WHEN ARE LINEAR TRANSFORMERS NEARLY RIDGE-OPTIMAL? TASK-VARYING NOISE AND COVARIANCE ADAPTATION FOR IN-CONTEXT REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time demonstrations for in-context learning may have noise levels that are shared within a task yet vary across tasks. We study this heterogeneity in in-context linear regression through a tractable one-layer linear-attention transformer. Under isotropic tasks, the population-optimal estimator reduces to scalar shrinkage of the empirical input-output cross-moment. We derive exact finite-sample bias–variance formulas and the optimal population scaling, then quantify two limitations of this rule. First, population tuning does not adapt to a new task's noise level. Although average risk depends only on the mean noise variance, the dispersion of excess conditional risk is exactly proportional to the variance of task noise, with a coefficient determined by normalized estimation variance. In the proportional limit, this coefficient peaks below the interpolation threshold and decreases throughout the overparameterized regime. Second, scalar shrinkage cannot adapt to the realized empirical covariance. We decompose its risk gap to optimally tuned ridge into a squared-bias reduction and a variance increase, and derive rates showing that the normalized additive gap vanishes in data-rich, highly overparameterized, and weak-signal regimes. Experiments with multi-layer softmax-attention transformers show partial recovery of the scalar-to-ridge gap where covariance adaptation offers substantial gains.

1 INTRODUCTION

Large language models based on the transformer architecture (Vaswani et al., 2017) have shown a remarkable capability for in-context learning (ICL): they can solve unseen tasks from input-output demonstrations given at test time, without parameter updates (Brown et al., 2020). Despite this empirical success, we still lack a clear understanding of how ICL uses in-context demonstrations, especially when the demonstrations are noisy (Min et al., 2022; Lyu et al., 2023; Yoo et al., 2022; Wei et al., 2023b; Shi et al., 2023; Wei et al., 2023a; Fei et al., 2023; Cheng et al., 2025; Gao et al., 2024). Prior work has primarily focused on the presence of noise in demonstrations, rather than systematic variation in noise levels across tasks.

We study the task-varying noise heterogeneity in an analytically tractable model of in-context linear regression with a simple linear-attention transformer (Schlag et al., 2021). The model is intentionally minimal. In the isotropic setting, the population-optimal rule in our reduced one-layer class is scalar shrinkage: it applies one coefficient to the empirical input-output cross-moment. This makes two limitations precise: if the coefficient is fixed at the population level, it cannot adjust to task-specific noise; and because the rule is scalar, it cannot vary its shrinkage across eigendirections of the empirical feature covariance matrix. We ask when these two limitations are large or small.

Figure 1 previews the answer. The left panel shows that, task-varying noise makes the excess conditional risk (with the irreducible noise term removed) fluctuate across tasks, and that this fluctuation is not monotone in the aspect ratio d/n, where d is the input dimension and n is the number of in-context demonstrations (Proposition 5.1). The right panel compares the simplified transformer (theory) with optimally tuned ridge regression, which uses the realized eigensystem of the empirical feature covariance matrix. The simplified transformer is not ridge-optimal in general, but it can be

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.

nearly ridge-optimal when the net value of covariance adaptation is small. The practical softmax-attention transformer experiments in the figure provide a qualitative check beyond the scalar model.

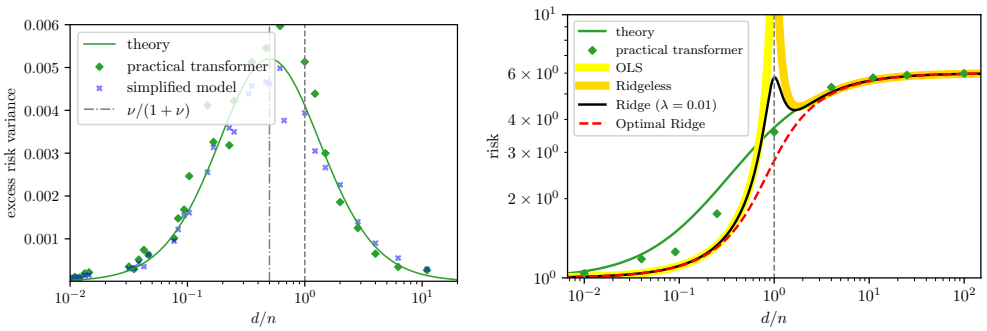

Figure 1: **Task-noise adaptation and covariance adaptation.** Left: variance across task-noise draws of the excess conditional risk $L(\theta^* \mid \xi) - \xi$ (the irreducible noise term $\xi = \sigma^2$ is excluded) at the population-optimal scaling θ^*. The theory curve is the asymptotic formula in Proposition 5.1. The effect is non-monotone in d/n: it increases up to $d/n = \nu/(1+\nu) < 1$ and decreases thereafter, in particular throughout the overparameterized regime where $d/n > 1$. Right: the simplified linear transformer (theory) differs from optimal ridge in general, but can be nearly ridge-optimal when $d/n \gg 1$ or when $d/n \ll 1$. Practical-transformer markers in both panels use fully parameterized 8-layer, 4-head softmax-attention transformers; experimental details are given in Appendix G.

The analysis is organized around a separation principle: **task-noise adaptation** and **feature-covariance adaptation** are distinct limitations of scalar linear attention. We first identify the population-optimal rule in the reduced linear-attention model: scalar shrinkage of the empirical input-output cross-moment, with exact finite-sample bias-variance formulas (Section 4). Holding this scalar form fixed, Section 5 shows that the conditional risk of this fixed scalar rule, given a task's own noise level, disperses across tasks in a way average risk cannot show: the dispersion is first order in the task-noise variance and governed by the normalized estimation variance. Section 6 then uses the optimal ridge as a benchmark to isolate the remaining value of feature-covariance adaptation. In the proportional limit, the scalar-to-ridge gap decomposes into a squared-bias reduction and a variance increase, and its rates explain why the simplified transformer is inferior to ridge in general yet nearly ridge-optimal when $n \gg d$, when $d \gg n$, or in low-signal regimes. The theory is stylized, but it separates the two adaptation mechanisms and gives two computable reference risks against which the practical-transformer experiments are measured.

2 Related Work

Transformers can be constructed to perform gradient-descent-like updates (Von Oswald et al., 2023) or to implement explicit in-context algorithms with Bayes-optimal behavior at sufficient width and depth (Akyürek et al., 2023). Complementing these construction results, *trained* transformers have been shown to perform preconditioned gradient descent (Ahn et al., 2023) and to recover linear estimators in-context (Zhang et al., 2024). For one layer of linear self-attention, Mahankali et al. (2024) prove that the global minimizer of the population pretraining loss implements one step of gradient descent, including with observation noise and non-isotropic features, and Ahn et al. (2023) characterize the minimizers in the noiseless Gaussian case; our Proposition 3.2 recovers their isotropic coefficient as the noiseless limit of the finite signal-to-noise optimum (Remark 3.3). Wu et al. (2024) study how many pretraining tasks one-layer linear attention needs on noisy linear regression, and Lu et al. (2025) give a random-matrix asymptotic theory of linear-attention ICL in a proportional regime related to ours. Li et al. (2024) analyze in-context linear estimation with correlated designs using one-layer linear attention and H3 models. Recent analyses of trained multi-head attention, including softmax attention, show mechanisms beyond single-head or scalar linear attention (Chen et al., 2024; He et al., 2025), and Bai et al. (2023) construct transformers that select among in-context algorithms, including the regularization level, from the demonstrations. Other perspectives study learnable function classes (Garg et al., 2022), generalization and stability (Li et al., 2023),

and Bayesian or task-diversity views of ICL (Xie et al., 2022; Panwar et al., 2024; Pan et al., 2023; Lin & Lee, 2024). Empirically, Raventós et al. (2023) show that above a pretraining task-diversity threshold, trained transformers' in-context regression behavior converges to ridge regression with a Gaussian task prior, and Carroll et al. (2025) find that trained transformers pass through a transient ridge-regression phase during training. Both are independent empirical evidence, at the scale of synthetic-task-trained transformers, for the same covariance-adaptive benchmark we compare against; neither examines task-varying noise or reports architecture-level diagnostics. Our analysis is complementary: rather than asking whether a rich transformer can implement a prespecified estimator, we ask what the population-optimal rule of the reduced one-layer class already achieves without either the task's own noise level or the realized feature covariance.

Closest to our setting, Vladymyrov et al. (2024) train multi-layer linear transformers on in-context regression tasks whose noise levels differ across tasks and find that the learned algorithm rescales its updates according to the noise level. Their results are empirical and mechanistic for deep linear models; ours are exact for one layer and quantify, through a task-wise oracle, how much such noise adaptation can be worth.

A substantial empirical literature studies how noise or irrelevant context in demonstrations influences ICL. Noisy labels can hurt ICL performance (Min et al., 2022; Lyu et al., 2023; Cheng et al., 2025), noisy annotations can harm generation tasks (Gao et al., 2024), and larger models can exhibit different sensitivity to noisy demonstrations (Yoo et al., 2022; Wei et al., 2023b; Shi et al., 2023). Related work also studies label bias and mitigation strategies (Fei et al., 2023). These papers mainly concern noisy or misleading demonstrations. Our focus is the distribution of noise levels across tasks, with one noise level shared by all demonstrations within a task.

Our ridge comparisons build on classical high-dimensional ridge and random matrix analyses (Dicker, 2016; Dobriban & Wager, 2018). The ridge formulas themselves are not the novelty. Here ridge is a diagnostic benchmark for in-context estimators: it separates the gain from scalar regularization from the additional gain of adapting to the realized eigensystem of the empirical feature covariance matrix. This viewpoint leads to the benchmark ordering and the scalar-to-ridge gap decompositions developed in Section 6. Because the tuned ridge is the Bayes posterior mean at the average noise level, its optimality among rules linear in the outputs is immediate; what we use from the random-matrix literature is the closed form that makes the gap and its components computable.

3 Model and Background

We consider an in-context regression instance encoded as

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} x^{(1)} & \cdots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & \cdots & y^{(n)} & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)}.$$

The first n columns are demonstrations, and the last column contains a query input whose output is masked. The prediction target is $y^{(n+1)}$. Each task is parameterized by a pair $(w, \sigma) \in \mathbb{R}^d \times \mathbb{R}_{>0}$, drawn from a distribution $\mathcal{D}_{\mathcal{W}}$ over $\mathbb{R}^d \times \mathbb{R}_{>0}$. These task parameters are latent: the learner observes the demonstrations and the query input, but not the realized values of w or σ. For $i = 1, \ldots, n+1$, the inputs are sampled as $x^{(i)} \sim \mathcal{D}_{\mathcal{X}}$, and the outputs are generated according to $y^{(i)} = w^\top x^{(i)} + \sigma\epsilon^{(i)}$, with $\epsilon^{(i)} \sim \mathcal{D}_{\mathcal{E}}$. Here, w and σ are shared across the demonstrations and the query within a task but **vary across tasks**. The vector w determines the signal, while the scale σ determines the **task-specific** noise level. This model captures a basic form of task variability in meta-learning and in-context learning: some tasks provide relatively clean demonstrations, while others are intrinsically noisier. In particular, task heterogeneity arises through both the regression parameter w and the noise scale σ. The variables $\{(x^{(i)}, \epsilon^{(i)})\}_{i\geq 1}$ are i.i.d. across i and independent of the task parameters (w, σ). Define

$$\bar{X} = [x^{(1)} \ \cdots \ x^{(n)}] \in \mathbb{R}^{d\times n}, \quad \bar{Y} = [y^{(1)} \ \cdots \ y^{(n)}] \in \mathbb{R}^{1\times n}, \quad \bar{\epsilon} = [\epsilon^{(1)}, \ldots, \epsilon^{(n)}]^\top \in \mathbb{R}^n.$$

Then $\bar{Y}^\top = \bar{X}^\top w + \sigma\bar{\epsilon}$.

Model. We study a one-layer linear self-attention model parameterized by matrices $P, Q \in \mathbb{R}^{(d+1)\times(d+1)}$, defined by

$$T_{P,Q}(Z) = \left[Z + \frac{1}{n}\mathrm{LSA}_{P,Q}(Z) \right]_{d+1,n+1}, \tag{1}$$

where

$$\mathrm{LSA}_{P,Q}(Z) = PZMZ^\top QZ, \quad M = \begin{bmatrix} I_n & 0_{n\times 1} \\ 0_{1\times n} & 0 \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}.$$

Reduced parameterization. Only the following entries of P and Q affect the predicted output $T_{P,Q}(Z)$, so it is enough to consider

$$P = \begin{bmatrix} 0_{d\times d} & 0_d \\ p^\top & \varphi \end{bmatrix}, \quad Q = \begin{bmatrix} \bar{Q} & 0_d \\ q^\top & 0 \end{bmatrix}, \tag{2}$$

where $p, q \in \mathbb{R}^d$, $\varphi \in \mathbb{R}$, and $\bar{Q} \in \mathbb{R}^{d\times d}$. This reduction is without loss for the query output and follows by direct multiplication of $PZMZ^\top QZ$. Under this parameterization, the model admits the linear representation $T_{P,Q}(Z) = \hat{w}^\top x^{(n+1)}$, where

$$\hat{w} = \bar{Q}^\top G_x p + (\varphi \bar{Q}^\top + qp^\top) G_{xy} + \varphi q G_y,$$

with empirical moments $G_x = \bar{X}\bar{X}^\top/n \in \mathbb{R}^{d\times d}$, $G_{xy} = \bar{X}\bar{Y}^\top/n \in \mathbb{R}^d$, and $G_y = \bar{Y}\bar{Y}^\top/n \in \mathbb{R}$.

Simplified model. Setting $p = q = 0_d$ and imposing $\bar{Q} = \kappa I_d$ yields $\hat{w} = \varphi\kappa G_{xy}$. We denote the resulting model by $T_{\varphi,\kappa}$. Its output is $T_{\varphi,\kappa}(Z) = \hat{w}^\top x^{(n+1)}$.

Assumptions. We use the following assumptions.

Assumption 3.1 (Data-generating process)**.** The data satisfy:

1. $y^{(i)} = w^\top x^{(i)} + \sigma\epsilon^{(i)}$.
2. $\{(x^{(i)}, \epsilon^{(i)})\}_{i\geq 1}$ are i.i.d., with $x^{(i)}$ independent of $\epsilon^{(i)}$, w independent of σ, and the pair $(x^{(i)}, \epsilon^{(i)})$ independent of (w, σ).
3. $x^{(i)} = z^{(i)}/\sqrt{d}$, where the coordinates of $z^{(i)}$ are i.i.d. with mean zero and unit variance, $\mathbb{E}[\epsilon^{(i)}] = 0$, and $\mathrm{Var}(\epsilon^{(i)}) = 1$.
4. $\mathbb{E}\|w\|^2 < \infty$, $\mathbb{E}[\sigma^2] < \infty$, $\mathbb{E}[\sigma^4] < \infty$, $\mathbb{E}\|x^{(i)}\|^4 < \infty$, and $\mathbb{E}[(\epsilon^{(i)})^4] < \infty$.
5. $\mathbb{E}[ww^\top] = \tau^2 I_d$ for some $\tau^2 > 0$.

Note that we do not make a *Gaussianity* assumption. Here, we define the signal-to-noise ratio $\nu := \frac{\tau^2}{\mathbb{E}[\sigma^2]}$.

Population prediction risk. For any estimator $\hat{w}$ constructed from the demonstrations, define the query predictor $\hat{y}^{(n+1)} = \hat{w}^\top x^{(n+1)}$ and the population prediction risk $L := \mathbb{E}[(y^{(n+1)} - \hat{y}^{(n+1)})^2]$. Equivalently, $L = \mathbb{E}[((w-\hat{w})^\top x^{(n+1)} + \sigma\epsilon^{(n+1)})^2]$; the cross-product term vanishes and we obtain

$$L = \frac{1}{d}\mathbb{E}_{w,\sigma,\bar{X},\bar{\epsilon}}\left[\|w - \hat{w}\|^2\right] + \mathbb{E}\left[\sigma^2\right].$$

Optimality of the simplified model. Isotropy and vanishing odd moments of w lead to the following optimality of the simplified model $\hat{w} = \theta G_{xy}$ (All proofs are deferred to the Appendix).

Proposition 3.2 (Optimality of the simplified model)**.** *Assuming that $\mathbb{E}[w_i] = 0$ and $\mathbb{E}[w_i w_j w_k] = 0$ for all i, j, k, the minimum population risk over the reduced parameterization $(p, q, \varphi, \bar{Q})$ is attained by an estimator of the form $\hat{w} = \theta G_{xy}$ for some $\theta \in \mathbb{R}$. Thus the reduced-class optimum equals the optimum over scalar rules $\{\theta G_{xy} : \theta \in \mathbb{R}\}$. If θ^* denotes a scalar minimizer, the minimum can be attained by choosing $p^* = q^* = 0_d$, $\bar{Q}^* = \kappa^* I_d$, and $\varphi^*\kappa^* = \theta^*$. The closed form for θ^* is computed in Proposition 4.1.*

Thus, within the reduced parameterization in (2), the simplified model $T_{\varphi,\kappa}$ is not a further restriction under isotropic tasks with vanishing odd moments. Nothing is assumed about the feature law beyond Assumption 3.1: the coordinates may be skewed, bounded, or heavy-tailed. We also show that this confinement of the reduced class to a fixed linear map of G_{xy} does not depend on isotropy (see Propositions F.1 and F.2 in Appendix F.1): for Gaussian features with any positive-definite covariance, and more generally for centrally symmetric feature and task laws, the optimum is still BG_{xy} for some $B \in \mathbb{R}^{d\times d}$, and isotropy is only what forces B to be a scalar multiple of I_d.

Remark 3.3 (Relation to single-layer linear-attention optima). The scalar form of the optimum is not new. Ahn et al. (2023, Theorem 1) characterize the global minimizers of the single-layer linear-attention pretraining objective for noiseless Gaussian linear regression, with a covariance-preconditioned-gradient interpretation in the non-isotropic case, and Mahankali et al. (2024) show that one step of gradient descent is the population optimum for one layer of linear self-attention including with noise. The (normalized) optimum θ^*/d of Proposition 4.1 has noiseless limit of $n/(n+d+1)$ for Gaussian features as $\nu \to \infty$, the isotropic coefficient of Ahn et al. (2023, Eq. (7)). What Propositions 3.2 and 4.1 add is the finite-ν optimum under task-varying noise with its exact squared-bias, variance, and risk components.

