative generates the two copies of the population resolvent, while $(1-\widehat{\gamma}_{1,\lambda}(F_S))^{-1}$ is the corresponding two-point stability factor.

We now average over F_S. Let $d_{1,\lambda} := \mathrm{df}_p(\kappa_{2,\lambda}), d_{2,\lambda} := \mathrm{df}_p^{(2)}(\kappa_{2,\lambda})$, where df_p and $\mathrm{df}_p^{(2)}$ are the Stage-I degrees-of-freedom functions. The feature average in (Atanasov et al., 2026, Eqs. (37)–(38)) introduces a second effective ridge

$$\kappa_{2,\lambda} = \bar{\kappa}_{1,\lambda} S_{G_S}(-d_{1,\lambda}), \tag{323}$$

where S_{G_S} is the limiting S-transform of $G_S = F_S F_S^\top$. For $F_S = \frac{1}{\sqrt{N_S}} U_S^\top$, the matrix G_S is a white Wishart Gram matrix with aspect ratio p/N_S. Its S-transform, evaluated at $-d_{1,\lambda}$, is

$$S_{G_S}(-d_{1,\lambda}) = \frac{1}{1-(p/N_S)d_{1,\lambda}}. \tag{324}$$

Consequently,

$$\kappa_{2,\lambda} = \frac{\bar{\kappa}_{1,\lambda}}{1-(p/N_S)d_{1,\lambda}}. \tag{325}$$

After averaging over F_S, introduce the deterministic first effective ridge

$$\bar{\kappa}_{1,\lambda} := \frac{\lambda}{1-(p/m)d_{1,\lambda}}. \tag{326}$$

Eliminating $\bar{\kappa}_{1,\lambda}$ between (325) and (326) gives the central nested-ridge equation

$$\lambda = \kappa_{2,\lambda}\left[1 - \frac{p}{m}d_{1,\lambda}\right]\left[1 - \frac{p}{N_S}d_{1,\lambda}\right], \qquad d_{1,\lambda} = \mathrm{df}_p(\kappa_{2,\lambda}). \tag{327}$$

The first bracket is the sample renormalization and the second is the random-feature renormalization. To state the feature-averaged mimic formula, define $\alpha_{S,\lambda} := \frac{d\log\kappa_{2,\lambda}}{d\log\bar{\kappa}_{1,\lambda}}$. The feature-averaged version of the sample stability parameter is

$$\bar{\gamma}_{1,\lambda} := \frac{p}{m}d_{1,\lambda}\left[1 - \frac{d_{1,\lambda} - d_{2,\lambda}}{d_{1,\lambda}}\alpha_{S,\lambda}\right]. \tag{328}$$

For a deterministic teacher vector b, define $T_b(\kappa) := b^\top\Sigma_p(\Sigma_p + \kappa I_p)^{-1}b$, so that $-T_b'(\kappa) = b^\top\Sigma_p(\Sigma_p + \kappa I_p)^{-2}b$. Define the positive-ridge conditional mimic risk by

$$\mathcal{R}^{\mathrm{mim}}_{S,\lambda}(b) := \mathbb{E}_S\left[d_{S,\lambda}^\top\Sigma_p d_{S,\lambda}\right]. \tag{329}$$

Specializing (Atanasov et al., 2026, Eq. (40)) to zero label noise gives the deterministic equivalent

$$\mathcal{R}^{\mathrm{mim}}_{S,\lambda}(b) \overset{\mathrm{DE}}{\simeq} -\frac{\kappa_{2,\lambda}^2 T_b'(\kappa_{2,\lambda})}{1-\bar{\gamma}_{1,\lambda}}\alpha_{S,\lambda} + \frac{\kappa_{2,\lambda}T_b(\kappa_{2,\lambda})}{1-\bar{\gamma}_{1,\lambda}}(1-\alpha_{S,\lambda}). \tag{330}$$

This formula is the explicit result of first averaging over the Stage-II samples and then averaging over the random features. The same two subordinations applied to the one-point quantity in (313) give

$$\mathbb{E}_S[Q_{S,\lambda}] \overset{\mathrm{DE}}{\simeq} \kappa_{2,\lambda}(\Sigma_p + \kappa_{2,\lambda}I_p)^{-1}. \tag{331}$$

Define the deterministic Stage-II ridgeless effective ridge by $\tau_{S,p} := \lim_{\lambda\downarrow 0}\kappa_{2,\lambda}$. Equation (327) shows algebraically that this limit can be reached in three distinct branches.