4 Bias and Variance Decomposition for Linear Transformers

We begin with the risk decomposition and the limits of its components at the optimal scaling. The risk is a quadratic in θ whose coefficients depend on the feature law through the single quantity $\psi_{d,n}$.

Proposition 4.1 (risk decomposition and optimal scaling)**.** *Under Assumption 3.1, for the simplified model $T_{\varphi,\kappa}$ in Section 3, define $\theta := \varphi\kappa$. The population prediction risk admits the decomposition*

$$\begin{aligned} L(\theta) &= \mathrm{Var}(\theta) + \mathrm{Bias}^2(\theta) + \mathbb{E}\big[\sigma^2\big] \\ &= \mathbb{E}[\sigma^2]\left[1 + \nu - \frac{2\nu}{d}\theta + \frac{1+\nu\psi_{d,n}}{dn}\,\theta^2\right], \end{aligned}$$

where

$$\mathrm{Var}(\theta) := \frac{1}{d}\,\mathbb{E}\big[\mathrm{Tr}\big(\mathrm{Cov}(\hat{w} \mid w,\sigma,\bar{X})\big)\big] = \frac{\theta^2}{dn}\,\mathbb{E}\big[\sigma^2\big]\,,$$

$$\mathrm{Bias}^2(\theta) := \frac{1}{d}\,\mathbb{E}_{w,\sigma,\bar{X}}\Big[\big\|w - \mathbb{E}_{\bar{\epsilon}}[\hat{w} \mid w,\sigma,\bar{X}]\big\|^2\Big] = \nu\mathbb{E}[\sigma^2]\left(1 - \frac{2\theta}{d} + \theta^2\frac{\psi_{d,n}}{dn}\right),$$

and

$$\nu := \frac{\tau^2}{\mathbb{E}[\sigma^2]}, \quad \psi_{d,n} := \; n\,\mathbb{E}[\mathrm{Tr}(G_x^2)] \;=\; \mathbb{E}\|x^{(1)}\|^4 + \frac{n-1}{d}.$$

Minimizing this quadratic risk gives the optimal scaling

$$\theta^* = \frac{\nu n}{1+\nu\psi_{d,n}}. \tag{3}$$

If, with ν fixed, $d, n \to \infty$ with $d/n \to \gamma \in (0,\infty)$, then $\psi_{d,n} \to 1 + 1/\gamma$, and the limiting normalized squared bias, variance, and minimum risk depend only on ν and γ, with

$$\frac{1}{\mathbb{E}[\sigma^2]}\big(\mathrm{Bias}^2(\theta^*), \mathrm{Var}(\theta^*), L(\theta^*)\big) \to \left(\frac{\nu\gamma\left(\gamma(1+\nu)^2+\nu^2\right)}{\left(\gamma(1+\nu)+\nu\right)^2}, \frac{\nu^2\gamma}{\left(\gamma(1+\nu)+\nu\right)^2}, 1 + \frac{\nu\gamma(1+\nu)}{\gamma(1+\nu)+\nu}\right).$$

For fixed ν, these limits imply that normalized squared bias and minimum risk increase with $\gamma = d/n$; fewer demonstrations relative to dimension therefore raise the scalar rule's minimum risk. The variance is hump-shaped: it increases up to $\gamma = \nu/(1+\nu)$ and decreases thereafter, so it is already decreasing throughout the overparameterized regime $\gamma > 1$.

Remark 4.2 (Empirical validation). Figure 2 compares the closed-form formulas (solid lines) with trained one-layer linear-attention transformers ($\times$) and practical transformers ($\blacklozenge$). The agreement with the simplified model is close: the markers confirm the predicted increase in squared bias and total risk as $\gamma = d/n$ varies, together with the hump-shaped variance component, and show the corresponding dependence on the signal-to-noise ratio ν. See Appendix E.4 for additional experimental validation and Appendix F for anisotropic feature analysis.

Overall, the formulas set up two remaining questions. First, what is lost because a population-tuned scalar rule cannot adapt to the task-specific noise variance σ^2, and when is this loss small? Second, what is lost because scalar shrinkage cannot adapt to the realized eigensystem of G_x, and when is this loss small? The next two sections answer these two in turn.

5 Task-Noise Heterogeneity: Risk Dispersion and Adaptation

Different tasks may have different noise levels. A population-tuned scalar transformer uses one scaling for all tasks, so for any fixed scaling the averaged risk depends on the noise law only through $\mathbb{E}[\sigma^2]$: task heterogeneity is invisible in average risk. It appears only in how the conditional risk $L(\theta \mid \xi)$, given a task's own noise level $\xi := \sigma^2$, varies from task to task.

We study the fluctuation attributable to the estimator: the excess conditional risk $L(\theta \mid \xi) - \xi$ of the population-tuned scalar rule. The proof of Proposition 4.1 gives $L(\theta \mid \xi) = \mathrm{Bias}^2(\theta) + \theta^2\xi/(dn) + \xi$, and the squared bias does not depend on ξ, so only the estimation-variance component fluctuates.

Proposition 5.1 (Dispersion of excess risk at the population scaling)**.** *Under Assumption 3.1, with* $\mathrm{Var}(\xi) > 0$*, with* ν *fixed and* $d, n \to \infty$ *with* $d/n \to \gamma \in (0, \infty)$*,*

$$\frac{\mathrm{Var}_\xi[L(\theta^* \mid \xi) - \xi]}{\mathrm{Var}(\xi)} \to \left(\frac{\nu^2\gamma}{(\gamma(1+\nu)+\nu)^2}\right)^2 = \begin{cases} \Theta(\gamma^2) & (\gamma \to 0), \\ \Theta(\gamma^{-2}) & (\gamma \to \infty), \end{cases}$$

which is maximized at $\gamma = \nu/(1+\nu) < 1$ *and strictly decreasing for* $\gamma > \nu/(1+\nu)$*.*

The excess-risk fluctuation is non-monotone in γ and decreases throughout the overparameterized regime; this is the pattern in Figure 1 (left). This differs from the classical double-descent variance peak of ordinary least squares and ridgeless regression at the interpolation threshold $\gamma = 1$ (Belkin et al., 2019; Hastie et al., 2022). Double descent fixes the task (hence the noise level) and asks how the risk of a single estimator behaves as γ varies; here γ is fixed and the task noise varies, and the object of interest is the *dispersion of conditional risk across tasks*, which peaks at the generally different point $\gamma = \nu/(1+\nu) < 1$ rather than at $\gamma = 1$.

Proposition 5.1 holds the scaling fixed. The same leading dispersion coefficient arises under oracle tuning along the local perturbation path of Assumption C.3. Write $L_{\mathrm{oracle}}(\xi) = \min_\theta L(\theta \mid \xi)$ for the task-wise oracle risk. Now, both the variance and bias terms depend on ξ because the optimal θ depends on ξ. Expanding the task-wise oracle risk around the average noise level $\bar{\xi} := \mathbb{E}[\sigma^2]$ gives, by the delta method, as $\eta \to 0$ with $\xi = \xi_\eta$ at fixed (d, n),

$$\frac{\mathrm{Var}_\xi[L_{\mathrm{oracle}}(\xi) - \xi]}{\mathrm{Var}(\xi)} = \left\{ \frac{d}{d\xi}\left(L_{\mathrm{oracle}}(\xi) - \xi\right)\Big|_{\xi=\bar{\xi}} \right\}^2 (1 + o(1)).$$

Moreover, the derivative is not an unknown quantity: it equals the normalized estimation variance of the oracle rule itself, for the one-parameter class of linear rules including the trained scalar transformer and oracle-tuned ridge-type rules considered in Section 6.

Proposition 5.2 (Risk derivative and normalized variance)**.** *Consider a one-parameter class of linear rules* $\hat{w}_\alpha = A_\alpha G_{xy}$*, where* A_α *depends on the demonstrations only through* $\bar{X}$ *and does not otherwise depend on* $\xi = \sigma^2$*. If* $\alpha^*(\xi)$ *is an interior differentiable minimizer, then the derivative of the excess risk is the normalized variance:*

$$\frac{d}{d\xi}\left(L_{\mathrm{oracle}}(\xi) - \xi\right)\Big|_{\xi=\bar{\xi}} = \frac{\mathrm{Var}(\hat{w}_{\alpha^*(\bar{\xi})} \mid \bar{\xi})}{\bar{\xi}}.$$

The coefficient is therefore the same object as in Proposition 5.1.

Noise adaptation and attainability. The dispersion above measures sensitivity to task noise, whereas the mean oracle regret $\Delta L = \mathbb{E}_\xi[L(\theta^* \mid \xi) - L_{\mathrm{oracle}}(\xi)]$ measures the average cost of using one population coefficient instead of tuning to each task's noise level. Under Assumption 3.1, Proposition C.1 gives an exact finite-sample expression for ΔL as a weighted second central moment of ξ, valid without a local perturbation assumption. Along the local perturbation path of

Assumption C.3, as $\eta \to 0$ at fixed (d, n), the mean regret is of order $\mathrm{Var}(\xi_\eta)$, while the standard deviation of excess conditional risk is of order $\mathrm{sd}(\xi_\eta)$; population and oracle tuning have the same leading excess-risk dispersion. Thus dispersion and the mean benefit of adaptation describe different aspects of task heterogeneity.

Appendix C.5 also studies whether this mean benefit can be attained from demonstrations. Under Assumption 3.1, with Gaussian demonstration noise, known τ^2, and a design matrix of full column rank almost surely with $n > d$, the residual estimate $\hat{\sigma}^2$ yields the plug-in rule $\theta^*(\hat{\sigma}^2)G_{xy}$. For fixed τ^2 and a fixed heterogeneous noise law, its average risk exceeds the task-wise oracle's average risk by $O(1/n)$ as $d, n \to \infty$ with $d/n \to \gamma \in (0, 1)$, attaining the oracle's mean gain to leading order. For $n \leq d$, a design of full row rank leaves zero residual, so this residual estimator is unavailable. Appendix C.6 gives the corresponding oracle comparisons for ridge.

We next turn from task-noise adaptation to a different limitation of scalar shrinkage: its inability to adapt to the realized eigensystem of G_x.

6 Transformer (Scalar Shrinkage) versus Ridge (Covariance Adaptation)

We now ask a different question: how much is gained by letting the estimator use the empirical covariance of the demonstrations, and how close does scalar shrinkage already get to that gain? Both scalar shrinkage and ridge use the same empirical cross-moment G_{xy}, but ridge also uses the eigensystem of G_x. This additional use of G_x is the form of **covariance adaptation** studied in this section.

6.1 Ridge benchmark

In the proportional regime $d, n \to \infty$ with $d/n \to \gamma \in (0, \infty)$ and $\nu = \tau^2/\mathbb{E}[\sigma^2]$ fixed, let μ_γ be the Marchenko-Pastur law (Marchenko & Pastur, 1967; Silverstein & Bai, 1995) and let

$$m_\gamma(z) := \int \frac{1}{x - z}\, d\mu_\gamma(x)$$

be its Stieltjes transform. Appendix Proposition D.1 gives the full asymptotic squared-bias, variance, and risk formulas for ridge. In particular, the optimal average-noise ridge penalty is $\lambda^* = 1/\nu = \mathbb{E}[\sigma^2]/\tau^2$, and the corresponding normalized risk is

$$\frac{L_{\lambda^*}}{\mathbb{E}[\sigma^2]} \to 1 + \gamma m_\gamma\Big(-\frac{\gamma}{\nu}\Big).$$

Here, for the Marchenko-Pastur law of dG_x, we use the coordinatewise independent features in Assumption 3.1(3), but this may be further relaxed.

6.2 Information-use hierarchy

The main comparison is organized by how much information each estimator uses:

$$\begin{aligned}
\hat{w}_{\mathrm{iso}} &= dG_{xy} && \text{(isotropic plug-in)},\\
\hat{w}_{\theta^*} &= \theta^* G_{xy} && \text{(optimal scalar shrinkage)},\\
\hat{w}_{\lambda^*} &= \left(G_x + \frac{\lambda^*}{n} I_d\right)^{-1} G_{xy} && \text{(covariance-adaptive ridge)}.
\end{aligned}$$

The first rule uses only the population isotropic covariance. The second adds the best scalar regularization available to the simplified transformer class; here θ^* is the population-optimal scalar from Proposition 4.1. The third adapts to the empirical eigensystem of G_x with $\lambda^* = 1/\nu$.

Theorem 6.1 (Information-use hierarchy)**.** *Under Assumption 3.1, the limiting normalized risks of the three rules above are, suppressing the common arguments (γ, ν),*

$$\big(\mathcal{R}_{\mathrm{iso}}, \mathcal{R}_{\mathrm{scalar}}, \mathcal{R}_{\mathrm{ridge}}\big) := 1 + \left(\gamma(1+\nu), \frac{\nu\gamma(1+\nu)}{\gamma(1+\nu)+\nu}, \gamma m_\gamma\Big(-\frac{\gamma}{\nu}\Big)\right),$$

Table 1: Closed forms collected from Section 6, in the proportional regime and normalized by $\mathbb{E}[\sigma^2]$. The last two rows summarize Proposition D.5 in Appendix D.5, with $\Delta_{\max}(\nu) := \sup_{\gamma>0} \Delta_{\text{adapt}}(\gamma, \nu)$ and $\gamma^*(\nu)$ a maximizing aspect ratio.

quantity	closed form	behaviour
$\mathcal{R}_{\text{iso}} - 1$	$\gamma(1+\nu)$	$\Theta(\gamma)$
$\mathcal{R}_{\text{scalar}} - 1$	$\dfrac{\nu\gamma(1+\nu)}{\gamma(1+\nu)+\nu}$	increases to ν as $\gamma \to \infty$
$\mathcal{R}_{\text{ridge}} - 1$	$\gamma\, m_\gamma(-\gamma/\nu)$	increases to ν as $\gamma \to \infty$
Δ_{reg}	$\dfrac{\gamma^2(1+\nu)^2}{\gamma(1+\nu)+\nu}$	grows linearly as $\gamma \to \infty$
Δ_{adapt}	$\dfrac{\nu\gamma(1+\nu)}{\gamma(1+\nu)+\nu} - \gamma\, m_\gamma(-\gamma/\nu)$	vanishes in all three limits
Δ_{adapt}, $\gamma \to 0$	$\nu\gamma + O(\gamma^2)$	$\Theta(\gamma)$
Δ_{adapt}, $\gamma \to \infty$	$\dfrac{\nu^4}{(1+\nu)^3}\gamma^{-2} + O(\gamma^{-3})$	$\Theta(\gamma^{-2})$
Δ_{adapt}, $\nu \to 0$	$\nu^4/\gamma^2 + O(\nu^5)$	$\Theta(\nu^4)$
$\Delta_{\max}(\nu)$	$\le 4\nu^2/27$	for every $\nu > 0$
$\Delta_{\max}(\nu)$, $\nu \to 0$	$4\nu^2/27 + O(\nu^3)$	$\gamma^*(\nu) \sim \nu/2$

where the scalar 1 *for the irreducible noise is added componentwise. Moreover, for every* $\gamma, \nu > 0$*,*

$$\mathcal{R}_{\text{ridge}}(\gamma, \nu) < \mathcal{R}_{\text{scalar}}(\gamma, \nu) < \mathcal{R}_{\text{iso}}(\gamma, \nu).$$

Theorem 6.1 is a benchmark ordering rather than a deep separation, and we present it as such. Under our moment and independence assumptions, ridge risk with $\lambda^* = 1/\nu$ minimizes prediction risk among estimators linear in the demonstration outputs, without requiring Gaussianity. It also admits a Bayes interpretation in the reference model with the corresponding Gaussian prior model with fixed noise variance $\mathbb{E}[\sigma^2]$: it is the posterior mean and minimizes Bayes prediction risk under squared loss. The substantive results of this section are the component decomposition in Theorem 6.2 and the rates in Theorem 6.3.

6.3 Risk-gap decomposition

The hierarchy separates two adjacent improvements. The scalar-regularization gain is

$$\Delta_{\text{reg}} := \mathcal{R}_{\text{iso}} - \mathcal{R}_{\text{scalar}} = \frac{\gamma^2(1+\nu)^2}{\gamma(1+\nu)+\nu},$$

which is strictly increasing in γ for fixed ν and grows linearly as $\gamma \to \infty$. The residual covariance-adaptation gain is

$$\Delta_{\text{adapt}} := \mathcal{R}_{\text{scalar}} - \mathcal{R}_{\text{ridge}} = \frac{\nu\gamma(1+\nu)}{\gamma(1+\nu)+\nu} - \gamma m_\gamma\Big(-\frac{\gamma}{\nu}\Big).$$

Let $B^2_\bullet$ and $V_\bullet$ denote limiting normalized squared bias and variance, and define the component gaps by

$$\Delta_{\text{reg},B^2} := B^2_{\text{iso}} - B^2_{\text{scalar}}, \qquad \Delta_{\text{reg},V} := V_{\text{iso}} - V_{\text{scalar}},$$
$$\Delta_{\text{adapt},B^2} := B^2_{\text{scalar}} - B^2_{\text{ridge}}, \quad \Delta_{\text{adapt},V} := V_{\text{scalar}} - V_{\text{ridge}}.$$

Theorem 6.2 (Risk-gap decomposition)**.** *Under Assumption 3.1, in the proportional regime, the limiting component gaps satisfy, for every* $\gamma, \nu > 0$*,*

$$\Delta_{\text{reg},V} > 0, \qquad \operatorname{sign}(\Delta_{\text{reg},B^2}) = \operatorname{sign}\{\gamma(1+\nu) + \nu - 1\},$$

while

$$\Delta_{\text{adapt},B^2} > 0, \qquad \Delta_{\text{adapt},V} < 0.$$

Thus scalar regularization always reduces variance, while its squared-bias effect depends on the amount of shrinkage. Covariance adaptation always reduces squared bias relative to optimized scalar shrinkage, but it increases variance. The scalar-to-ridge risk gain is therefore a tradeoff rather than a componentwise improvement.

6.4 SMALL COVARIANCE-ADAPTATION GAPS

We finally ask when the residual scalar-to-ridge gap is small. Write Δ, Δ_{B^2}, and Δ_V for Δ_{adapt}, $\Delta_{\mathrm{adapt},B^2}$, and $\Delta_{\mathrm{adapt},V}$, so that $\Delta = \Delta_{B^2} + \Delta_V$.

Theorem 6.3 (Adaptation gap rates)**.** *Under Assumption 3.1, the following componentwise expansions hold:*

$$\gamma \to 0: \quad (\Delta, \Delta_{B^2}, \Delta_V) = \big(\nu\gamma - (\nu+3)\gamma^2,\ \nu\gamma - \nu\gamma^2,\ -3\gamma^2\big) + O(\gamma^3),$$

$$\gamma \to \infty: \quad (\Delta, \Delta_{B^2}, \Delta_V) = \left(\frac{\nu^4}{(1+\nu)^3},\ \frac{\nu^4(\nu+4)}{(1+\nu)^4},\ -\frac{3\nu^4}{(1+\nu)^4}\right)\gamma^{-2} + O(\gamma^{-3}),$$

$$\nu \to 0: \quad (\Delta, \Delta_{B^2}, \Delta_V) = (1, 4, -3)\frac{\nu^4}{\gamma^2} + O(\nu^5).$$

When $\gamma = d/n \to 0$, the number of demonstrations is much larger than the dimension, so the empirical covariance is already accurate and scalar shrinkage is close to ridge. In the large-γ and small-ν regimes, the comparison exhibits a bias–variance tradeoff: ridge reduces squared bias relative to optimal scalar shrinkage but increases estimation variance. Both component gaps vanish at the rates given in Theorem 6.3, with the variance increase partially offsetting the bias reduction.

Appendix D.5 further shows that the covariance-adaptation gap is at most $4\nu^2/27$ uniformly over all aspect ratios, with this bound asymptotically sharp as $\nu \to 0$.

Figure 2 gives an empirical check beyond the simplified scalar transformer. As $\gamma = d/n$ and ν vary, the risk of the trained softmax-attention transformer generally lies between the scalar and ridge benchmarks.

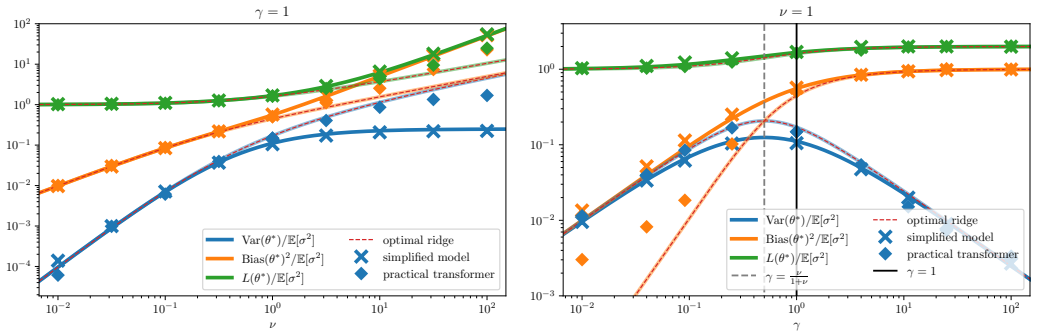

Figure 2: Scalar-transformer theory (solid curves) and finite experiment ($\times$) for normalized variance, squared bias, and risk, compared with ridge benchmarks (curves overlaid with red dashed curves) and practical-transformer comparison ($\blacklozenge$). For the practical transformer, we use a 6-layer, 2-head softmax-attention transformer trained under squared loss. Left: ν varies at $\gamma = 1$. Right: $\gamma = d/n$ varies at $\nu = 1$. See Appendix G for details.

7 LIMITATIONS AND FUTURE WORK

Our theory is intentionally stylized: sharp comparisons concern a one-layer linear-attention model with isotropic features and tasks, and the ridge comparison is asymptotic in the proportional regime. It isolates two adaptation mechanisms rather than modeling trained transformers; Appendix F relaxes isotropy and Appendix G examines multi-layer transformers.

The practical-transformer experiments compare two computable reference risks on one synthetic family; they do not identify internal mechanisms and are not systematic across depth, width, or heads. Relevance to pretrained models requires experiments with realistic tasks or data.

The task-noise oracle is a value-of-information benchmark. Appendix C.5 analyzes a residual-based plug-in estimator for $n > d$, and Appendix G.7 provides evidence consistent with partial adaptation to task-specific noise levels.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, to find related work, to draft some sections of the paper, to prove mathematical claims.

REPRODUCIBILITY STATEMENT

Our theoretical assumptions are stated alongside the results and summarized in Appendix E.1. Proofs and supplementary derivations are provided in Appendices A–D, with anisotropic extensions in Appendix F. All experiments use synthetically generated regression tasks. Appendices E–G describe the data-generating distributions, estimators, model architectures, training and evaluation protocols, and reported measures of variability. The experimental descriptions distinguish analytical benchmarks from estimated quantities and state the limitations of the numerical checks.

ETHICS STATEMENT

This work studies in-context linear regression through mathematical analysis and controlled experiments on synthetic data. It involves no human participants or personal data. The results characterize statistical behavior under stated modeling assumptions and do not establish the safety, fairness, or reliability of deployed transformers. Applications to real-world tasks require separate empirical and ethical assessment.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45614–45650, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/8ed3d610ea4b68e7afb30ea7d01422c6-Abstract-Conference.html.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211, 2023.

Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993. doi: 10.1214/aop/1176989118.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer, New York, 2 edition, 2010. doi: 10.1007/978-1-4419-0661-8.

Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 18(2):425–442, 2008.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp.

1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Liam Carroll, Jesse Hoogland, Matthew Farrugia-Roberts, and Daniel Murfet. Dynamics of transient structure in in-context linear regression transformers. *arXiv preprint arXiv:2501.17745*, 2025.

Xingwu Chen, Lei Zhao, and Difan Zou. How transformers utilize multi-head attention in in-context learning? a case study on sparse linear regression. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://openreview.net/forum?id=fG8TukiXa5.

Chen Cheng, Xinzhi Yu, Haodong Wen, Jingsong Sun, Guanzhang Yue, Yihao Zhang, and Zeming Wei. Exploring the robustness of in-context learning with noisy labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10889281. URL https://doi.org/10.1109/ICASSP49660.2025.10889281.

Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016. doi: 10.3150/14-BEJ609. URL https://doi.org/10.3150/14-BEJ609.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018. doi: 10.1214/17-AOS1549. URL https://doi.org/10.1214/17-AOS1549.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14014–14031. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.783. URL https://aclanthology.org/2023.acl-long.783/.

Hongfu Gao, Feipeng Zhang, Wenyu Jiang, Jun Shu, Feng Zheng, and Hongxin Wei. On the noise robustness of in-context learning for text generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 16569–16600, 2024. doi: 10.52202/079017-0528. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1dd8ea647d80e38b3702cf01a0855bed-Abstract-Conference.html.

Shivam Garg, Dimitris Tsipras, Percy S. Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

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

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22686–22742. PMLR, 2025. URL https://proceedings.mlr.press/v267/he25q.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23l.html.

Yingcong Li, Ankit Singh Rawat, and Samet Oymak. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=lYPAYmfQqm.

Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ElVHUWyL3n.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2304–2317. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.129. URL https://aclanthology.org/2023.acl-long.129/.

Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8p3fu56lKc.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994. URL https://www.mathnet.ru/eng/sm4101.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.527. URL https://aclanthology.org/2023.findings-acl.527/.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the Bayesian prism. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HX5ujdsSon.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of Non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/2e10b2c2e1aa4f8083c37dfe269873f8-Abstract-Conference.html.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366. PMLR, 2021. URL https://proceedings.mlr.press/v139/schlag21a.html.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Jack W. Silverstein and Zhidong Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995. doi: 10.1006/jmva.1995.1051. URL https://doi.org/10.1006/jmva.1995.1051.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Max Vladymyrov, Johannes von Oswald, Mark Sandler, and Rong Ge. Linear transformers are versatile in-context learners. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2402.14180.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc Le. Symbol tuning improves in-context learning in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 968–979, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.61. URL https://aclanthology.org/2023.emnlp-main.61/.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023b. doi: 10.48550/arXiv.2303.03846. URL https://arxiv.org/abs/2303.03846.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter L. Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vSh5ePa0ph.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2422–2437. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.155. URL https://aclanthology.org/2022.emnlp-main.155/.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL http://jmlr.org/papers/v25/23-1042.html.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Model and Background **3**

4 Bias and Variance Decomposition for Linear Transformers **5**

5 Task-Noise Heterogeneity: Risk Dispersion and Adaptation **6**

6 Transformer (Scalar Shrinkage) versus Ridge (Covariance Adaptation) **7**

6.1 Ridge benchmark 7

6.2 Information-use hierarchy 7

6.3 Risk-gap decomposition 8

6.4 Small covariance-adaptation gaps 9

7 Limitations and Future Work **9**

A Proofs and Supplementary Results for Section 3 **16**

A.1 Proof of Proposition 3.2 16

B Proofs and Supplementary Results for Section 4 **17**

B.1 The fourth-moment constant 17

B.2 Proof of Proposition 4.1 18

B.3 Empirical Estimation of the Scalar Coefficient 19

C Proofs and Supplementary Results for Section 5 **20**

C.1 Proof of Proposition 5.1 20

C.2 Proof of Proposition 5.2 20

C.3 Exact Identities for Task-Wise Oracle Risk 21

C.4 Local Expansions under Noise Heterogeneity 22

C.4.1 Monte Carlo Check of Oracle-Risk Dispersion 24

C.5 Residual-Based Estimation of Task Noise 24

C.6 Task-Noise Adaptation for Ridge 27

D Proofs and Supplementary Results for Section 6 **29**

D.1 Asymptotic Bias, Variance, and Risk of Ridge 29

D.2 Proof of Theorem 6.1 (Information-Use Hierarchy) 31

D.3 Proof of Theorem 6.2 (Risk-Gap Decomposition) 31

D.4 Proof of Theorem 6.3 (Adaptation-Gap Rates) 32

D.5 Global Bounds and Peak Behavior of the Covariance-Adaptation Gap 34

D.5.1 Relative Gaps in Total and Excess Risk 36
D.6 OLS and Ridgeless Benchmarks at the Interpolation Threshold 38

E Beyond Gaussianity: Assumptions, Universality, and Numerical Checks 38
E.1 Assumptions by Result 39
E.2 Universality of Scalar-Risk Limits 40
E.3 Universality of Spectral Limits 40
E.4 Numerical Check of Scalar Risk (Proposition 4.1) 40
E.5 Numerical Check of Scalar Optimality (Proposition 3.2) 42
E.6 Numerical Check of Task-Noise Dispersion (Proposition 5.1) 43
E.7 Scope of the Non-Gaussian Extensions 46

F Beyond Isotropy: Theory and Numerical Checks 46
F.1 Fixed-Map Optimality under Anisotropy 46
F.2 Scalar Risk and Task-Noise Dispersion under Anisotropy 48
F.3 Scalar-to-Ridge Gaps under ReLU Random Features 51
F.4 Numerical Check of Non-Scalar Fixed Maps 52
F.5 Numerical Check of Fixed-Map Optimality (Proposition F.2) 53
F.6 Numerical Check of Scalar Risk under Gaussian Anisotropy (Proposition F.4) . . . 54
F.7 Numerical check of Proposition F.4 beyond Gaussian features: the fourth moment at work 56
F.8 Numerical Check of Task-Noise Dispersion under Gaussian Anisotropy (Proposition F.5) 57
F.9 Numerical Check of Task-Noise Dispersion under Non-Gaussian Anisotropy (Proposition F.5) 58

G Multi-Layer Softmax Transformers: Experiments and Diagnostics 61
G.1 Shared Data-Generation and Evaluation Conventions 61
G.2 Model Architectures 61
G.3 Protocol for the Main-Text Experiments 61
G.4 Protocol for the Seeded Experiments 62
G.5 Recovery of the Scalar-to-Ridge Risk Gap 62
G.6 Effective-Estimator Diagnostics 63
G.7 Task-Noise Adaptation Diagnostics 64
G.8 The information: the two adaptations compound 65
G.9 Additional Parameter Sweeps 66

Organization and notation. Appendices A–D follow the order of the main-text theory (Sections 3–6). Appendix A proves the scalar-optimality result of Section 3. Appendix B derives the fourth-moment constant and the scalar risk decomposition, and analyzes finite-task empirical minimization. Appendix C develops task-noise dispersion, exact oracle identities, and local expansions with an exploratory Monte Carlo check, then treats residual-based noise estimation and noise adaptation for ridge. Appendix D proves the ridge risk formulas, information-use hierarchy, and risk-gap decomposition and rates; it also gives global bounds and peak asymptotics for the covariance-adaptation gap and records OLS and ridgeless benchmarks. Appendix E summarizes result-specific assumptions, gives conditions for proportional-limit universality, and reports **isotropic non-Gaussian** checks of the risk formula, scalar optimality, and task-noise dispersion. Appendix F extends fixed-map optimality, scalar risk, and task-noise dispersion **beyond isotropy**, with Gaussian and non-Gaussian numerical checks and a scalar-to-ridge comparison under ReLU random features. Appendix G begins with shared experimental conventions, architectures, and protocols for **multi-layer softmax transformers**, followed by seeded risk-gap comparisons, effective-estimator and task-noise adaptation diagnostics, the information 2×2, and additional parameter sweeps.

Throughout, $\xi := \sigma^2$ denotes a task's noise variance, $\bar{\xi} := \mathbb{E}[\xi] = \mathbb{E}[\sigma^2]$ its mean across tasks, $\nu = \tau^2/\bar{\xi}$ the signal-to-noise ratio, $\gamma = d/n$ the aspect ratio, and $\psi_{d,n} = n\,\mathbb{E}[\mathrm{Tr}(G_x^2)]$ the fourth-moment constant of Proposition 4.1. We also write

$$\alpha_{d,n} := \tau^2 \psi_{d,n}, \qquad \rho := 1 + \nu\psi_{d,n}, \qquad R_\gamma := \gamma(1+\nu) + \nu,$$

so that, in the isotropic setting, $\theta^* = \nu n/\rho = \tau^2 n/(\bar{\xi} + \alpha_{d,n})$. The conditional risk of a scalar rule $\hat{w} = \theta G_{xy}$ given a task's noise variance is

$$L(\theta \mid \xi) := \mathbb{E}\Big[\big(y^{(n+1)} - \theta G_{xy}^\top x^{(n+1)}\big)^2 \,\Big|\, \sigma^2 = \xi\Big],$$

so that $L(\theta) = \mathbb{E}_\xi[L(\theta \mid \xi)]$, while $L(\theta \mid \xi) - \xi$ denotes excess conditional risk, excluding irreducible query noise. Proportional-limit prediction risks normalized by $\bar{\xi}$ are written $\mathcal{R}_\bullet$ as in Section 6; thus $\mathcal{R}_\bullet - 1$ denotes normalized excess risk. The law μ_γ is the Marchenko–Pastur limit for dG_x, and m_γ is its Stieltjes transform.

A Proofs and Supplementary Results for Section 3

A.1 Proof of Proposition 3.2

Proof of Proposition 3.2. Under the reduced parameterization,

$$\hat{w} = \bar{Q}^\top G_x p + (\varphi \bar{Q}^\top + q p^\top) G_{xy} + \varphi q\, G_y,$$

so every estimator in the reduced class lies in

$$\mathcal{F} = \big\{ \hat{w}_{A,b,B,c} = A G_x b + B G_{xy} + c\, G_y : \; A, B \in \mathbb{R}^{d\times d},\; b, c \in \mathbb{R}^d \big\},$$

with $A = \bar{Q}^\top$, $b = p$, $B = \varphi\bar{Q}^\top + qp^\top$ and $c = \varphi q$. It therefore suffices to minimize $R := \mathbb{E}\|w - \hat{w}_{A,b,B,c}\|^2$ over $\mathcal{F}$, since $L = R/d + \mathbb{E}[\sigma^2]$.

By Assumption 3.1(2), w is independent of $(\bar{X}, \bar{\epsilon}, \underline{\sigma})$. Write $u := AG_x b$, which is $\bar{X}$-measurable, $v := cG_y$, and $G_{xy} = G_x w + \sigma G_{x\epsilon}$ with $G_{x\epsilon} := \bar{X}\bar{\epsilon}/n$.