No bottleneck. If $p < \min\{m, N_S\}$, both brackets in (327) remain positive at $\kappa_{2,\lambda} = 0$. Hence $\tau_{S,p} = 0$. In fact, the conclusion is exact: F_S has full row rank and $\tilde{X}$ has full column rank almost surely, so $\widehat{\beta}_S = b$.

Feature bottleneck. If $N_S < \min\{m, p\}$, the feature bracket vanishes while the sample bracket remains positive. Thus $\tau_{S,p} > 0$ is the unique solution of $\mathrm{df}_p(\tau_{S,p}) = \frac{N_S}{p}$. In this branch,

$$\bar{\kappa}_{1,\lambda} \downarrow 0, \qquad \kappa_{2,\lambda} \longrightarrow \tau_{S,p} > 0,$$

so

$$\alpha_{S,\lambda} \longrightarrow 0, \qquad \bar{\gamma}_{1,\lambda} \longrightarrow \frac{N_S}{m}.$$

Taking the ridgeless limit in (330) yields

$$\mathcal{R}^{\mathrm{mim}}_{S}(b) \overset{\mathrm{DE}}{\simeq} \frac{\tau_{S,p}T_b(\tau_{S,p})}{1 - N_S/m}. \tag{332}$$

This is exactly the zero-noise specialization of (Atanasov et al., 2026, Eq. (42)).

Sample bottleneck. If $m < \min\{p, N_S\}$, the sample bracket vanishes while the feature bracket remains positive. Therefore $\tau_{S,p} > 0$ is the unique solution of $\mathrm{df}_p(\tau_{S,p}) = \frac{m}{p}$. Let $\gamma^{\mathrm{samp}}_{S,p} := \frac{p}{m}\mathrm{df}_p^{(2)}(\tau_{S,p})$. The overparameterized ridgeless formula (Atanasov et al., 2026, Eq. (44)), with zero label noise, gives

$$\mathcal{R}^{\mathrm{mim}}_{S}(b) \overset{\mathrm{DE}}{\simeq} \frac{-\tau_{S,p}^2 T_b'(\tau_{S,p})}{1-\gamma^{\mathrm{samp}}_{S,p}} + \tau_{S,p}T_b(\tau_{S,p})\left.\frac{d\log S_{G_S}(-d)}{d\log d}\right|_{d=m/p}. \tag{333}$$

For the white random features, $S_{G_S}(-d) = \frac{1}{1-(p/N_S)d}$, and therefore

$$\left.\frac{d\log S_{G_S}(-d)}{d\log d}\right|_{d=m/p} = \left.\frac{(p/N_S)d}{1-(p/N_S)d}\right|_{d=m/p} = \frac{m/N_S}{1-m/N_S}. \tag{334}$$

Substitution gives

$$\mathcal{R}_S^{\mathrm{mim}}(b) \overset{\mathrm{DE}}{\simeq} \frac{-\tau_{S,p}^2 T_b'(\tau_{S,p})}{1-(p/m)\,\mathrm{df}_p^{(2)}(\tau_{S,p})} + \frac{\tau_{S,p} T_b(\tau_{S,p})(m/N_S)}{1-m/N_S}. \tag{335}$$

The first term is the sample-projection contribution. Its stability denominator comes from the two-point/source differentiation in the data average. The second term is the finite-random-feature contribution and comes from the logarithmic derivative of the white Wishart S-transform. In particular, the denominator $1-\frac{p}{m}\,\mathrm{df}_p^{(2)}(\tau_{S,p})$ does not follow merely by differentiating the scalar function $\mathrm{df}_p(\tau)$; it is the two-point stability factor in (Atanasov et al., 2026, Eq. (44)).

Define the positive-ridge conditional cross term by

$$\mathcal{C}_{S,\lambda}^{\mathrm{cross}}(b) := 2\mathbb{E}_S\left[(b-\beta^\star)^\top \Sigma_p d_{S,\lambda}\right] = -2(b-\beta^\star)^\top \Sigma_p \mathbb{E}_S[Q_{S,\lambda}]b, \tag{336}$$

Using (331) and then sending $\lambda \downarrow 0$,

$$\mathcal{C}_S^{\mathrm{cross}}(b) \overset{\mathrm{DE}}{\simeq} -2\tau_{S,p}(b-\beta^\star)^\top \Sigma_p(\Sigma_p+\tau_{S,p}I_p)^{-1}b. \tag{337}$$

Set $b=\widehat{\beta}_T$. Throughout this subsection, $\mathbb{E}_T$ denotes expectation over the Stage-I randomness U_T. Define the exact Stage-I-averaged Stage-II contributions by

$$\begin{aligned}\mathcal{M}_{S,p} &:= \mathbb{E}_T\left[\mathcal{R}_S^{\mathrm{mim}}(\widehat{\beta}_T)\right],\\ \mathcal{C}_{S,p}^{\mathrm{cross}} &:= \mathbb{E}_T\left[\mathcal{C}_S^{\mathrm{cross}}(\widehat{\beta}_T)\right].\end{aligned} \tag{338}$$

Recall $S_{T,p} = \mathbb{E}\|\theta_{T,\mathrm{str}}\|_2^2, V_{T,p} = \mathbb{E}\|\theta_{T,\mathrm{wk}}\|_2^2$, where $\theta_T = \Sigma_p^{1/2}\widehat{\beta}_T$. Because $\widehat{\beta}_{T,\mathrm{str}} = \theta_{T,\mathrm{str}}, \delta_p\mathbb{E}\|\widehat{\beta}_{T,\mathrm{wk}}\|_2^2 = V_{T,p}$, the conditional target functions average exactly as

$$\mathbb{E}_T[T_{\widehat{\beta}_T}(\tau)] = \frac{S_{T,p}}{1+\tau} + \frac{V_{T,p}}{\delta_p+\tau} =: \mathcal{T}_{T,p}^{(1)}(\tau), \tag{339}$$

$$\mathbb{E}_T[-T'_{\widehat{\beta}_T}(\tau)] = \frac{S_{T,p}}{(1+\tau)^2} + \frac{V_{T,p}}{(\delta_p+\tau)^2} =: \mathcal{T}_{T,p}^{(2)}(\tau). \tag{340}$$

The cross target has a stronger exact simplification. Define

$$\mathcal{B}_{T,p}(\tau) := \mathbb{E}_T\left[(\widehat{\beta}_T-\beta^\star)^\top \Sigma_p(\Sigma_p+\tau I_p)^{-1}\widehat{\beta}_T\right].$$

The exact Stage-I projection identity gives $(\theta_{T,\mathrm{str}}-\beta^\star_{\mathrm{str}})^\top\theta_{T,\mathrm{str}} = -\|\theta_{T,\mathrm{wk}}\|_2^2$. Hence the strong part of $\mathcal{B}_{T,p}(\tau)$ is $-\frac{V_{T,p}}{1+\tau}$, whereas the weak part is $\frac{V_{T,p}}{\delta_p+\tau}$. Therefore

$$\mathcal{B}_{T,p}(\tau) = V_{T,p}\Delta_p(\tau), \qquad \Delta_p(\tau) := \frac{1}{\delta_p+\tau} - \frac{1}{1+\tau}. \tag{341}$$

Since $0<\delta_p<1$, one has $\Delta_p(\tau)>0$. Let $\mathcal{T}_{T,p}^{(r),\mathrm{DE}}(\tau), r\in\{1,2\}$, denote (339)–(340) with $S_{T,p}, V_{T,p}$ replaced by their Stage-I deterministic equivalents.

$$\mathcal{M}_{S,p}^{\mathrm{DE}} = \begin{cases} 0, & p<\min\{m,N_S\},\\ \dfrac{\tau_{S,p}\mathcal{T}_{T,p}^{(1),\mathrm{DE}}(\tau_{S,p})}{1-N_S/m}, & N_S<\min\{m,p\},\\ \dfrac{\tau_{S,p}^2\mathcal{T}_{T,p}^{(2),\mathrm{DE}}(\tau_{S,p})}{1-(p/m)\,\mathrm{df}_p^{(2)}(\tau_{S,p})} + \dfrac{\tau_{S,p}\mathcal{T}_{T,p}^{(1),\mathrm{DE}}(\tau_{S,p})(m/N_S)}{1-m/N_S}, & m<\min\{p,N_S\}. \end{cases} \tag{342}$$