Step 1: two isotropy identities. By Assumption 3.1(3), $\mathbb{E}[G_x] = \mathbb{E}[x^{(1)} x^{(1)\top}] = I_d/d$. Next,

$$\mathbb{E}[G_x^2] = \frac{1}{n^2}\Big(n\,\mathbb{E}\big[\|x^{(1)}\|^2 x^{(1)} x^{(1)\top}\big] + n(n-1)\,\frac{I_d}{d^2}\Big),$$

where the $i \neq j$ terms factor by independence into $\mathbb{E}[x^{(1)}x^{(1)\top}]^2 = I_d/d^2$. For $j \neq k$ the entry $\mathbb{E}[\|x^{(1)}\|^2 x_j^{(1)} x_k^{(1)}] = \sum_l \mathbb{E}[x_l^2 x_j x_k]$ vanishes: each summand factors over independent coordinates with at least one lone factor $\mathbb{E}[x_j] = 0$ or $\mathbb{E}[x_k] = 0$. The diagonal entries are equal because the coordinates are identically distributed. Hence $\mathbb{E}[G_x^2]$ is a multiple of I_d, and so, using $\mathbb{E}[ww^\top] = \tau^2 I_d$ (Assumption 3.1(5)), $\mathbb{E}[\epsilon^{(i)}] = 0$ and $\mathrm{Var}(\epsilon^{(i)}) = 1$,

$$\mathbb{E}\big[G_{xy} G_{xy}^\top\big] = \tau^2 \mathbb{E}[G_x^2] + \frac{\mathbb{E}[\sigma^2]}{n}\mathbb{E}[G_x] = m\, I_d, \qquad m = \frac{\tau^2}{d}\mathbb{E}[\mathrm{Tr}(G_x^2)] + \frac{\mathbb{E}[\sigma^2]}{nd} > 0.$$

Step 2: the odd moments of w decouple the directional terms. $\mathbb{E}[w^\top u] = \mathbb{E}[w]^\top \mathbb{E}[u] = 0$ because w is independent of $\bar{X}$ and $\mathbb{E}[w] = 0$. For v, expanding $G_y = n^{-1}\sum_i (w^\top x^{(i)} + \sigma\epsilon^{(i)})^2$ gives

$$\mathbb{E}[w\, G_y] = \frac{1}{d}\mathbb{E}\big[\|w\|^2 w\big] + 2\,\mathbb{E}\big[w\,\sigma\epsilon^{(i)} w^\top x^{(i)}\big] + \mathbb{E}[w]\,\mathbb{E}\big[\sigma^2(\epsilon^{(i)})^2\big] = 0,$$

the first term by the vanishing third moments of w and the second because $\epsilon^{(i)}$ is centred and independent of $(w, \bar{X}, \sigma)$. The same two facts give $\mathbb{E}[u^\top B G_{xy}] = 0$ and $\mathbb{E}[G_y G_{xy}] = 0$: grouping $y^{(j)2} y^{(i)} x^{(i)}$ by its degree in w, the cubic part is a third moment of w, the quadratic and linear parts carry a lone centred ϵ or a lone $\mathbb{E}[w]$, and the part free of w carries a lone $\mathbb{E}[x^{(i)}] = 0$. Finally

$$\mathbb{E}\big[w^\top B G_{xy}\big] = \operatorname{Tr}\big(B\,\mathbb{E}[G_x]\,\mathbb{E}[ww^\top]\big) = \frac{\tau^2}{d}\operatorname{Tr} B.$$

Step 3: minimization. Collecting Steps 1 and 2,

$$R = \mathbb{E}\|w\|^2 - \frac{2\tau^2}{d}\operatorname{Tr} B + m\|B\|_F^2 + \mathbb{E}\|u + v\|^2.$$

The last term is nonnegative and equals zero at $b = c = 0_d$. Writing $B = \theta I_d + B_\perp$ with $\operatorname{Tr} B_\perp = 0$ gives $\operatorname{Tr} B = d\theta$ and $\|B\|_F^2 = d\theta^2 + \|B_\perp\|_F^2$, so

$$R = \mathbb{E}\|w\|^2 - 2\tau^2\theta + md\,\theta^2 + m\|B_\perp\|_F^2,$$

which is strictly minimized at $B_\perp = 0$ since $m > 0$. The minimizer over θ is $\theta^* = \tau^2/(md)$, and Proposition 4.1 records it in closed form. Every minimizer of R over $\mathcal{F}$ therefore has the form $\hat{w} = \theta^* G_{xy}$, attained in the reduced parameterization by $p^* = q^* = 0_d$, $\bar{Q}^* = \kappa^* I_d$ and $\varphi^*\kappa^* = \theta^*$. □

Remark A.1 (What the proof uses). The argument uses no rotational invariance and no condition on the feature law beyond Assumption 3.1. The law of w enters only through $\mathbb{E}[ww^\top] = \tau^2 I_d$ and the vanishing of its first and third moments, both implied by symmetry about the origin; a mean-zero but skewed task prior leaves $\mathbb{E}[w\, G_y] \neq 0$, and the directional term cG_y then strictly improves on every scalar rule. Step 1 is the only place where the coordinate structure of Assumption 3.1(3) is used, and it needs two facts only: the off-diagonal entries of $\mathbb{E}[\|x\|^2 xx^\top]$ vanish and its diagonal entries are equal. Both hold whenever the law of $x^{(i)}$ is invariant under sign changes and permutations of its coordinates, whether or not the coordinates are independent. Features drawn uniformly from the sphere of radius one, or from any spherically symmetric law with $\mathbb{E}\|x\|^2 = 1$, therefore satisfy the conclusion of Proposition 3.2 although their coordinates are dependent and fall outside the letter of Assumption 3.1(3). This extension concerns Proposition 3.2 alone; which other results extend to such designs, and under which separate conditions, is recorded in Appendix E.

B Proofs and Supplementary Results for Section 4

B.1 The fourth-moment constant $\psi_{d,n}$

The finite-sample formulas of Proposition 4.1 depend on the feature law only through $\psi_{d,n} = n\,\mathbb{E}[\operatorname{Tr}(G_x^2)] = \mathbb{E}\|x^{(1)}\|^4 + (n-1)/d$. The following lemma shows that the fourth moment moves the finite-sample constants but leaves every proportional-limit statement of the paper unchanged.

Lemma B.1 (The feature law washes out of the proportional limit)**.** *Let Assumption 3.1(3) hold with $\kappa := \mathbb{E}[(z_1^{(1)})^4] < \infty$. Then $\mathbb{E}\|x^{(1)}\|^4 = 1 + (\kappa - 1)/d$, and as $d, n \to \infty$ with $d/n \to \gamma \in (0, \infty)$,*

$$\psi_{d,n} = 1 + \frac{\kappa - 1}{d} + \frac{n-1}{d} \;\longrightarrow\; 1 + \frac{1}{\gamma},$$

which does not depend on κ. Consequently $\rho = 1 + \nu\psi_{d,n} \to R_\gamma/\gamma$ and $\theta^*/d \to \nu/R_\gamma = \nu/\{\gamma(1+\nu) + \nu\}$.

Proof. With i.i.d. coordinates of mean zero and unit variance, $\mathbb{E}\|z\|^4 = d\kappa + d(d-1)$, so $\mathbb{E}\|x\|^4 = \mathbb{E}\|z\|^4/d^2 = 1 + (\kappa - 1)/d$. Substituting into the definition of $\psi_{d,n}$ and using $n/d \to 1/\gamma$ gives the limit; the expressions for ρ and θ^*/d then follow from (3). □

The same limit holds for isotropic designs with dependent coordinates whenever $\mathbb{E}\|x^{(1)}\|^4 = 1 + O(1/d)$; for features uniform on the unit sphere, $\|x^{(1)}\|^2 = 1$ exactly and $\psi_{d,n} = 1 + (n-1)/d$. Appendix E.4 checks the design dependence of $\psi_{d,n}$ numerically.

B.2 Proof of Proposition 4.1

Proof of Proposition 4.1. Step 1: conditional decomposition. Under the simplified model,

$$\hat{w} = \theta G_{xy}, \qquad G_{xy} = G_x w + \sigma G_{x\epsilon}, \qquad G_{x\epsilon} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}\epsilon^{(i)} \in \mathbb{R}^d.$$

Conditioning on $(w, \sigma, \bar{X})$ and decomposing the squared error over the demonstration noises,

$$\mathbb{E}_{\bar{\epsilon}}\big[\|w - \hat{w}\|^2 \mid w, \sigma, \bar{X}\big] = \mathrm{Tr}\big(\mathrm{Var}_{\bar{\epsilon}}(\hat{w} \mid w, \sigma, \bar{X})\big) + \big\|w - \mathbb{E}_{\bar{\epsilon}}[\hat{w} \mid w, \sigma, \bar{X}]\big\|^2.$$

By Assumption 3.1(2)–(3), $\mathbb{E}_{\bar{\epsilon}}[G_{x\epsilon} \mid w, \sigma, \bar{X}] = 0$, so $\mathbb{E}_{\bar{\epsilon}}[\hat{w} \mid w, \sigma, \bar{X}] = \theta G_x w$, and by independence of the $\epsilon^{(i)}$ across i with $\mathrm{Var}(\epsilon^{(i)}) = 1$,

$$\mathrm{Var}_{\bar{\epsilon}}(\hat{w} \mid w, \sigma, \bar{X}) = \frac{\theta^2\sigma^2}{n} G_x.$$

Hence

$$\mathbb{E}\big[\|w - \hat{w}\|^2 \mid \sigma^2 = \xi\big] = \mathbb{E}_{w,\bar{X}}\big[\|(I_d - \theta G_x)w\|^2\big] + \frac{\theta^2\xi}{n}\,\mathbb{E}[\mathrm{Tr}(G_x)], \tag{4}$$

where we used that σ is independent of $(w, \bar{X}, \bar{\epsilon})$ by Assumption 3.1(2). Since $\mathbb{E}[\mathrm{Tr}(G_x)] = \mathrm{Tr}(\mathbb{E}[G_x]) = \mathrm{Tr}(I_d/d) = 1$ by Assumption 3.1(3), averaging over ξ and dividing by d gives

$$\mathrm{Var}(\theta) = \frac{\theta^2}{dn}\,\mathbb{E}[\sigma^2], \qquad \mathrm{Bias}^2(\theta) = \frac{1}{d}\mathbb{E}_{w,\bar{X}}\big[\|(I_d - \theta G_x)w\|^2\big],$$

and the bias term does not depend on σ. Dividing (4) by d and adding the irreducible term also gives the conditional risk given the task's noise variance,

$$L(\theta \mid \xi) = \mathrm{Bias}^2(\theta) + \frac{\theta^2}{dn}\,\xi + \xi, \tag{5}$$

which is used throughout Appendix C.

Step 2: the squared bias in closed form. Since w is independent of $\bar{X}$ and $\mathbb{E}[ww^\top] = \tau^2 I_d$,

$$\mathbb{E}_w\big[\|(I_d - \theta G_x)w\|^2 \mid \bar{X}\big] = \tau^2\,\mathrm{Tr}\big((I_d - \theta G_x)^2\big) = \tau^2\left(d - 2\theta\,\mathrm{Tr}(G_x) + \theta^2\,\mathrm{Tr}(G_x^2)\right).$$

For the second moment, let x' be an independent copy of $x^{(1)}$. Then

$$\mathbb{E}[\mathrm{Tr}(G_x^2)] = \frac{1}{n^2}\Big(n\mathbb{E}\|x^{(1)}\|^4 + n(n-1)\mathbb{E}[(x^{(1)\top}x')^2]\Big) = \frac{1}{n}\mathbb{E}\|x^{(1)}\|^4 + \frac{n-1}{nd} = \frac{\psi_{d,n}}{n},$$

where the second step uses only $\mathbb{E}[(x^{(1)\top}x')^2] = \mathrm{Tr}(\mathbb{E}[x^{(1)}x^{(1)\top}]^2) = 1/d$. No property of the feature law beyond its second and fourth moments is involved; under Gaussian features $\mathbb{E}\|x^{(1)}\|^4 = 1 + 2/d$ and $\psi_{d,n} = (d + n + 1)/d$. Hence

$$\mathrm{Bias}^2(\theta) = \tau^2\left(1 - \frac{2\theta}{d} + \frac{\theta^2\psi_{d,n}}{dn}\right) = \nu\mathbb{E}[\sigma^2]\left(1 - \frac{2\theta}{d} + \frac{\theta^2\psi_{d,n}}{dn}\right).$$

Step 3: risk and optimal scaling. Adding the variance and the irreducible term,

$$L(\theta) = \mathbb{E}[\sigma^2]\left[1 + \nu - \frac{2\nu}{d}\theta + \frac{1 + \nu\psi_{d,n}}{nd}\theta^2\right] = \mathbb{E}[\sigma^2]\left(1 + \nu - \frac{2\nu}{d}\theta + \frac{\rho}{nd}\theta^2\right),$$

with $\rho = 1 + \nu\psi_{d,n} > 0$. The quadratic is strictly convex, and $\partial L/\partial\theta = 0$ gives the unique minimizer $\theta^* = \nu n/\rho$, which is (3). At the optimum,

$$L(\theta^*) = \mathbb{E}[\sigma^2]\left(1 + \nu - \frac{\nu^2 n}{d\rho}\right), \qquad \mathrm{Var}(\theta^*) = \frac{\nu^2 n\,\mathbb{E}[\sigma^2]}{d\rho^2},$$

$$\mathrm{Bias}^2(\theta^*) = \mathbb{E}[\sigma^2]\left[\nu - \frac{\nu^2 n(\rho+1)}{d\rho^2}\right].$$

Step 4: proportional limit. Suppose $d, n \to \infty$ with $d/n \to \gamma \in (0, \infty)$. By Lemma B.1, $\rho \to R_\gamma/\gamma$ with $R_\gamma = \gamma(1+\nu) + \nu$, so

$$\frac{\mathrm{Var}(\theta^*)}{\mathbb{E}[\sigma^2]} = \frac{\nu^2 n}{d\rho^2} \to \frac{\nu^2\gamma}{R_\gamma^2},$$
$$\frac{L(\theta^*)}{\mathbb{E}[\sigma^2]} = 1 + \nu - \frac{\nu^2 n}{d\rho} \to 1 + \nu - \frac{\nu^2}{R_\gamma} = 1 + \frac{\nu\gamma(1+\nu)}{R_\gamma},$$

and the normalized squared bias is the normalized risk minus the normalized variance and one:

$$\frac{\mathrm{Bias}^2(\theta^*)}{\mathbb{E}[\sigma^2]} \to \frac{\nu\gamma(1+\nu)}{R_\gamma} - \frac{\nu^2\gamma}{R_\gamma^2} = \frac{\nu\gamma\{\gamma(1+\nu)^2 + \nu^2\}}{R_\gamma^2}.$$

Step 5: monotonicity. Differentiating the one-dimensional limits,

$$\frac{d}{d\gamma}\left\{1 + \frac{\nu\gamma(1+\nu)}{R_\gamma}\right\} = \frac{\nu^2(1+\nu)}{R_\gamma^2} > 0,$$
$$\frac{d}{d\gamma}\left\{\frac{\nu^2\gamma}{R_\gamma^2}\right\} = \frac{\nu^2\{\nu - \gamma(1+\nu)\}}{R_\gamma^3},$$

so the normalized risk is strictly increasing while the normalized variance increases for $\gamma < \nu/(1+\nu)$ and decreases for $\gamma > \nu/(1+\nu)$. The derivative of the normalized squared bias is the difference of the two,

$$\frac{\nu^2\{\gamma(1+\nu)(2+\nu) + \nu^2\}}{R_\gamma^3} > 0.$$

Thus the normalized squared bias and minimum risk are strictly increasing in γ, while the normalized variance is maximized at $\gamma = \nu/(1+\nu)$. □

B.3 Empirical Estimation of the Scalar Coefficient

The results of Section 4 concern population risk. If the scalar coefficient is instead fitted by least squares on N independent training tasks, the minimizer is a ratio of sample moments whose population counterparts are the two coefficients of the quadratic risk, so it converges to θ^*; a $\sqrt{N}$ central limit theorem holds under moment conditions stronger than those of Assumption 3.1. This concerns exact minimization over the scalar, not gradient-based training of the factorized architecture, for which Appendix G gives empirical evidence only.

Proposition B.2 (Empirical scalar coefficient over N tasks)**.** *Under Assumption 3.1, let $(Z_j, Y_j)_{j=1}^N$ be i.i.d. copies of $Z := x^{(n+1)\top} G_{xy}$ and $Y := y^{(n+1)}$, one pair per training task, and let $\hat\theta_N := \arg\min_\theta N^{-1}\sum_j (Y_j - \theta Z_j)^2 = \sum_j Z_j Y_j / \sum_j Z_j^2$. Then*

$$\mathbb{E}[ZY] = \frac{\tau^2}{d}, \qquad \mathbb{E}[Z^2] = \mathbb{E}[\sigma^2]\,\frac{1 + \nu\psi_{d,n}}{nd}, \qquad \frac{\mathbb{E}[ZY]}{\mathbb{E}[Z^2]} = \theta^*,$$

with θ^* as in (3). Hence, for fixed (d, n), as $N \to \infty$, $\hat\theta_N \to \theta^*$ and $N^{-1}\sum_{j=1}^N (Y_j - \hat\theta_N Z_j)^2 \to L(\theta^*)$ almost surely. If, in addition, $\mathbb{E}\|w\|^4 < \infty$ and $\mathbb{E}\|x^{(1)}\|^8 < \infty$ (for i.i.d. coordinates, $\mathbb{E}[(z_1^{(1)})^8] < \infty$), then

$$\sqrt{N}(\hat\theta_N - \theta^*) \Rightarrow \mathcal{N}(0, V_\theta), \qquad V_\theta = \frac{\mathrm{Var}[Z(Y - \theta^* Z)]}{\mathbb{E}[Z^2]^2},$$
$$\sqrt{N}\Big\{N^{-1}\textstyle\sum_{j=1}^N (Y_j - \hat\theta_N Z_j)^2 - L(\theta^*)\Big\} \Rightarrow \mathcal{N}(0, V_L), \qquad V_L = \mathrm{Var}[(Y - \theta^* Z)^2],$$

and both variances are finite.

Proof. Write $G_{xy} = G_x w + \sigma G_{x\epsilon}$ with $G_{x\epsilon} = n^{-1}\bar X \bar\epsilon$, and $Y = w^\top x^{(n+1)} + \sigma\epsilon^{(n+1)}$. Since $x^{(n+1)}$ is independent of $(w, \sigma, \bar X, \bar\epsilon, \epsilon^{(n+1)})$ with $\mathbb{E}[x^{(n+1)} x^{(n+1)\top}] = I_d/d$, and $\epsilon^{(n+1)}$ is centred and independent of everything else,

$$\mathbb{E}[ZY] = \frac{1}{d}\,\mathbb{E}[G_{xy}^\top w] = \frac{1}{d}\,\mathbb{E}[w^\top G_x w] = \frac{\tau^2}{d}\,\mathbb{E}[\mathrm{Tr}\, G_x] = \frac{\tau^2}{d}.$$

Similarly, $\mathbb{E}[Z^2] = d^{-1}\mathbb{E}\|G_{xy}\|^2 = d^{-1}\{\tau^2\mathbb{E}[\mathrm{Tr}(G_x^2)] + \mathbb{E}[\sigma^2]\mathbb{E}[\mathrm{Tr}\, G_x]/n\}$, and $\mathbb{E}[\mathrm{Tr}(G_x^2)] = \psi_{d,n}/n$ from Step 2 of the proof of Proposition 4.1, which gives the displayed value. The ratio equals $\nu n/\{1+\nu\psi_{d,n}\} = \theta^*$. For consistency, $\mathbb{E}[Z^2] = d^{-1}\{\tau^2\mathbb{E}[\mathrm{Tr}(G_x^2)] + \mathbb{E}[\sigma^2]/n\}$ is finite under Assumption 3.1, because $\mathbb{E}[\mathrm{Tr}(G_x^2)] = \psi_{d,n}/n$ needs only $\mathbb{E}\|x^{(1)}\|^4 < \infty$, and $\mathbb{E}|ZY| \le (\mathbb{E}[Z^2]\mathbb{E}[Y^2])^{1/2} < \infty$; the strong law applied to the two sample means then gives $\hat\theta_N \to \theta^*$ and the almost-sure limit of the empirical loss. The central limit theorems need $\mathrm{Var}[Z(Y-\theta^*Z)] < \infty$ and $\mathrm{Var}[(Y-\theta^*Z)^2] < \infty$, hence finite $\mathbb{E}[Z^4]$ and $\mathbb{E}[Z^2Y^2]$, and these do *not* follow from Assumption 3.1: $Z = x^{(n+1)\top}G_x w + \sigma x^{(n+1)\top}G_{x\epsilon}$ is quadratic in the demonstration features and linear in w, so Z^4 contains eighth powers of the features and fourth powers of w, whereas Assumption 3.1 provides only $\mathbb{E}\|x^{(1)}\|^4 < \infty$ and $\mathbb{E}\|w\|^2 < \infty$. Under the additional hypotheses $\mathbb{E}\|w\|^4 < \infty$ and $\mathbb{E}\|x^{(1)}\|^8 < \infty$, repeated applications of the Cauchy–Schwarz inequality bound $\mathbb{E}[Z^4]$ and $\mathbb{E}[Z^2Y^2]$ by finite combinations of $\mathbb{E}\|w\|^4$, $\mathbb{E}\|x^{(1)}\|^8$, $\mathbb{E}[\sigma^4]$ and $\mathbb{E}[(\epsilon^{(1)})^4]$. The Lindeberg–Lévy central limit theorem and the delta method applied to the ratio of sample means then give the first statement, and the central limit theorem for $(Y-\theta^*Z)^2$ together with $\hat\theta_N \to \theta^*$ gives the second. □

C Proofs and Supplementary Results for Section 5

Except in the anisotropic extension of Appendix F.2, $L(\theta \mid \xi)$ in this appendix is the isotropic conditional risk (5) of the scalar rule $\hat w = \theta G_{xy}$ given the task's noise variance. Inserting the closed form of the squared bias from Step 2 of the proof of Proposition 4.1 and using $\tau^2\psi_{d,n} = \alpha_{d,n}$,

$$L(\theta \mid \xi) = \xi + \tau^2 - \frac{2\tau^2}{d}\theta + \frac{\alpha_{d,n}+\xi}{dn}\theta^2. \tag{6}$$

Completing the square gives, for every θ and every $\xi > 0$,

$$L(\theta \mid \xi) = L_{\mathrm{oracle}}(\xi) + \frac{\alpha_{d,n}+\xi}{dn}\big(\theta - \theta^*(\xi)\big)^2, \tag{7}$$

where

$$\theta^*(\xi) := \frac{\tau^2 n}{\alpha_{d,n}+\xi}, \qquad L_{\mathrm{oracle}}(\xi) := \min_\theta L(\theta \mid \xi) = \tau^2 + \xi - \frac{\tau^4 n}{d(\alpha_{d,n}+\xi)}.$$

Here $\theta^*(\xi)$ is the task-wise oracle coefficient, $\theta^*(\bar\xi) = \nu n/(1+\nu\psi_{d,n})$ is the population optimum (3), and $L_{\mathrm{oracle}}(\bar\xi) = L(\theta^*)$. The function $\xi \mapsto L_{\mathrm{oracle}}(\xi)$ is the task-wise oracle curve of Section 5.