Likewise, (337) and (341) give

$$\mathcal{C}_{S,p}^{\mathrm{cross}} \overset{\mathrm{DE}}{\simeq} \mathcal{C}_{S,p}^{\mathrm{cross,DE}} := -2\tau_{S,p}V_{T,p}^{\mathrm{DE}}\Delta_p(\tau_{S,p}). \tag{343}$$

Theorem D.4 (Stage-II risk and weak-to-strong gain)**.** *Under the proportional-regime assumptions of the Stage-I theorem and the linear-random-feature deterministic-equivalent assumptions of (Atanasov et al., 2026, Sec. IV),*

$$\overline{\mathcal{R}}_{S,p} \overset{\mathrm{DE}}{\simeq} \overline{\mathcal{R}}^{\mathrm{DE}}_{S,p} := \overline{\mathcal{R}}^{\mathrm{DE}}_{T,p} + \mathcal{M}^{\mathrm{DE}}_{S,p} + \mathcal{C}^{\mathrm{cross,DE}}_{S,p}. \tag{344}$$

Consequently,

$$\mathsf{W2SG}_p \overset{\mathrm{DE}}{\simeq} \mathsf{W2SG}^{\mathrm{DE}}_p := \overline{\mathcal{R}}^{\mathrm{DE}}_{T,p} - \overline{\mathcal{R}}^{\mathrm{DE}}_{S,p}. \tag{345}$$

Proof. Averaging (316) over Stage I gives the exact identity $\overline{\mathcal{R}}_{S,p} = \overline{\mathcal{R}}_{T,p} + \mathcal{M}_{S,p} + \mathcal{C}^{\mathrm{cross}}_{S,p}$. Substitute the Stage-I risk deterministic equivalent, (342), and (343) to obtain (344). Finally, from $\mathsf{W2SG}_p = \overline{\mathcal{R}}_{T,p} - \overline{\mathcal{R}}_{S,p}$, replacing the teacher and student risks by their deterministic equivalents yields

$$\mathsf{W2SG}_p \overset{\mathrm{DE}}{\simeq} \mathsf{W2SG}^{\mathrm{DE}}_p = \overline{\mathcal{R}}^{\mathrm{DE}}_{T,p} - \overline{\mathcal{R}}^{\mathrm{DE}}_{S,p},$$

which is (345). □

D.5 Positive-signal-fraction regime: assumptions and phase diagram

Assumption D.5 (Proportional asymptotic scaling)**.** *Let $d_p := p - k$ and assume*

$$\frac{k}{p} \to \vartheta, \qquad d_p\delta_p \to \eta, \qquad \frac{N_T}{p} \to \rho_T, \qquad \frac{m}{p} \to \rho_m, \qquad \frac{N_S}{p} \to \rho_S, \tag{346}$$

where $\vartheta, \rho_T \in (0,1)$, $\eta \in (0,\infty)$, and $\rho_m, \rho_S \in (0,\infty)$. We consider the regular fixed-ratio regime in which $\rho_T \neq \vartheta$, $\rho_m \neq \rho_S$, $\rho_m, \rho_S \neq \vartheta$, and $\rho_m, \rho_S \neq 1$. We exclude these critical boundaries from the fixed-ratio analysis; the relevant finite-dimensional and critical-scaling phenomena are discussed separately in Section D.6.

Theorem D.6 (Positive-signal-fraction phase diagram within the master DE)**.** *Under Assumption D.5 and the conditions of the Stage-I and Stage-II deterministic equivalents, let $\mathsf{W2SG}^{\mathrm{DE}}_p$ denote the master deterministic equivalent of the teacher-minus-student prediction risk. Define*

$$\Lambda_T := \begin{cases} \vartheta - \rho_T, & 0 < \rho_T < \vartheta, \\ \dfrac{(1-\vartheta)(\rho_T - \vartheta)}{1 - \rho_T}, & \vartheta < \rho_T < 1. \end{cases} \tag{347}$$

(a) Student bottleneck below the signal dimension. *If $\min\{\rho_m, \rho_S\} < \vartheta$, then $\mathsf{W2SG}^{\mathrm{DE}}_p < 0$ for all sufficiently large p.*

(b) Sample bottleneck above the signal dimension. *Suppose $\vartheta < \rho_m < \min\{1, \rho_S\}$. If the left-hand side of*

$$(\rho_m - 2\vartheta)\rho_S - \rho_m(2\rho_m - 3\vartheta + \Lambda_T) > 0 \tag{348}$$

is nonzero, then $\mathsf{W2SG}^{\mathrm{DE}}_p > 0$ for all sufficiently large p if and only if (348) holds. Equivalently, $\mathsf{W2SG}^{\mathrm{DE}}_p > 0$ if and only if

$$2\rho_m^2 - (\rho_S + 3\vartheta - \Lambda_T)\rho_m + 2\vartheta\rho_S < 0. \tag{349}$$

Define $D_{\mathrm{SB}} := (\rho_S + 3\vartheta - \Lambda_T)^2 - 16\vartheta\rho_S$. If $D_{\mathrm{SB}} > 0$, let

$$\rho_{m,\pm} := \frac{\rho_S + 3\vartheta - \Lambda_T \pm \sqrt{D_{\mathrm{SB}}}}{4}.$$

Then the open positive phase on the sample-bottleneck branch is $\rho_m \in (\rho_{m,-}, \rho_{m,+}) \cap (\vartheta, \min\{1, \rho_S\})$. If $D_{\mathrm{SB}} \leq 0$, or if this intersection is empty, there is no open positive phase on this branch.

(c) Feature bottleneck above the signal dimension. *Suppose $\vartheta < \rho_S < \min\{1, \rho_m\}$. If the left-hand side of*

$$(\rho_m - 2\rho_S)(\rho_S - \vartheta) - \rho_m\Lambda_T > 0 \tag{350}$$

is nonzero, then $\mathsf{W2SG}_p^{\mathrm{DE}} > 0$ *for all sufficiently large* p *if and only if* (350) *holds. Equivalently, if* $\rho_S - \vartheta - \Lambda_T > 0$*, the open positive phase is*

$$\rho_m > \frac{2\rho_S(\rho_S - \vartheta)}{\rho_S - \vartheta - \Lambda_T}. \tag{351}$$

If $\rho_S - \vartheta - \Lambda_T \le 0$*, then* $\mathsf{W2SG}_p^{\mathrm{DE}} < 0$ *for all sufficiently large* p *throughout the feature-bottleneck branch.*

(d) No student bottleneck. *If* $1 < \min\{\rho_m, \rho_S\}$*, then, under the almost-sure full-rank conditions of the two-stage model,*

$$\widehat{\beta}_S = \widehat{\beta}_T, \qquad \mathsf{W2SG}_p = 0$$

almost surely for all sufficiently large p*.*

Proof. Under Assumption D.5, $\delta_p \sim \eta/((1-\vartheta)p) \to 0$. By the Stage-II master deterministic equivalent,

$$\mathsf{W2SG}_p^{\mathrm{DE}} = 2\tau_{S,p} V_{T,p}^{\mathrm{DE}} \Delta_p(\tau_{S,p}) - \mathcal{M}_{S,p}^{\mathrm{DE}}. \tag{352}$$

Stage-I asymptotics. The equation $\mathrm{df}_p(\tau_{T,p}) = N_T/p$ is equivalent to

$$\frac{N_T}{p}\tau_{T,p}^2 + \left[\frac{N_T}{p} - \frac{k}{p} + \delta_p\left(\frac{N_T}{p} + \frac{k}{p} - 1\right)\right]\tau_{T,p} - \left(1 - \frac{N_T}{p}\right)\delta_p = 0.$$

If $0 < \rho_T < \vartheta$, its positive root satisfies $\tau_{T,p} \to (\vartheta - \rho_T)/\rho_T$, whereas if $\vartheta < \rho_T < 1$, writing $\tau_{T,p} = \delta_p t_{T,p}$ gives $\tau_{T,p}/\delta_p \to (1-\rho_T)/(\rho_T - \vartheta)$. Substitution into the Stage-I deterministic equivalents yields

$$\left(S_{T,p}^{\mathrm{DE}}, \frac{V_{T,p}^{\mathrm{DE}}}{\delta_p}\right) \longrightarrow \begin{cases} \left(\dfrac{\rho_T}{\vartheta}, \dfrac{(1-\vartheta)\rho_T}{\vartheta(\vartheta - \rho_T)}\right), & 0 < \rho_T < \vartheta, \\ \left(1, \dfrac{1-\rho_T}{\rho_T - \vartheta}\right), & \vartheta < \rho_T < 1. \end{cases}$$