C.1 Proof of Proposition 5.1

Proof of Proposition 5.1. By (5), for fixed θ the excess conditional risk $L(\theta \mid \xi) - \xi = \mathrm{Bias}^2(\theta) + \theta^2\xi/(dn)$ is affine in ξ with slope $\theta^2/(dn)$, and $L(\theta \mid \xi)$ is affine with slope $1+\theta^2/(dn)$; the squared bias does not depend on ξ because w and $\bar X$ are independent of σ. Since $\mathrm{Var}(\xi) = \mathrm{Var}(\sigma^2) < \infty$ by Assumption 3.1(4), the variance of an affine function of ξ gives both identities. At $\theta = \theta^*$, Step 1 of the proof of Proposition 4.1 gives $\mathrm{Var}(\theta^*) = \theta^{*2}\mathbb{E}[\sigma^2]/(dn)$, so $\theta^{*2}/(dn) = \mathrm{Var}(\theta^*)/\mathbb{E}[\sigma^2]$ is the normalized estimation variance. In the proportional limit, Step 4 of the same proof gives $\theta^{*2}/(dn) = \nu^2 n/(d\rho^2) \to \nu^2\gamma/R_\gamma^2$, and squaring yields the displayed limit. By Step 5, $\gamma \mapsto \nu^2\gamma/R_\gamma^2$ is maximized at $\gamma = \nu/(1+\nu)$ and strictly decreasing beyond it, and so is its square. Finally, $\nu^2\gamma/R_\gamma^2 \sim \gamma$ as $\gamma \to 0$ and $\nu^2\gamma/R_\gamma^2 \sim \nu^2/\{(1+\nu)^2\gamma\}$ as $\gamma \to \infty$, which gives the rates $\Theta(\gamma^2)$ and $\Theta(\gamma^{-2})$ for the square. □

C.2 Proof of Proposition 5.2

Proof of Proposition 5.2. For fixed $\xi = \sigma^2$, write

$$G_{xy} = G_x w + \sigma G_{x\epsilon}, \qquad G_{x\epsilon} := \frac{1}{n}\bar X\bar\epsilon.$$

Conditional on $(w, \bar{X})$,

$$\mathbb{E}_{\bar{\epsilon}}[A_\alpha G_{xy} \mid w, \bar{X}] = A_\alpha G_x w, \qquad \mathrm{Var}_{\bar{\epsilon}}(A_\alpha G_{xy} \mid w, \bar{X}) = \frac{\xi}{n} A_\alpha G_x A_\alpha^\top,$$

because A_α depends on the demonstrations only through $\bar{X}$. After integrating out the query input and using $\mathbb{E}[x^{(n+1)} x^{(n+1)\top}] = I_d/d$, the estimation-variance component of the prediction risk is therefore

$$\mathrm{Var}(\hat{w}_\alpha \mid \xi) = \frac{\xi}{dn} \mathbb{E}_{\bar{X}}\left[\mathrm{Tr}(A_\alpha G_x A_\alpha^\top)\right] = \xi C(\alpha), \qquad C(\alpha) := \frac{1}{dn} \mathbb{E}_{\bar{X}}\left[\mathrm{Tr}(A_\alpha G_x A_\alpha^\top)\right].$$

The squared-bias component is

$$B(\alpha) = \frac{1}{d} \mathbb{E}_{w,\bar{X}}\left[\|(I_d - A_\alpha G_x) w\|^2\right],$$

which does not depend on ξ because w and $\bar{X}$ are independent of σ. Hence

$$L(\alpha \mid \xi) = B(\alpha) + \xi C(\alpha) + \xi.$$

Let $\alpha^*(\xi)$ be an interior differentiable minimizer, so that $\partial_\alpha L(\alpha^*(\xi) \mid \xi) = 0$. By the chain rule,

$$\frac{d}{d\xi} L(\alpha^*(\xi) \mid \xi) = \partial_\xi L(\alpha^*(\xi) \mid \xi) + \partial_\alpha L(\alpha^*(\xi) \mid \xi) \frac{d\alpha^*(\xi)}{d\xi} = C(\alpha^*(\xi)) + 1.$$

Since $\mathrm{Var}(\hat{w}_{\alpha^*(\xi)} \mid \xi) = \xi C(\alpha^*(\xi))$, the displayed identity follows. □

For the scalar class $A_\theta = \theta I_d$ the proposition can be checked directly on (7): $L'_{\mathrm{oracle}}(\xi) - 1 = \tau^4 n / \{d(\alpha_{d,n} + \xi)^2\} = \theta^*(\xi)^2/(dn)$, which is the normalized variance of the oracle rule at level ξ. Appendix C.6 applies the same identity to ridge.

C.3 Exact Identities for Task-Wise Oracle Risk

The two propositions of Section 5 describe the estimator-attributable fluctuation at a fixed scaling and the local slope of the oracle curve. This subsection records what can be said about the oracle curve exactly, at every finite (d, n) and for every task-noise law with $\mathrm{Var}(\xi) < \infty$.

Proposition C.1 (Exact identities for the task-wise oracle)**.** *Under Assumption 3.1, write $\alpha = \alpha_{d,n}$, let ξ' be an independent copy of ξ, and define*

$$\Delta L := L_{\mathrm{oracle}}(\bar{\xi}) - \mathbb{E}[L_{\mathrm{oracle}}(\xi)] = \mathbb{E}_\xi\left[L(\theta^* \mid \xi) - L_{\mathrm{oracle}}(\xi)\right], \qquad D_{\mathrm{oracle}} := \mathrm{Var}(L_{\mathrm{oracle}}(\xi)).$$

Then

$$\Delta L = \frac{\tau^4 n}{d(\bar{\xi} + \alpha)^2} \mathbb{E}\left[\frac{(\xi - \bar{\xi})^2}{\xi + \alpha}\right] \qquad \textit{and} \qquad D_{\mathrm{oracle}} = \frac{1}{2} \mathbb{E}\left[(\xi - \xi')^2 \left(1 + \frac{\tau^4 n}{d(\xi + \alpha)(\xi' + \alpha)}\right)^2\right],$$

and both quantities vanish if and only if $\mathrm{Var}(\xi) = 0$. Moreover, with $c := \tau^4 n/d$,

$$0 \le \Delta L \le \frac{c}{\alpha(\bar{\xi} + \alpha)^2} \mathrm{Var}(\xi), \qquad \mathrm{Var}(\xi) \le D_{\mathrm{oracle}} \le \left(1 + \frac{c}{\alpha^2}\right)^2 \mathrm{Var}(\xi).$$

The second expression for ΔL says what it measures: the mean, over tasks, of the cost of using the population coefficient θ^* instead of the task's own $\theta^*(\xi)$, since $\mathbb{E}_\xi[L(\theta^* \mid \xi)] = L(\theta^*) = L_{\mathrm{oracle}}(\bar{\xi})$. The mean gap is a weighted second central moment, not a function of $\mathrm{Var}(\xi)$ alone; lower task-noise deviations receive larger weight through $(\xi + \alpha)^{-1}$. The bounds are nonlocal: they hold for skewed, dispersed, or multimodal noise laws, and give $\Delta L = O(\mathrm{Var}(\xi))$ and $(\Delta L)^2/D_{\mathrm{oracle}} \to 0$ as $\mathrm{Var}(\xi) \to 0$ at fixed $(d, n, \tau, \bar{\xi})$ without any perturbation path.

Proof of the mean-gap identity. For any $s > 0$, the pointwise identity

$$\frac{1}{s + \alpha} = \frac{1}{\bar{\xi} + \alpha} - \frac{s - \bar{\xi}}{(\bar{\xi} + \alpha)^2} + \frac{(s - \bar{\xi})^2}{(s + \alpha)(\bar{\xi} + \alpha)^2}$$

holds by direct calculation. Taking expectations with $s = \xi$ in (7), using $\mathbb{E}[\xi - \bar{\xi}] = 0$ and $L_{\mathrm{oracle}}(\bar{\xi}) = L(\theta^*)$, yields the stated formula. The equality condition follows because the integrand is strictly positive on the event $\{\xi \neq \bar{\xi}\}$. The upper bound uses $\xi + \alpha \ge \alpha$. □

Proof of the dispersion identity. The constant term τ^2 does not affect variance, so

$$D_{\text{oracle}} = \text{Var}\left(\xi - \frac{\tau^4 n}{d(\xi+\alpha)}\right).$$

Using $\text{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z-Z')^2]$ for independent copies,

$$D_{\text{oracle}} = \frac{1}{2}\mathbb{E}\left[\left\{\xi - \xi' - \frac{\tau^4 n}{d}\left(\frac{1}{\xi+\alpha} - \frac{1}{\xi'+\alpha}\right)\right\}^2\right],$$

and the expression inside braces equals

$$(\xi - \xi')\left(1 + \frac{\tau^4 n}{d(\xi+\alpha)(\xi'+\alpha)}\right),$$

which gives the identity. The multiplier in parentheses lies in $[1,\, 1 + c/\alpha^2]$, which gives the two-sided bound, and it is strictly positive, so $D_{\text{oracle}} = 0$ holds if and only if $\xi = \xi'$ almost surely, equivalently $\text{Var}(\xi) = 0$. □

Remark C.2 (Squared fluctuation around the average-noise benchmark). Since $\Delta L = L_{\text{oracle}}(\bar{\xi}) - \mathbb{E}[L_{\text{oracle}}(\xi)]$, the identity $\mathbb{E}[(Y-c)^2] = \text{Var}(Y) + (\mathbb{E}[Y]-c)^2$ with $Y = L_{\text{oracle}}(\xi)$ and $c = L_{\text{oracle}}(\bar{\xi})$ gives

$$\mathbb{E}\big[\{L_{\text{oracle}}(\xi) - L_{\text{oracle}}(\bar{\xi})\}^2\big] = D_{\text{oracle}} + (\Delta L)^2.$$

This is the quantity estimated by Monte Carlo in Figure 3; by the bounds above, the squared mean gap is a lower-order correction to the dispersion.

How to read the oracle. The task-wise oracle knows the realized ξ. It is a value-of-information benchmark inside the scalar class: ΔL is the largest mean gain that any rule which infers the noise level from the demonstrations can achieve within that class, and it is positive whenever the task noise is heterogeneous, at every aspect ratio. Whether the gain can be realized is a separate question. Appendix C.5 analyzes the natural residual-based plug-in rule exactly: when $n > d$ it attains the oracle's mean gain to leading order in the proportional limit, and when $n \le d$ the residual route is closed.

C.4 Local Expansions under Noise Heterogeneity

The remainder of this subsection extracts sharp leading coefficients from the identities above along a regular sensitivity path around homogeneous noise.

Assumption C.3 (Local perturbation of task noise)**.** For the local approximations, let $\xi_\eta = \bar{\xi} + \eta\zeta$, where $\mathbb{E}[\zeta] = 0$, $0 < \text{Var}(\zeta) < \infty$, and there exists $\eta_0 > 0$ such that $\xi_\eta > 0$ almost surely for every $|\eta| \le \eta_0$.

Assumption C.3 is a regularity condition on a sensitivity path, not on the data: it is what lets the oracle curve be expanded around homogeneous noise. The population-tuned dispersion result in Proposition 5.1 does not require it. Appendix C.4.1 compares the leading expressions with Monte Carlo calculations from the exact limiting oracle-risk curve across four task-noise families. This is an exploratory check over the plotted ranges, not a uniform approximation result.

Proposition C.4 (Local mean cost of non-adaptation)**.** *Under Assumptions 3.1 and C.3, let ΔL_η denote the mean gap computed with $\xi = \xi_\eta$. Then, as $\eta \to 0$,*

$$\Delta L_\eta = \frac{\tau^4 n}{d(\bar{\xi} + \alpha_{d,n})^3}\text{Var}(\xi_\eta) + o(\text{Var}(\xi_\eta)).$$

Consequently, with ν fixed and $d, n \to \infty$ with $d/n \to \gamma \in (0,\infty)$,

$$\lim_{d,n\to\infty}\lim_{\eta\to 0}\frac{\Delta L_\eta/\bar{\xi}}{\text{Var}(\xi_\eta)/\bar{\xi}^2} = \frac{\nu^2\gamma^2}{\{\gamma(1+\nu)+\nu\}^3}.$$

Read plainly, this says that the population-tuned scalar rule loses little in mean to the task-wise oracle. The cost is of the order of $\mathrm{Var}(\xi_\eta)$, that is $O(\eta^2)$ along the path, and its normalized coefficient $\nu^2\gamma^2/\{\gamma(1+\nu)+\nu\}^3$ is maximized at $\gamma = 2\nu/(1+\nu)$ with value $4\nu/\{27(1+\nu)^2\} \le 1/27$, vanishing as $\gamma \to 0$ and as $\gamma \to \infty$. The dispersion of Proposition C.5 below is of the same order $\mathrm{Var}(\xi_\eta)$, so the two are not separated by their order in η; what separates them is that the *standard deviation* of conditional risk is of order η whereas the mean gap is of order η^2, and that the coefficient of the mean gap is small. This is the one-scale form of the statement that heterogeneity is a dispersion rather than a mean phenomenon, and Proposition C.1 gives it without any perturbation path: its bounds imply $\Delta L/\sqrt{D_{\mathrm{oracle}}} \le c\,\mathrm{sd}(\xi)/\{\alpha(\bar\xi+\alpha)^2\}$.

Proof of Proposition C.4. Applying the mean-gap identity of Proposition C.1 with ξ replaced by ξ_η gives

$$\Delta L_\eta = \frac{\tau^4 n}{d(\bar\xi+\alpha_{d,n})^2}\mathbb{E}\left[\frac{\eta^2\zeta^2}{\xi_\eta+\alpha_{d,n}}\right].$$

For all sufficiently small $|\eta|$, Assumption C.3 gives $\xi_\eta > 0$ almost surely, so $\xi_\eta + \alpha_{d,n} \ge \alpha_{d,n} > 0$. Since $\mathbb{E}[\zeta^2] < \infty$, dominated convergence yields

$$\mathbb{E}\left[\frac{\zeta^2}{\xi_\eta+\alpha_{d,n}}\right] \to \frac{\mathbb{E}[\zeta^2]}{\bar\xi+\alpha_{d,n}}.$$

Because $\mathbb{E}[\zeta] = 0$, $\mathrm{Var}(\xi_\eta) = \eta^2\mathbb{E}[\zeta^2]$, and the claimed expansion follows. Dividing the leading term by $\bar\xi$ and normalizing by $\mathrm{Var}(\xi_\eta)/\bar\xi^2$ gives the finite-sample coefficient $\tau^4 n\bar\xi/\{d(\bar\xi+\alpha_{d,n})^3\}$. If $d/n \to \gamma$ and $\nu = \tau^2/\bar\xi$, then $\alpha_{d,n} = \tau^2\psi_{d,n} \to \nu\bar\xi(1+1/\gamma)$ by Lemma B.1, so

$$\frac{\tau^4 n\bar\xi}{d(\bar\xi+\alpha_{d,n})^3} \to \frac{\nu^2\gamma^2}{\{\gamma(1+\nu)+\nu\}^3}.$$

□

Proposition C.5 (Local oracle-risk dispersion)**.** *Under Assumptions 3.1 and C.3, as $\eta \to 0$,*

$$\mathrm{Var}(L_{\mathrm{oracle}}(\xi_\eta)) = \left(1+\frac{\tau^4 n}{d(\bar\xi+\alpha_{d,n})^2}\right)^2 \mathrm{Var}(\xi_\eta) + o(\mathrm{Var}(\xi_\eta)).$$

Consequently, with ν fixed and $d, n \to \infty$ with $d/n \to \gamma \in (0,\infty)$,

$$\lim_{d,n\to\infty}\lim_{\eta\to 0}\frac{\mathrm{Var}(L_{\mathrm{oracle}}(\xi_\eta))/\bar\xi^2}{\mathrm{Var}(\xi_\eta)/\bar\xi^2} = \left(1+\frac{\nu^2\gamma}{\{\gamma(1+\nu)+\nu\}^2}\right)^2.$$

The term beyond one inside the parentheses is exactly the normalized variance of Proposition 4.1 at the optimum, $\tau^4 n/\{d(\bar\xi+\alpha_{d,n})^2\} = \theta^{*2}/(dn)$. Therefore, locally,

$$\mathrm{Var}[L_{\mathrm{oracle}}(\xi_\eta)] \approx \left(\frac{\mathrm{Var}(\hat w_{\theta^*(\bar\xi)} \mid \bar\xi)}{\bar\xi}+1\right)^2 \mathrm{Var}(\xi_\eta),$$

which is the delta-method statement of Section 5 with the coefficient identified by Proposition 5.2.