Hence, with Λ_T as in (347),

$$\frac{\delta_p S_{T,p}^{\mathrm{DE}}}{V_{T,p}^{\mathrm{DE}}} \longrightarrow \frac{\Lambda_T}{1-\vartheta}, \qquad V_{T,p}^{\mathrm{DE}} \longrightarrow 0, \qquad \frac{\delta_p^2 S_{T,p}^{\mathrm{DE}}}{V_{T,p}^{\mathrm{DE}}} \longrightarrow 0. \tag{353}$$

Bottleneck below the signal dimension. Write $s_T := \lim_{p\to\infty} S_{T,p}^{\mathrm{DE}} > 0$. First suppose $\rho_m < \min\{\vartheta, \rho_S\}$. Then $\tau_{S,p} \to (\vartheta - \rho_m)/\rho_m$ and $\mathrm{df}_p^{(2)}(\tau_{S,p}) \to \rho_m^2/\vartheta$. The beneficial term in (352) tends to zero, while substitution into the Stage-II mimicry term gives

$$\mathcal{M}_{S,p}^{\mathrm{DE}} \longrightarrow s_T \frac{\vartheta - \rho_m}{\vartheta}\frac{\rho_S}{\rho_S - \rho_m},$$

and therefore

$$\mathsf{W2SG}_p^{\mathrm{DE}} \longrightarrow -s_T \frac{\vartheta - \rho_m}{\vartheta}\frac{\rho_S}{\rho_S - \rho_m} < 0.$$

If instead $\rho_S < \min\{\vartheta, \rho_m\}$, then $\tau_{S,p} \to (\vartheta - \rho_S)/\rho_S$, and the same substitution on the feature-bottleneck branch gives

$$\mathsf{W2SG}_p^{\mathrm{DE}} \longrightarrow -s_T \frac{\vartheta - \rho_S}{\vartheta}\frac{\rho_m}{\rho_m - \rho_S} < 0.$$

This proves part (a).

Bottleneck above the signal dimension. For an active Stage-II bottleneck ratio $\rho \in (\vartheta, 1)$, the fixed-point equation gives

$$\frac{\tau_{S,p}}{\delta_p} \longrightarrow \frac{1-\rho}{\rho - \vartheta}, \qquad \tau_{S,p}\Delta_p(\tau_{S,p}) \longrightarrow A(\rho) := \frac{1-\rho}{1-\vartheta}.$$

Using (353),

$$\frac{\tau_{S,p}\mathcal{T}_{T,p}^{(1),\mathrm{DE}}(\tau_{S,p})}{V_{T,p}^{\mathrm{DE}}} \longrightarrow A(\rho)\left(1+\frac{\Lambda_T}{\rho-\vartheta}\right), \qquad \frac{\tau_{S,p}^2\mathcal{T}_{T,p}^{(2),\mathrm{DE}}(\tau_{S,p})}{V_{T,p}^{\mathrm{DE}}} \longrightarrow A(\rho)^2. \tag{354}$$

On the sample-bottleneck branch $\vartheta < \rho_m < \min\{1, \rho_S\}$, take $\rho = \rho_m$. Moreover,

$$\mathrm{df}_p^{(2)}(\tau_{S,p}) \longrightarrow \vartheta + \frac{(\rho_m-\vartheta)^2}{1-\vartheta}, \qquad 1-\frac{p}{m}\,\mathrm{df}_p^{(2)}(\tau_{S,p}) \longrightarrow \frac{(\rho_m-\vartheta)(1-\rho_m)}{\rho_m(1-\vartheta)}.$$

Together with $m/(N_S-m) \to \rho_m/(\rho_S-\rho_m)$, substitution into (352) gives

$$\frac{\mathsf{W2SG}_p^{\mathrm{DE}}}{V_{T,p}^{\mathrm{DE}}A(\rho_m)} \longrightarrow 2-\frac{\rho_m}{\rho_m-\vartheta}-\frac{\rho_m}{\rho_S-\rho_m}\left(1+\frac{\Lambda_T}{\rho_m-\vartheta}\right) = \frac{(\rho_m-2\vartheta)\rho_S-\rho_m(2\rho_m-3\vartheta+\Lambda_T)}{(\rho_m-\vartheta)(\rho_S-\rho_m)}. \tag{355}$$

Since $V_{T,p}^{\mathrm{DE}} > 0$, $A(\rho_m) > 0$, and $(\rho_m-\vartheta)(\rho_S-\rho_m) > 0$, the eventual sign of $\mathsf{W2SG}_p^{\mathrm{DE}}$ is the sign of the numerator. This gives (348), equivalently (349); the discriminant and root characterization in part (b) follow from the quadratic formula.

On the feature-bottleneck branch $\vartheta < \rho_S < \min\{1, \rho_m\}$, take $\rho = \rho_S$ in (354). Since $m/(m-N_S) \to \rho_m/(\rho_m-\rho_S)$, substitution into (352) yields

$$\frac{\mathsf{W2SG}_p^{\mathrm{DE}}}{V_{T,p}^{\mathrm{DE}}A(\rho_S)} \longrightarrow 2-\frac{\rho_m}{\rho_m-\rho_S}\left(1+\frac{\Lambda_T}{\rho_S-\vartheta}\right) = \frac{(\rho_m-2\rho_S)(\rho_S-\vartheta)-\rho_m\Lambda_T}{(\rho_S-\vartheta)(\rho_m-\rho_S)}. \tag{356}$$

Since $V_{T,p}^{\mathrm{DE}} > 0$, $A(\rho_S) > 0$, and $(\rho_S-\vartheta)(\rho_m-\rho_S) > 0$, the eventual sign of $\mathsf{W2SG}_p^{\mathrm{DE}}$ is again the sign of the numerator. This gives (350). Rearranging yields (351) when $\rho_S-\vartheta-\Lambda_T > 0$; if this quantity is nonpositive, the limiting numerator is strictly negative throughout the feature-bottleneck branch. This proves parts (b) and (c).

No student bottleneck. If $1 < \min\{\rho_m, \rho_S\}$, then eventually $m > p$ and $N_S > p$. Under the stated full-rank conditions, $\widetilde{X}$ has full column rank and F_S has full row rank. Hence F_S is surjective, so there exists v_0 with $F_S v_0 = \widehat{\beta}_T$, and the minimum Stage-II residual is zero. Any least-squares minimizer $\widehat{v}_S$ therefore satisfies $\widetilde{X}(F_S\widehat{v}_S - \widehat{\beta}_T) = 0$. Since $\ker(\widetilde{X}) = \{0\}$, it follows that $F_S\widehat{v}_S = \widehat{\beta}_T$. Hence

$$\widehat{\beta}_S = \widehat{\beta}_T, \qquad \mathsf{W2SG}_p = 0$$

almost surely for all sufficiently large p. This proves part (d). □

D.6 Critical Boundaries and Finite-Dimensional Effects

The regular fixed-ratio phase diagram excludes the interpolation boundary, the signal-rank threshold, and the ambient no-bottleneck boundary. We record here the two boundary phenomena most relevant to the W2SG phase diagram.

Exact interpolation singularity. Let $r_p := \min\{m, N_S\}$, $g_p := |m-N_S|$, and assume that the Stage-I coefficient $b = \widehat{\beta}_T$ is nonzero almost surely and has finite population risk.

Theorem D.7 (Exact Stage-II interpolation-boundary classification)**.** *The following finite-dimensional statements hold.*

1. *If $r_p \geq p$, then both Stage-II maps have sufficient rank to identify the teacher coefficient: $\widehat{\beta}_S = \widehat{\beta}_T$ almost surely. Consequently, $\overline{\mathcal{R}}_{S,p} = \overline{\mathcal{R}}_{T,p}$, and $\mathsf{W2SG}_p = 0$.*
2. *If $r_p < p$ and $g_p \leq 1$, then the relevant inverse-Wishart first moment diverges. Under the nondegeneracy conditions of the two-stage model, $\overline{\mathcal{R}}_{S,p} = +\infty$ and $\mathsf{W2SG}_p = -\infty$.*
3. *If $r_p < p$ and $g_p \geq 2$, the Stage-II expected risk is finite. Its excess contribution contains the inverse-Wishart amplification factor*

$$\frac{1}{g_p-1}.$$

Proof. If $m \geq p$ and $N_S \geq p$, then $\widetilde{X}$ has full column rank and the student feature map has full row rank almost surely. The Stage-II normal equations therefore identify the teacher coefficient uniquely, giving $\widehat{\beta}_S = \widehat{\beta}_T$.