Proof of Proposition C.5. Let ζ' be an independent copy of ζ and set $\xi'_\eta = \bar\xi + \eta\zeta'$. The dispersion identity of Proposition C.1 with ξ replaced by ξ_η gives

$$\mathrm{Var}(L_{\mathrm{oracle}}(\xi_\eta)) = \frac{\eta^2}{2}\mathbb{E}\left[(\zeta-\zeta')^2\left(1+\frac{\tau^4 n}{d(\xi_\eta+\alpha_{d,n})(\xi'_\eta+\alpha_{d,n})}\right)^2\right].$$

For all sufficiently small $|\eta|$, both ξ_η and ξ'_η are positive almost surely. Hence the squared multiplier is bounded by $(1+\tau^4 n/(d\alpha_{d,n}^2))^2$ and converges almost surely to $(1+\tau^4 n/\{d(\bar\xi+\alpha_{d,n})^2\})^2$. Dominated convergence therefore gives

$$\frac{\mathrm{Var}(L_{\mathrm{oracle}}(\xi_\eta))}{\eta^2} \to \frac{1}{2}\mathbb{E}[(\zeta-\zeta')^2]\left(1+\frac{\tau^4 n}{d(\bar\xi+\alpha_{d,n})^2}\right)^2.$$

Since $\frac{1}{2}\mathbb{E}[(\zeta-\zeta')^2] = \mathrm{Var}(\zeta)$ and $\mathrm{Var}(\xi_\eta) = \eta^2\mathrm{Var}(\zeta)$, the result follows. The substitutions $d/n \to \gamma$, $\tau^2 = \nu\bar{\xi}$, and $\alpha_{d,n} \to \nu\bar{\xi}(1+1/\gamma)$ give

$$\frac{\tau^4 n}{d(\bar{\xi}+\alpha_{d,n})^2} \to \frac{\nu^2\gamma}{\{\gamma(1+\nu)+\nu\}^2},$$

which yields the proportional-limit coefficient. □

C.4.1 Monte Carlo Check of Oracle-Risk Dispersion

The local dispersion result of Proposition C.5 is derived for perturbations ξ_η around the fixed average-noise level $\bar{\xi}$. Figure 3 checks whether the same leading coefficient remains informative when the task-noise variance ξ is sampled directly from several families. It evaluates the proportional-limit scalar oracle curve $f(\xi) = \xi + (\xi+1)/(\xi+2)$ at $\gamma = \tau^2 = 1$, so $f'(\xi) = 1+1/(\xi+2)^2$. The Monte Carlo calculations approximate expectations of this exact limiting curve, not a finite-(d,n) oracle formula; L_{oracle} below denotes this limiting curve. For the uniform, exponential, and Gamma panels, it compares Monte Carlo estimates of

$$\mathbb{E}\big[\{L_{\mathrm{oracle}}(\xi) - L_{\mathrm{oracle}}(\bar{\xi})\}^2\big] = D_{\mathrm{oracle}} + (\Delta L)^2 \quad \text{and} \quad D_{\mathrm{oracle}} = \mathrm{Var}(L_{\mathrm{oracle}}(\xi))$$

(Remark C.2) with the plug-in approximation

$$\left(\frac{\mathrm{Var}(\hat{w}_{\theta^*(\bar{\xi})} \mid \bar{\xi})}{\bar{\xi}} + 1\right)^2 \mathrm{Var}(\xi),$$

for $\sigma^2 \sim \mathrm{Unif}(0,B)$, $\mathrm{Exp}(\lambda)$, $\mathrm{Gamma}(\alpha, \mathrm{scale}=1)$, and χ^2_k, with 10,000 sampled noise variances per marker. The saved plotting notebook uses 100 linearly spaced B values on $[0.1, 10]$, 50 logarithmically spaced rates $\lambda \in [0.1, 2]$ and shapes $\alpha \in [1, 10]$, and 50 linearly spaced degrees of freedom $k \in [2, 50]$. In the first three panels the blue statistic centers $f(\xi)$ at $f(\hat{\bar{\xi}})$, using the sample mean $\hat{\bar{\xi}}$. The chi-square panel is different: its blue statistic in that notebook is $\widehat{\mathbb{E}}[f'(\xi)^2]\,\widehat{\mathrm{Var}}(\xi)$, despite the inherited squared-fluctuation legend. It must not be interpreted as the same blue quantity as in the first three panels. The red statistic is the sample variance of $f(\xi)$ in all four panels, and the green curve is $f'(\bar{\xi})^2\mathrm{Var}(\xi)$. These are exploratory comparisons over the plotted ranges; the local expansion supplies no uniform approximation guarantee over these families.

C.5 Residual-Based Estimation of Task Noise

Section 5 summarizes how residual-based estimation of task noise can approach the task-wise oracle's mean risk. This subsection provides the exact plug-in risk identity and its proportional-limit consequence under the conditions stated below. The population-tuned rule $\theta^* = \theta^*(\bar{\xi})$ already presupposes the constants τ^2 and $\bar{\xi}$; the rule below keeps τ^2 and replaces $\bar{\xi}$ by the classical residual estimate of the task's own noise variance, so that only what varies across tasks is estimated.

Fix one task and write $X := \bar{X}^\top \in \mathbb{R}^{n\times d}$ for its design matrix, with rows $x^{(i)\top}$, so that $\bar{Y}^\top = Xw + \sigma\bar{\epsilon}$, $G_x = X^\top X/n$ and $G_{xy} = X^\top\bar{Y}^\top/n$. When $n > d$ and X has full column rank, let $P_X := X(X^\top X)^{-1}X^\top$ and define

$$\hat{\sigma}^2 := \frac{\|(I_n - P_X)\bar{Y}^\top\|^2}{n-d}, \qquad \hat{w}_{\mathrm{plug}} := \theta^*(\hat{\sigma}^2)\, G_{xy}, \qquad \theta^*(s) = \frac{\tau^2 n}{\alpha_{d,n}+s},$$

and let $L_{\mathrm{plug}}(\xi) := \mathbb{E}[(y^{(n+1)} - \hat{w}_{\mathrm{plug}}^\top x^{(n+1)})^2 \mid \sigma^2 = \xi]$ be its conditional risk. Write also

$$P(\xi) := L(\theta^* \mid \xi) - L_{\mathrm{oracle}}(\xi) = \frac{\tau^4 n\,(\xi-\bar{\xi})^2}{d(\alpha_{d,n}+\bar{\xi})^2(\alpha_{d,n}+\xi)}$$

for the cost of population tuning at a task with noise variance ξ, which is (7) at $\theta = \theta^*$; by Proposition C.1, $\mathbb{E}_\xi[P(\xi)] = \Delta L$.

Proposition C.6 (Residual-based plug-in adaptation)**.** *Let Assumption 3.1 hold with Gaussian demonstration noise $\epsilon^{(i)} \sim \mathcal{N}(0,1)$, let $n > d$, and suppose X has full column rank almost surely. Write $\alpha = \alpha_{d,n}$ and $k := n - d$.*

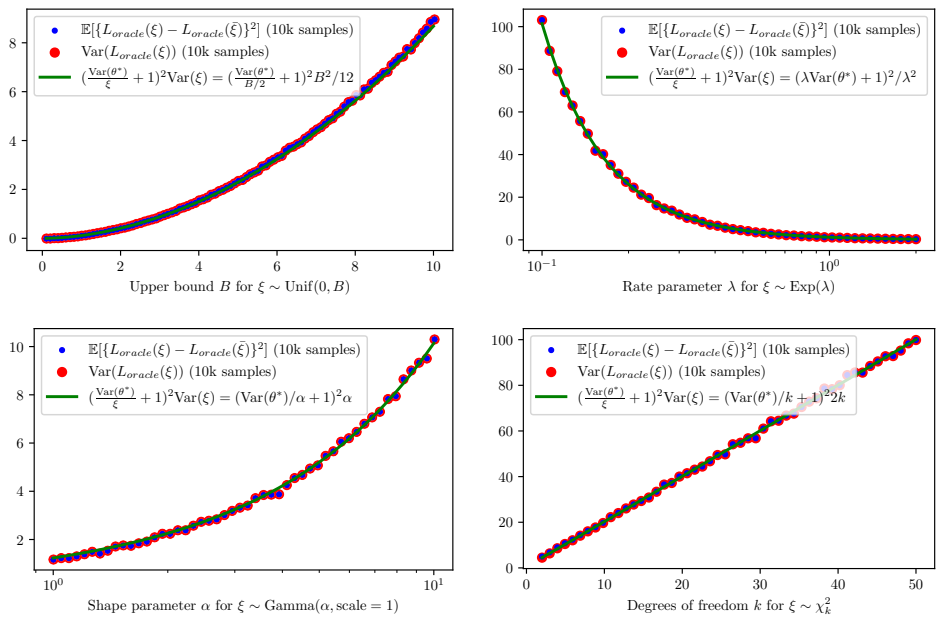

Figure 3: Monte Carlo check of the squared risk fluctuation and oracle-risk dispersion for four distributions $\sigma^2 \sim \mathrm{Unif}(0, B)$, $\mathrm{Exp}(\lambda)$, $\mathrm{Gamma}(\alpha, \mathrm{scale} = 1)$, χ^2_k. The limiting oracle curve $f(\xi) = \xi + (\xi + 1)/(\xi + 2)$ is used. Blue markers in the first three panels use the sample-mean-centered squared fluctuation of $f(\xi)$; in the chi-square panel they instead use $\widehat{\mathbb{E}}[f'(\xi)^2]\,\widehat{\mathrm{Var}}(\xi)$, correcting the inherited legend's interpretation. Red markers estimate $D_{\mathrm{oracle}} = \mathrm{Var}(L_{\mathrm{oracle}}(\xi))$, and green curves plot the leading approximation in Proposition C.5. Each marker is based on 10,000 sampled task-noise variances, and $\gamma = \tau^2 = 1$.

1. *Conditional on (X, w, σ), $k\hat{\sigma}^2/\sigma^2 \sim \chi^2_k$; in particular $\mathbb{E}[\hat{\sigma}^2 \mid X, w, \sigma] = \sigma^2$ and $\mathrm{Var}(\hat{\sigma}^2 \mid X, w, \sigma) = 2\sigma^4/k$. Moreover, conditional on σ, $\hat{\sigma}^2$ is independent of (X, w, G_{xy}).*

2. *For every $\xi > 0$,*

$$C(\xi) := L_{\mathrm{plug}}(\xi) - L_{\mathrm{oracle}}(\xi) = \frac{\tau^4 n}{d(\alpha + \xi)}\,\mathbb{E}\left[\frac{(\hat{\sigma}^2 - \xi)^2}{(\alpha + \hat{\sigma}^2)^2}\right],$$

where the expectation is over $\hat{\sigma}^2 = \xi V/k$ with $V \sim \chi^2_k$.

3. *$C(\xi) > 0$, and*

$$C(\xi) \le \frac{2\tau^4 n\,\xi^2}{d\,k\,\alpha^2(\alpha + \xi)},$$
$$C(\xi) = \frac{2\tau^4 n\,\xi^2}{d\,k\,(\alpha + \xi)^3}\big(1 + O(k^{-1})\big),$$

where the constant in $O(k^{-1})$ depends only on α/ξ.

4. *The plug-in rule has smaller conditional risk than the population-tuned rule at a task with noise variance ξ if and only if $C(\xi) < P(\xi)$, and smaller average risk if and only if $\mathbb{E}_\xi[C(\xi)] < \Delta L$.*

Proof. Item 1. Conditional on X, the vector $(I_n - P_X)\bar{\epsilon}$ is Gaussian with covariance $I_n - P_X$, an orthogonal projection of rank k, so $\|(I_n - P_X)\bar{\epsilon}\|^2 \sim \chi^2_k$, and $(I_n - P_X)\bar{Y}^\top = \sigma(I_n - P_X)\bar{\epsilon}$ because $(I_n - P_X)X = 0$. This gives the law of $k\hat{\sigma}^2/\sigma^2$ and its first two moments. The vectors $P_X\bar{\epsilon}$ and $(I_n - P_X)\bar{\epsilon}$ are jointly Gaussian given X with cross-covariance $P_X(I_n - P_X) = 0$, hence independent given X. Since $X^\top = X^\top P_X$, the cross-moment $G_{xy} = (X^\top X w + \sigma X^\top P_X\bar{\epsilon})/n$ is a function

of $(X, w, \sigma, P_X\bar{\epsilon})$, so $\hat{\sigma}^2$ and G_{xy} are independent conditional on (X, w, σ). Finally, the conditional law of $\hat{\sigma}^2$ given (X, w, σ) depends on σ alone, so for bounded f and g, $\mathbb{E}[f(\hat{\sigma}^2)g(X, w, G_{xy}) \mid \sigma] = \mathbb{E}\big[\mathbb{E}[f(\hat{\sigma}^2) \mid X, w, \sigma]\,\mathbb{E}[g(X, w, G_{xy}) \mid X, w, \sigma] \mid \sigma\big] = \mathbb{E}[f(\hat{\sigma}^2) \mid \sigma]\,\mathbb{E}[g(X, w, G_{xy}) \mid \sigma]$, which is the claimed conditional independence.

Item 2. Write $\hat{\theta} := \theta^*(\hat{\sigma}^2)$. The query pair $(x^{(n+1)}, \epsilon^{(n+1)})$ is independent of everything else, so, as in Section 3, $L_{\mathrm{plug}}(\xi) = d^{-1}\mathbb{E}[\|w - \hat{\theta}G_{xy}\|^2 \mid \xi] + \xi$. By item 1, $\hat{\theta}$ is independent of (w, G_{xy}) conditional on $\sigma^2 = \xi$, so conditioning on $\hat{\theta}$ first,

$$\mathbb{E}\big[\|w - \hat{\theta}G_{xy}\|^2 \mid \xi\big] = \mathbb{E}\big[h(\hat{\theta}) \mid \xi\big], \qquad h(\theta) := \mathbb{E}\big[\|w - \theta G_{xy}\|^2 \mid \xi\big] = d\{L(\theta \mid \xi) - \xi\},$$

the last equality being (5). Hence $L_{\mathrm{plug}}(\xi) = \mathbb{E}[L(\hat{\theta} \mid \xi)]$, and (7) gives

$$L_{\mathrm{plug}}(\xi) - L_{\mathrm{oracle}}(\xi) = \frac{\alpha + \xi}{dn}\,\mathbb{E}\big[(\theta^*(\hat{\sigma}^2) - \theta^*(\xi))^2\big].$$

Since $\theta^*(\hat{\sigma}^2) - \theta^*(\xi) = \tau^2 n(\xi - \hat{\sigma}^2)/\{(\alpha + \hat{\sigma}^2)(\alpha + \xi)\}$, squaring and simplifying yields the displayed formula, and item 1 gives the law of $\hat{\sigma}^2$.

Item 3. The integrand is positive unless $\hat{\sigma}^2 = \xi$, an event of probability zero, so $C(\xi) > 0$. The bound follows from $(\alpha + \hat{\sigma}^2)^2 \geq \alpha^2$ and $\mathbb{E}[(\hat{\sigma}^2 - \xi)^2] = 2\xi^2/k$. For the expansion, substitute $\hat{\sigma}^2 = \xi u$ with $u = V/k$ and $a := \alpha/\xi$:

$$\frac{(\hat{\sigma}^2 - \xi)^2}{(\alpha + \hat{\sigma}^2)^2} = \frac{\xi^2}{(\alpha + \xi)^2}\,h_a(u), \qquad h_a(u) := \frac{(a+1)^2(u-1)^2}{(a+u)^2}.$$

The function h_a is smooth on $[0, \infty)$ with bounded derivatives of every order, $h_a(1) = h_a'(1) = 0$ and $h_a''(1) = 2$. Since $\mathbb{E}[(u-1)^2] = 2/k$, $\mathbb{E}[(u-1)^3] = 8/k^2$, and $\mathbb{E}[(u-1)^4] = 12/k^2 + 48/k^3$ for $V \sim \chi^2_k$, a third-order Taylor expansion of h_a around 1 with a remainder bounded by a constant times $(u-1)^4$ gives $\mathbb{E}[h_a(u)] = 2/k + O(k^{-2})$, with constants depending on a only. Multiplying by $\tau^4 n\xi^2/\{d(\alpha + \xi)^3\}$ gives the expansion.

Item 4. Immediate from the definitions of C and P and from $\mathbb{E}_\xi[P(\xi)] = \Delta L$. □

The identity in item 2 has a simple reading. Because the residual is independent of the cross-moment, plugging an estimate of the noise level into the oracle coefficient costs exactly the curvature of the conditional risk, $(\alpha+\xi)/(dn)$, times the mean squared error of the resulting coefficient, and nothing else; no bias is introduced by the dependence between the estimate and the data, since there is none. The cost is a pure estimation-variance term, of order $1/k$ in the residual degrees of freedom, whereas the cost of population tuning $P(\xi)$ is a squared-bias term of order $(\xi - \bar{\xi})^2$. Comparing the two leading terms, the plug-in rule improves on population tuning at a task with noise variance ξ when

$$\frac{|\xi - \bar{\xi}|}{\alpha + \bar{\xi}} \gtrsim \sqrt{\frac{2}{n-d}}\,\frac{\xi}{\alpha + \xi},$$

that is, when the task's relative deviation from the average noise level exceeds the relative standard error of the residual variance estimate.

Corollary C.7 (Proportional limit)**.** *Under the conditions of Proposition C.6, let $d, n \to \infty$ with $d/n \to \gamma \in (0, 1)$ and ν fixed. Then $\sup_{\xi \leq K} C(\xi) = O(1/n)$ for every $K < \infty$, and $\mathbb{E}_\xi[C(\xi)] \to 0$, whereas*

$$\Delta L \;\longrightarrow\; \frac{\tau^4}{\gamma(\bar{\xi} + \alpha_\infty)^2}\,\mathbb{E}\left[\frac{(\xi - \bar{\xi})^2}{\xi + \alpha_\infty}\right] > 0, \qquad \alpha_\infty := \tau^2\Big(1 + \frac{1}{\gamma}\Big),$$

whenever $\mathrm{Var}(\xi) > 0$. *Consequently* $\lim \mathbb{E}_\xi[L_{\mathrm{plug}}(\xi)] = \lim \mathbb{E}_\xi[L_{\mathrm{oracle}}(\xi)] < \lim L(\theta^*)$*: the plug-in rule attains the oracle's mean gain to leading order.*

Proof. By Lemma B.1, $\alpha_{d,n} \to \alpha_\infty > 0$, so $\alpha_{d,n}$ is bounded below by a positive constant for all large n. The bound in item 3 of Proposition C.6 then gives $C(\xi) \leq c_0\,\xi^2\,n/\{d(n-d)\}$ for a constant c_0, and $n/\{d(n-d)\} = (n/d)/(n-d) \to 0$ at rate $1/n$ because $d/n \to \gamma < 1$; this proves the uniform statement and, since $\mathbb{E}[\xi^2] < \infty$ by Assumption 3.1(4), $\mathbb{E}_\xi[C(\xi)] \to 0$. For ΔL, apply Proposition C.1: the prefactor converges because $n/d \to 1/\gamma$, and the expectation converges by dominated convergence with dominating function $(\xi - \bar{\xi})^2/\inf_n \alpha_{d,n}$. The limit is positive when $\mathrm{Var}(\xi) > 0$, and $\lim L(\theta^*) - \lim \mathbb{E}[L_{\mathrm{oracle}}(\xi)] = \lim \Delta L$ gives the last statement. □

Table 2: Residual-based plug-in adaptation against population tuning, from the exact identities of Proposition C.6, with Gaussian features, $\tau^2 = 1$ and $\bar{\xi} = 1$ (so $\nu = 1$). $\Delta L = \mathbb{E}_\xi[P(\xi)]$ is the mean cost of population tuning relative to the task-wise oracle, $\mathbb{E}_\xi[C(\xi)]$ the mean cost of the plug-in rule relative to the same oracle, and the fraction is the share of tasks at which the plug-in rule has the smaller conditional risk. The left block uses $\xi \sim \mathrm{Unif}(0, 2)$, the task-noise law of Figure 1; the right block uses the two-point law $\xi \in \{0.1, 1.9\}$ with equal probability, a bimodal mixture of clean and noisy tasks with the same mean. All entries are normalized by $\bar{\xi} = 1$ and computed by numerical integration.

			$\xi \sim \mathrm{Unif}(0,2)$			$\xi \in \{0.1, 1.9\}$	
(n, d)	γ	$n - d$	ΔL	$\mathbb{E}_\xi[C(\xi)]$	plug-in better	ΔL	$\mathbb{E}_\xi[C(\xi)]$
$(200, 10)$	0.05	190	0.0006	$< 10^{-4}$	89%	0.0015	$< 10^{-4}$
$(40, 10)$	0.25	30	0.0060	0.0012	73%	0.0146	0.0014
$(20, 10)$	0.5	10	0.0100	0.0054	54%	0.0247	0.0055
$(40, 20)$	0.5	20	0.0104	0.0029	68%	0.0257	0.0029
$(80, 40)$	0.5	40	0.0106	0.0015	77%	0.0262	0.0015
$(160, 80)$	0.5	80	0.0107	0.0008	84%	0.0264	0.0008
$(20, 15)$	0.75	5	0.0119	0.0115	33%	0.0295	0.0116
$(20, 18)$	0.9	2	0.0124	0.0250	25%	0.0308	0.0258

Reading Table 2. Two features stand out. First, the plug-in cost halves each time n doubles at fixed $\gamma = 1/2$ (the four middle rows), which is the $O(1/n)$ rate of Corollary C.7, while ΔL stays put; at $(n, d) = (20, 10)$ the plug-in rule already recovers about half of the oracle's mean gain, and at $(160, 80)$ more than 90%. Second, the plug-in cost is insensitive to the shape of the task-noise law, whereas ΔL grows with its dispersion: for the bimodal law the oracle's mean gain is 2.5 times larger and the plug-in rule recovers about four fifths of it at $(n, d) = (20, 10)$. This is the direction in which the exact identities matter and the local expansions of Appendix C.4 do not: the value of noise information is a nonlocal, second-central-moment quantity, and a rule that estimates the noise level pays only a variance price that does not depend on how heterogeneous the tasks are. However, the plug-in rule can underperform population tuning when the residual degrees of freedom $n - d$ are small, as in the uniform-noise setting at $(n, d) = (20, 18)$.

The boundary $n \le d$, and what the experiments add. When $n \le d$ and X has full row rank, $P_X = I_n$, the fixed-effects fit interpolates and the residual vanishes; residual-based estimation of ξ is unavailable without further structure. Under a Gaussian random-effects likelihood $\bar{Y}^\top \mid X, \sigma^2 \sim \mathcal{N}(0, \tau^2 X X^\top + \sigma^2 I_n)$ the noise variance remains identifiable for $n \le d$, because it shifts every eigenvalue of the covariance, but one prompt supplies a single n-vector, the precision can be poor, and the argument needs both known τ^2 and Gaussian noise. We do not claim an identifiability rate under Assumption 3.1 in that regime. Accordingly, ΔL is positive at every aspect ratio, including the overparameterized regime where the residual route to attaining it is closed, and whether a trained architecture attains part of it there is an empirical question. Appendix G.7 reports, at $\gamma = 1/2$, that a softmax-attention transformer trained on a mixture of noise levels lies below the task-wise oracle scalar at low noise and above it at high noise; whether this reflects use of the noise level or covariance adaptation alone is examined there against ridge references.

C.6 Task-Noise Adaptation for Ridge

The following identities are the ridge analogues of Proposition C.1 and of the local expansions above. They show that the limitation studied in Section 5 is a property of tuning once across tasks and not of the scalar class. They are functionals of the limiting task-wise oracle-risk curve of ridge,

$$L^{\mathrm{ridge}}_{\mathrm{oracle}}(\xi) = \xi \left\{ 1 + \gamma m_\gamma \left(-\frac{\gamma \xi}{\tau^2} \right) \right\},$$

and are not meant as uniform finite-sample limits for random oracle penalties without additional tail conditions.

Proposition C.8 (Ridge analogue of the noise-adaptation identities)**.** *In the proportional regime of Section 6, under Assumption 3.1, define the ridge mean non-adaptation gap for the limiting oracle-risk curve by*

$$\Delta L_{\mathrm{ridge}} := L^{\mathrm{ridge}}_{\mathrm{oracle}}(\bar{\xi}) - \mathbb{E}\Big[L^{\mathrm{ridge}}_{\mathrm{oracle}}(\xi)\Big].$$

Then

$$\Delta L_{\mathrm{ridge}} = \tau^2 \int \mathbb{E}\left[\frac{(\tau^2 x/\gamma)(\xi-\bar{\xi})^2}{(\bar{\xi}+\tau^2 x/\gamma)^2(\xi+\tau^2 x/\gamma)}\right] d\mu_\gamma(x).$$

In particular, $\Delta L_{\mathrm{ridge}} \geq 0$*, with equality if and only if* $\mathrm{Var}(\xi) = 0$*. Let* ξ' *be an independent copy of* ξ *and define the ridge oracle-risk dispersion* $D_{\mathrm{ridge}} := \mathrm{Var}(L^{\mathrm{ridge}}_{\mathrm{oracle}}(\xi))$*. Then*

$$D_{\mathrm{ridge}} = \frac{1}{2}\mathbb{E}\left[(\xi-\xi')^2\left(1+\tau^2\int \frac{\tau^2 x/\gamma}{(\xi+\tau^2 x/\gamma)(\xi'+\tau^2 x/\gamma)} d\mu_\gamma(x)\right)^2\right],$$

and $D_{\mathrm{ridge}} = 0$ *if and only if* $\mathrm{Var}(\xi) = 0$*.*

Proof. The conditional limiting risk follows from the proof of Proposition D.1 by replacing $\mathbb{E}[\sigma^2]$ with the fixed task-noise variance ξ:

$$L_{\mathrm{ridge}}(\lambda \mid \xi) = \xi\{1+\gamma[m_\gamma(-\gamma\lambda) - \gamma\lambda m'_\gamma(-\gamma\lambda)]\} + \tau^2(\gamma\lambda)^2 m'_\gamma(-\gamma\lambda).$$

Let $t = \gamma\lambda$. Differentiating $L_{\mathrm{ridge}}(t/\gamma \mid \xi)$ with respect to t gives

$$\frac{d}{dt}L_{\mathrm{ridge}}(t/\gamma \mid \xi) = (\tau^2 t - \xi\gamma)\left\{2m'_\gamma(-t) - t m''_\gamma(-t)\right\},$$
$$2m'_\gamma(-t) - t m''_\gamma(-t) = 2\int \frac{x}{(x+t)^3}\, d\mu_\gamma(x) > 0$$

for $t > 0$, because the Marchenko–Pastur law has positive mass on $(0,\infty)$. Hence the unique minimizer satisfies $t = \xi\gamma/\tau^2$, or $\lambda^*(\xi) = \xi/\tau^2$, and the oracle-risk curve is $L^{\mathrm{ridge}}_{\mathrm{oracle}}(\xi) = \xi\{1+\gamma m_\gamma(-\gamma\xi/\tau^2)\}$. Equivalently, writing $\beta_x := \tau^2 x/\gamma$,

$$L^{\mathrm{ridge}}_{\mathrm{oracle}}(\xi) = \xi + \tau^2 \int \frac{\xi}{\xi+\beta_x}\, d\mu_\gamma(x).$$

The linear term ξ has no Jensen gap, so

$$\Delta L_{\mathrm{ridge}} = \tau^2 \int \left\{\frac{\bar{\xi}}{\bar{\xi}+\beta_x} - \mathbb{E}\left[\frac{\xi}{\xi+\beta_x}\right]\right\} d\mu_\gamma(x).$$

For each $\beta_x \geq 0$,

$$\frac{\bar{\xi}}{\bar{\xi}+\beta_x} - \mathbb{E}\left[\frac{\xi}{\xi+\beta_x}\right] = \mathbb{E}\left[\frac{\beta_x(\xi-\bar{\xi})^2}{(\bar{\xi}+\beta_x)^2(\xi+\beta_x)}\right],$$

where the case $\beta_x = 0$ is immediate and, for $\beta_x > 0$, the identity follows from

$$\mathbb{E}\left[\frac{(\xi-\bar{\xi})^2}{\xi+\beta_x}\right] = (\bar{\xi}+\beta_x)^2\mathbb{E}\left[\frac{1}{\xi+\beta_x}\right] - (\bar{\xi}+\beta_x)$$

together with $\mathbb{E}[\xi] = \bar{\xi}$. This form keeps the factor β_x inside the expectation, so the atom at zero in μ_γ when $\gamma > 1$ contributes zero without requiring a negative moment of ξ. Substituting $\beta_x = \tau^2 x/\gamma$ proves the stated formula for $\Delta L_{\mathrm{ridge}}$. The integrand is nonnegative, and it is strictly positive on the positive part of the Marchenko–Pastur support whenever $\mathrm{Var}(\xi) > 0$, which gives the equality condition.

For the dispersion identity, for any $s, t > 0$,

$$L^{\mathrm{ridge}}_{\mathrm{oracle}}(s) - L^{\mathrm{ridge}}_{\mathrm{oracle}}(t) = (s-t)\left\{1+\tau^2\int \frac{\beta_x}{(s+\beta_x)(t+\beta_x)} d\mu_\gamma(x)\right\}.$$

Using $\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z-Z')^2]$ and substituting $s = \xi$, $t = \xi'$ gives the displayed dispersion formula. The multiplier in braces is at least one, so the dispersion is zero if and only if $\xi = \xi'$ almost surely, equivalently $\mathrm{Var}(\xi) = 0$. □

Corollary C.9 (Local ridge noise-adaptation cost)**.** *Under the conditions of Proposition C.8 and Assumption C.3, let $\Delta L_{\mathrm{ridge},\eta}$ and $D_{\mathrm{ridge},\eta}$ denote the mean gap and ridge oracle-risk dispersion in Proposition C.8 computed under ξ_η. Then, as $\eta \to 0$,*

$$\frac{\Delta L_{\mathrm{ridge},\eta}}{\bar{\xi}} = \left\{ \frac{\nu^2}{\gamma} \int \frac{x}{(1+\nu x/\gamma)^3} d\mu_\gamma(x) \right\} \frac{\mathrm{Var}(\xi_\eta)}{\bar{\xi}^2} + o\left(\frac{\mathrm{Var}(\xi_\eta)}{\bar{\xi}^2} \right),$$

and

$$\frac{D_{\mathrm{ridge},\eta}}{\bar{\xi}^2} = \left(1 + \frac{\nu^2}{\gamma} \int \frac{x}{(1+\nu x/\gamma)^2} d\mu_\gamma(x) \right)^2 \frac{\mathrm{Var}(\xi_\eta)}{\bar{\xi}^2} + o\left(\frac{\mathrm{Var}(\xi_\eta)}{\bar{\xi}^2} \right),$$

where $\nu = \tau^2/\bar{\xi}$.

The term added to one in the dispersion coefficient is the limiting normalized variance of optimally tuned ridge, $V_{\mathrm{ridge}} = \gamma\{m_\gamma(-\gamma/\nu) - (\gamma/\nu)m'_\gamma(-\gamma/\nu)\}$ of Proposition D.1, in agreement with Proposition 5.2; since $V_{\mathrm{ridge}} > V_{\mathrm{scalar}}$ by Theorem 6.2, the conditional risk of ridge is more sensitive to the task's noise level than that of the scalar rule.

Proof. Set $\beta_x = \tau^2 x/\gamma$. By Assumption C.3, $\xi_\eta = \bar{\xi} + \eta\zeta$ and $\mathrm{Var}(\xi_\eta) = \eta^2 \mathrm{Var}(\zeta)$. Applying Proposition C.8 gives

$$\frac{\Delta L_{\mathrm{ridge},\eta}}{\eta^2} = \tau^2 \int \mathbb{E}\left[\frac{\beta_x \zeta^2}{(\bar{\xi}+\beta_x)^2(\xi_\eta+\beta_x)} \right] d\mu_\gamma(x).$$

Because Assumption C.3 is two-sided in η, there exists $\eta_0 > 0$ such that $\bar{\xi}+\eta_0\zeta > 0$ and $\bar{\xi}-\eta_0\zeta > 0$ almost surely, so $|\zeta| < \bar{\xi}/\eta_0$ almost surely, and after shrinking the neighborhood of zero there is a constant $c > 0$ such that $\xi_\eta \geq c$ almost surely. The integrand is then dominated by a constant multiple of ζ^2, uniformly in x and small $|\eta|$, and dominated convergence yields

$$\frac{\Delta L_{\mathrm{ridge},\eta}}{\mathrm{Var}(\xi_\eta)} \to \tau^2 \int \frac{\beta_x}{(\bar{\xi}+\beta_x)^3} d\mu_\gamma(x).$$

Dividing by $\bar{\xi}$ and using $\tau^2 = \nu\bar{\xi}$ and $\beta_x = \nu\bar{\xi}x/\gamma$ gives the displayed formula for the mean gap.

For the dispersion, let ζ' be an independent copy of ζ and set $\xi'_\eta = \bar{\xi} + \eta\zeta'$. Proposition C.8 gives

$$\frac{D_{\mathrm{ridge},\eta}}{\eta^2} = \frac{1}{2}\mathbb{E}\left[(\zeta-\zeta')^2 \left(1 + \tau^2 \int \frac{\beta_x}{(\xi_\eta+\beta_x)(\xi'_\eta+\beta_x)} d\mu_\gamma(x) \right)^2 \right].$$

After shrinking the neighborhood of zero if necessary, Assumption C.3 gives a uniform positive lower bound for both ξ_η and ξ'_η, so the squared multiplier is bounded by a constant independent of small $|\eta|$, and the remaining factor $(\zeta-\zeta')^2$ is integrable. Dominated convergence gives

$$\frac{D_{\mathrm{ridge},\eta}}{\mathrm{Var}(\xi_\eta)} \to \left(1 + \tau^2 \int \frac{\beta_x}{(\bar{\xi}+\beta_x)^2} d\mu_\gamma(x) \right)^2,$$

because $\frac{1}{2}\mathbb{E}[(\zeta-\zeta')^2] = \mathrm{Var}(\zeta)$. Dividing by $\bar{\xi}^2$ and using $\tau^2 = \nu\bar{\xi}$ gives the displayed local dispersion formula. □

D Proofs and Supplementary Results for Section 6

All statements in this appendix are in the proportional regime of Section 6: $d, n \to \infty$ with $d/n \to \gamma \in (0,\infty)$ and ν fixed, under Assumption 3.1. The feature law enters only through the limiting spectral distribution μ_γ of $S_d := dG_x$ and through its second moment $\int x^2 \, d\mu_\gamma(x) = 1+\gamma$; Appendix E records why this requires no Gaussianity.