If $N_S < m$ and $N_S < p$, conditioning on the student feature map reduces Stage II to an N_S-dimensional Gaussian least-squares problem. Its variance contains the factor

$$\frac{N_S}{m - N_S - 1}.$$

If $m < N_S$ and $m < p$, conditioning instead on the Stage-II sample gives the dual factor

$$\frac{1}{N_S - m - 1}.$$

These expectations are infinite when $|m - N_S| \leq 1$ and finite when $|m - N_S| \geq 2$, proving the result. □

In particular, $m = N_S < p$ gives $\overline{\mathcal{R}}_{S,p} = +\infty$, whereas the exceptional point $m = N_S = p$ belongs to the exact no-bottleneck regime and satisfies

$$\widehat{\beta}_S = \widehat{\beta}_T, \qquad \mathsf{W2SG}_p = 0.$$

Thus, the internal interpolation singularity terminates when both Stage-II resources reach ambient dimension.

The signal threshold has a $\sqrt{p}$ critical window. Assume $\vartheta_p := \frac{k}{p} \to \vartheta \in (0,1), (p-k)\delta_p \to \eta \in (0,\infty)$, and $\varepsilon_p := \sqrt{\delta_p}$. Define

$$A(z) := \frac{\sqrt{z^2 + 4\vartheta(1-\vartheta)} - z}{2\vartheta}, \qquad L(z) := \sqrt{z^2 + 4\vartheta(1-\vartheta)}.$$

Let ρ_p denote a dimension-to-ambient ratio, such as m/p, N_S/p, or N_T/p. If

$$\frac{\rho_p - \vartheta_p}{\varepsilon_p} \to z$$

and the corresponding effective resolution satisfies $\mathrm{df}_p(\tau_p) = \rho_p$, then $\frac{\tau_p}{\varepsilon_p} \to A(z)$. Because

$$p\varepsilon_p \sim \sqrt{\frac{\eta p}{1-\vartheta}},$$

the transition around the signal dimension occurs in a dimension window of order $\sqrt{p}$. We next consider the jointly critical regime in which both the teacher width and the active Stage-II resource approach the signal dimension within their respective $\sqrt{p}$ windows. Suppose

$$\frac{N_T - k}{p\varepsilon_p} \to z_T, \qquad \frac{\min\{m, N_S\} - k}{p\varepsilon_p} \to z_S,$$

and define

$$A_S := A(z_S), \qquad L_S := L(z_S), \qquad v_T^{\mathrm{crit}} := \frac{1-\vartheta}{L(z_T)}.$$

Within the master deterministic equivalent, the two Stage-II bottleneck branches then behave as follows.

If m is the active bottleneck and $N_S/p \to \rho_S > \vartheta$, then

$$\mathsf{W2SG}_p^{\mathrm{DE}} \longrightarrow -\frac{\vartheta v_T^{\mathrm{crit}}}{L_S} < 0.$$

If N_S is the active bottleneck and $m/p \to \rho_m > \vartheta$, then

$$\frac{\mathsf{W2SG}_p^{\mathrm{DE}}}{\varepsilon_p} \longrightarrow 2v_T^{\mathrm{crit}} - \frac{\rho_m}{\rho_m - \vartheta}\left(A_S + v_T^{\mathrm{crit}}\right).$$

Consequently, the leading critical feature-bottleneck gain is positive exactly when

$$(\rho_m - 2\vartheta)v_T^{\mathrm{crit}} > \rho_m A_S.$$

Thus joint signal criticality is asymmetric: the sample-bottleneck branch remains unfavorable, while a nonempty positive W2SG region may survive on the feature-bottleneck branch. These statements concern the master deterministic equivalent; equality in the displayed sign condition requires a higher-order expansion.