D.1 Asymptotic Bias, Variance, and Risk of Ridge

Proposition D.1 (Asymptotic ridge bias, variance, and risk)**.** *In the proportional regime, under Assumption 3.1, let*

$$\hat{w}_\lambda = \left(G_x + \frac{\lambda}{n} I_d \right)^{-1} G_{xy}$$

and let L_λ denote its population prediction risk. For each fixed $\lambda > 0$,

$$\frac{\mathrm{Bias}^2(\hat{w}_\lambda)}{\mathbb{E}[\sigma^2]} \to \nu(\gamma\lambda)^2 m'_\gamma(-\gamma\lambda),$$
$$\frac{\mathrm{Var}(\hat{w}_\lambda)}{\mathbb{E}[\sigma^2]} \to \gamma\{m_\gamma(-\gamma\lambda) - \gamma\lambda m'_\gamma(-\gamma\lambda)\},$$
$$\frac{L_\lambda}{\mathbb{E}[\sigma^2]} \to 1 + \gamma\{m_\gamma(-\gamma\lambda) - \gamma\lambda m'_\gamma(-\gamma\lambda)\} + \nu(\gamma\lambda)^2 m'_\gamma(-\gamma\lambda).$$

The limiting ridge risk is uniquely minimized at $\lambda^* = 1/\nu = \mathbb{E}[\sigma^2]/\tau^2$. At this value,

$$\frac{\mathrm{Bias}^2(\hat{w}_{\lambda^*})}{\mathbb{E}[\sigma^2]} \to \frac{\gamma^2}{\nu} m'_\gamma\Big(-\frac{\gamma}{\nu}\Big),$$
$$\frac{\mathrm{Var}(\hat{w}_{\lambda^*})}{\mathbb{E}[\sigma^2]} \to \gamma\left\{m_\gamma\Big(-\frac{\gamma}{\nu}\Big) - \frac{\gamma}{\nu} m'_\gamma\Big(-\frac{\gamma}{\nu}\Big)\right\},$$
$$\frac{L_{\lambda^*}}{\mathbb{E}[\sigma^2]} \to 1 + \gamma m_\gamma\Big(-\frac{\gamma}{\nu}\Big).$$

Proof of Proposition D.1. Write $H_\lambda := \big(G_x + \frac{\lambda}{n} I_d\big)^{-1}$. Since $G_{xy} = G_x w + \sigma G_{x\epsilon}$ with $G_{x\epsilon} := \frac{1}{n}\bar{X}\bar{\epsilon}$, conditioning on $(w, \sigma, \bar{X})$ gives

$$\mathbb{E}[\hat{w}_\lambda - w \mid w, \sigma, \bar{X}] = (H_\lambda G_x - I_d)w = -\frac{\lambda}{n} H_\lambda w$$

and

$$\mathbb{E}\Big[\big\|\hat{w}_\lambda - \mathbb{E}[\hat{w}_\lambda \mid w, \sigma, \bar{X}]\big\|^2 \mid w, \sigma, \bar{X}\Big] = \frac{\sigma^2}{n}\,\mathrm{Tr}(G_x H_\lambda^2),$$

using $\mathbb{E}[G_{x\epsilon} \mid \bar{X}] = 0$ and $\mathrm{Var}(G_{x\epsilon} \mid \bar{X}) = G_x/n$. The prediction risk includes the factor $1/d$ coming from the isotropic test covariance. Therefore, because $\mathbb{E}[ww^\top] = \tau^2 I_d$ and (w, σ) is independent of $\bar{X}$,

$$\mathrm{Bias}^2(\hat{w}_\lambda) = \frac{\lambda^2\tau^2}{dn^2}\mathbb{E}[\mathrm{Tr}(H_\lambda^2)], \qquad \mathrm{Var}(\hat{w}_\lambda) = \frac{\mathbb{E}[\sigma^2]}{dn}\mathbb{E}[\mathrm{Tr}(G_x H_\lambda^2)],$$

and $L_\lambda = \mathrm{Bias}^2(\hat{w}_\lambda) + \mathrm{Var}(\hat{w}_\lambda) + \mathbb{E}[\sigma^2]$.

Let $S_d := dG_x$ and $\gamma_{d,n} := d/n$. Then

$$G_x + \frac{\lambda}{n} I_d = \frac{1}{d}(S_d + \gamma_{d,n}\lambda I_d), \qquad H_\lambda = d(S_d + \gamma_{d,n}\lambda I_d)^{-1}.$$

The empirical spectral distribution of S_d converges to μ_γ (Appendix E.3). Since $\gamma_{d,n}\lambda \to \gamma\lambda > 0$, the normalized trace limits give

$$\frac{1}{d}\,\mathrm{Tr}((S_d + \gamma_{d,n}\lambda I_d)^{-2}) \to m'_\gamma(-\gamma\lambda), \qquad \frac{1}{d}\,\mathrm{Tr}(S_d(S_d + \gamma_{d,n}\lambda I_d)^{-2}) \to m_\gamma(-\gamma\lambda) - \gamma\lambda m'_\gamma(-\gamma\lambda).$$

Because $\gamma_{d,n}\lambda \to \gamma\lambda > 0$, there is a constant $c > 0$ such that $\gamma_{d,n}\lambda \geq c$ for all large n, so the test functions are uniformly bounded and the same limits hold after taking expectations. Substituting them into the finite-sample formulas gives the displayed ridge limits.

Finally, let the limiting ridge risk be denoted by $F_\gamma(\lambda; \nu) = 1 + \gamma m_\gamma(-\gamma\lambda) + \{\nu(\gamma\lambda)^2 - \gamma^2\lambda\} m'_\gamma(-\gamma\lambda)$. Differentiating,

$$F'_\gamma(\lambda; \nu) = \gamma^2(\nu\lambda - 1)\{2m'_\gamma(-\gamma\lambda) - \gamma\lambda m''_\gamma(-\gamma\lambda)\},$$
$$2m'_\gamma(-\gamma\lambda) - \gamma\lambda m''_\gamma(-\gamma\lambda) = 2\int \frac{x}{(x + \gamma\lambda)^3}\, d\mu_\gamma(x) > 0,$$

so the unique minimizer is $\lambda^* = 1/\nu = \mathbb{E}[\sigma^2]/\tau^2$. Substituting λ^* into the ridge bias and variance limits gives the displayed optimal bias and variance, and their sum plus the irreducible noise gives the optimal risk. □

Remark D.2 (In what sense ridge is optimal). The description of optimally tuned ridge as Bayes-optimal in Section 6 is a statement within the class of rules linear in the demonstration outputs. For a Gaussian prior $w \sim \mathcal{N}(0, \tau^2 I_d)$ and a fixed noise variance $\bar{\xi}$, the posterior mean is ridge with penalty $\lambda^*/n = \bar{\xi}/(n\tau^2)$, and for every rule linear in $\bar{Y}$ the risk depends on the task-noise law only through $\bar{\xi}$, so this ridge minimizes risk over that class under any task-noise law with mean $\bar{\xi}$. It is not the unrestricted Bayes rule under heterogeneous latent noise: with σ^2 random, the posterior mean of w given the demonstrations is a mixture of ridge estimators weighted by the posterior over σ^2, and it is nonlinear in $\bar{Y}$. The ordering of Theorem 6.1 should therefore be read within the output-linear class, and the task-wise oracle of Appendix C as the value of the noise level to rules in that class.

D.2 Proof of Theorem 6.1 (Information-Use Hierarchy)

Proof of Theorem 6.1. First consider scalar rules $\hat{w} = \theta G_{xy}$ and write $\theta = dt$, so that t is the scalar shrinkage level on the spectrum of $S_d = dG_x$. Using $\int x\, d\mu_\gamma(x) = 1$ and $\int x^2\, d\mu_\gamma(x) = 1 + \gamma$, the normalized proportional-limit risk of a scalar rule is

$$1 + \nu\{1 - 2t + (1+\gamma)t^2\} + \gamma t^2.$$

This strictly convex quadratic is minimized at $t^* = \nu/\{\gamma(1+\nu)+\nu\}$, which gives $\mathcal{R}_{\text{scalar}}$; it agrees with Lemma B.1. Setting $t = 1$, equivalently $\hat{w}_{\text{iso}} = dG_{xy}$, gives $\mathcal{R}_{\text{iso}}$. The expression for $\mathcal{R}_{\text{ridge}}$ is the optimally tuned ridge risk from Proposition D.1.

It remains to prove that the inequalities are strict. A spectral rule with a bounded profile h has the form $\hat{w} = d\, h(S_d) G_{xy}$. The same bias-variance calculation gives normalized risk $1 + \Phi(h)$, where the normalized excess risk is

$$\Phi(h) = \nu \int (1 - xh(x))^2\, d\mu_\gamma(x) + \gamma \int xh(x)^2\, d\mu_\gamma(x).$$

For each $x > 0$, the integrand $\nu(1-xh)^2 + \gamma x h^2$ is uniquely minimized at $h_{\text{ridge}}(x) = 1/(x+\gamma/\nu)$, which is the shrinkage profile of optimally tuned ridge. At $x = 0$, the integrand equals ν and is independent of $h(0)$, so a possible atom at zero when $\gamma > 1$ does not affect the minimization. Scalar rules correspond to constant profiles $h(x) \equiv t$. The positive part of the Marchenko–Pastur law is nondegenerate for every $\gamma > 0$, whereas h_{ridge} is not constant on that positive support. Therefore no scalar profile can attain the pointwise spectral minimum, and the inequality $\mathcal{R}_{\text{ridge}} < \mathcal{R}_{\text{scalar}}$ is strict.

For the second inequality, the scalar risk above is strictly convex in t. The isotropic plug-in rule corresponds to $t = 1$, while $t^* = \nu/\{\gamma(1+\nu)+\nu\} < 1$ for every $\gamma, \nu > 0$. Therefore $\mathcal{R}_{\text{scalar}} < \mathcal{R}_{\text{iso}}$. □

D.3 Proof of Theorem 6.2 (Risk-Gap Decomposition)

Proof of Theorem 6.2. Let Λ have the Marchenko–Pastur law μ_γ, and write $c = \gamma/\nu$ and $M = 1 + \gamma$. All expectations in this proof are with respect to μ_γ, unless stated otherwise. We use the spectral notation from the proof of Theorem 6.1: a spectral rule has the form $\hat{w}_r = d\, r(S_d) G_{xy}$ with $S_d = dG_x$. In this notation, the optimized scalar rule corresponds to the constant profile $r(x) \equiv t$, where $\theta/d \to t$ and

$$t = \frac{\nu}{\gamma(1+\nu)+\nu} = \frac{1}{M+c},$$

and the optimally tuned ridge profile is $h(x) := h_{\text{ridge}}(x) = 1/(x+c)$.

For a generic spectral rule, using $G_{xy} = G_x w + \sigma n^{-1} \bar{X}\bar{\epsilon}$ and $S_d = dG_x$,

$$\hat{w}_r = r(S_d) S_d w + \frac{\sigma d}{n} r(S_d) \bar{X}\bar{\epsilon}.$$

Therefore, conditional on $\bar{X}$, the squared-bias and variance components of the prediction risk are

$$\frac{\tau^2}{d} \operatorname{Tr}\left[\{I - S_d r(S_d)\}^2\right], \qquad \frac{\sigma^2}{n} \operatorname{Tr}\left\{S_d r(S_d)^2\right\}.$$

After normalizing by $\mathbb{E}[\sigma^2]$, using $\nu = \tau^2/\mathbb{E}[\sigma^2]$ and $d/n \to \gamma$, these components converge to

$$B_r^2 = \nu\, \mathbb{E}(1 - \Lambda r(\Lambda))^2, \qquad V_r = \gamma\, \mathbb{E}\{\Lambda r(\Lambda)^2\}.$$

For the regularization gap, the isotropic plug-in profile is $r(x) \equiv 1$. Hence

$$\Delta_{\mathrm{reg},V} = \gamma(1-t^2) = \frac{\gamma^2(1+\nu)\{\gamma(1+\nu)+2\nu\}}{\{\gamma(1+\nu)+\nu\}^2} > 0,$$

because $0 < t < 1$. For squared bias, using $B^2_{\mathrm{iso}} = \nu\gamma$ and $B^2_{\mathrm{scalar}} = \nu\{1-2t+Mt^2\}$ gives

$$\Delta_{\mathrm{reg},B^2} = \frac{\nu\gamma^2(1+\nu)\{\gamma(1+\nu)+\nu-1\}}{\{\gamma(1+\nu)+\nu\}^2}.$$

Thus the sign of $\Delta_{\mathrm{reg},B^2}$ is exactly the sign of $\gamma(1+\nu)+\nu-1$.

Theorem 6.1 gives $\Delta_{\mathrm{adapt}} > 0$. The pointwise comparison behind this inequality is

$$\nu(1-xt)^2 + \gamma x t^2 - \{\nu(1-xh(x))^2 + \gamma x h(x)^2\} = (\nu x^2 + \gamma x)\{t-h(x)\}^2,$$

which is nonnegative for every x. After integration over μ_γ, the inequality is strict because $h(x)$ is not constant on the nondegenerate positive support of μ_γ.

It remains to compare variance components. Since $t = (M+c)^{-1}$,

$$h(x) - t = \frac{M-x}{(x+c)(M+c)}.$$

Therefore

$$\frac{V_{\mathrm{ridge}} - V_{\mathrm{scalar}}}{\gamma} = \mathbb{E}\{\Lambda(h(\Lambda)-t)(h(\Lambda)+t)\} = \frac{1}{M+c}\,\mathbb{E}\{\Lambda(M-\Lambda)\psi(\Lambda)\},$$

where

$$\psi(x) := \frac{1}{x+c}\left\{\frac{1}{x+c} + \frac{1}{M+c}\right\}.$$

The Marchenko–Pastur moments give $\mathbb{E}\Lambda = 1$ and $\mathbb{E}\Lambda^2 = M$. Let Q be the probability measure $dQ(x) = x\,d\mu_\gamma(x)$, and let X denote the identity random variable under Q. Then

$$\mathbb{E}\{\Lambda(M-\Lambda)\psi(\Lambda)\} = \mathrm{Cov}_Q\{M-X, \psi(X)\},$$

because $\mathbb{E}_Q(M-X) = M - \mathbb{E}\Lambda^2/\mathbb{E}\Lambda = 0$. If X' is an independent copy of X under Q, then

$$\mathrm{Cov}_Q\{M-X,\psi(X)\} = \frac{1}{2}\mathbb{E}_Q[\{(M-X)-(M-X')\}\{\psi(X)-\psi(X')\}].$$

Under Q, both $M-X$ and $\psi(X)$ are strictly decreasing functions of X on the positive support. The product inside the expectation is therefore nonnegative, and it is positive with positive probability because the positive part of the Marchenko–Pastur law is nondegenerate. Hence $V_{\mathrm{ridge}} > V_{\mathrm{scalar}}$, so $\Delta_{\mathrm{adapt},V} < 0$. Since $\Delta_{\mathrm{adapt}} = \Delta_{\mathrm{adapt},B^2} + \Delta_{\mathrm{adapt},V}$ and $\Delta_{\mathrm{adapt}} > 0$, it follows that $\Delta_{\mathrm{adapt},B^2} > 0$. □

D.4 Proof of Theorem 6.3 (Adaptation-Gap Rates)

Proof of Theorem 6.3. Throughout the proof, B^2_{scalar} and B^2_{ridge} denote the normalized squared-bias components of the scalar and ridge risks. Using the explicit Marchenko–Pastur Stieltjes transform, with the positive square root for $z < 0$,

$$m_\gamma(z) = \frac{-(z-1+\gamma) - \sqrt{(z-(1+\gamma))^2 - 4\gamma}}{2\gamma z},$$

we have

$$\gamma m_\gamma\Big(-\frac{\gamma}{\nu}\Big) = \frac{\nu}{2\gamma}\left[\sqrt{\Big(1+\gamma+\frac{\gamma}{\nu}\Big)^2 - 4\gamma} - 1 + \gamma - \frac{\gamma}{\nu}\right].$$

For the component expansions below, we use the displayed formulas for B^2_{ridge} and V_{ridge} in Proposition D.1, with m'_γ obtained by differentiating the same explicit transform before substituting $z = -\gamma/\nu$.

First let $\gamma \to 0$ with $\nu > 0$ fixed. Taylor expansion gives

$$\gamma m_\gamma\Big(-\frac{\gamma}{\nu}\Big) = \gamma + \left(1 - \frac{1}{\nu}\right)\gamma^2 + O(\gamma^3),$$
$$\frac{\nu\gamma(1+\nu)}{\gamma(1+\nu)+\nu} = (1+\nu)\gamma - \frac{(1+\nu)^2}{\nu}\gamma^2 + O(\gamma^3).$$

Subtracting gives $\Delta_{\mathrm{adapt}}(\gamma,\nu) = \nu\gamma - (\nu+3)\gamma^2 + O(\gamma^3)$. For the components, the same Taylor expansion gives

$$B^2_{\mathrm{scalar}} = \nu\gamma + \left(\frac{1}{\nu} - \nu\right)\gamma^2 + O(\gamma^3), \qquad B^2_{\mathrm{ridge}} = \frac{\gamma^2}{\nu} + O(\gamma^3),$$

and

$$V_{\mathrm{scalar}} = \gamma - \left(2 + \frac{2}{\nu}\right)\gamma^2 + O(\gamma^3), \qquad V_{\mathrm{ridge}} = \gamma + \left(1 - \frac{2}{\nu}\right)\gamma^2 + O(\gamma^3).$$

Subtracting the ridge components from the scalar components yields

$$\Delta_{\mathrm{adapt},B^2} = \nu\gamma - \nu\gamma^2 + O(\gamma^3), \qquad \Delta_{\mathrm{adapt},V} = -3\gamma^2 + O(\gamma^3).$$

Next let $\gamma \to \infty$ with $\nu > 0$ fixed. Expanding the same expression in powers of $1/\gamma$ yields

$$\gamma m_\gamma\Big(-\frac{\gamma}{\nu}\Big) = \nu - \frac{\nu^2}{1+\nu}\frac{1}{\gamma} + \frac{\nu^3}{(1+\nu)^3}\frac{1}{\gamma^2} + O(\gamma^{-3}),$$
$$\frac{\nu\gamma(1+\nu)}{\gamma(1+\nu)+\nu} = \nu - \frac{\nu^2}{1+\nu}\frac{1}{\gamma} + \frac{\nu^3}{(1+\nu)^2}\frac{1}{\gamma^2} + O(\gamma^{-3}).$$

The constant and $1/\gamma$ terms cancel, leaving $\Delta_{\mathrm{adapt}}(\gamma,\nu) = \nu^4(1+\nu)^{-3}\gamma^{-2} + O(\gamma^{-3})$. The component expansions from the same formulas are

$$B^2_{\mathrm{scalar}} = \nu - \frac{\nu^2(\nu+2)}{(1+\nu)^2}\frac{1}{\gamma} + \frac{\nu^3(\nu+3)}{(1+\nu)^3}\frac{1}{\gamma^2} + O(\gamma^{-3}),$$
$$B^2_{\mathrm{ridge}} = \nu - \frac{\nu^2(\nu+2)}{(1+\nu)^2}\frac{1}{\gamma} + \frac{3\nu^3}{(1+\nu)^4}\frac{1}{\gamma^2} + O(\gamma^{-3}),$$

and

$$V_{\mathrm{scalar}} = \frac{\nu^2}{(1+\nu)^2}\frac{1}{\gamma} - \frac{2\nu^3}{(1+\nu)^3}\frac{1}{\gamma^2} + O(\gamma^{-3}),$$
$$V_{\mathrm{ridge}} = \frac{\nu^2}{(1+\nu)^2}\frac{1}{\gamma} + \frac{\nu^3(\nu-2)}{(1+\nu)^4}\frac{1}{\gamma^2} + O(\gamma^{-3}).$$

Therefore

$$\Delta_{\mathrm{adapt},B^2} = \frac{\nu^4(\nu+4)}{(1+\nu)^4}\frac{1}{\gamma^2} + O(\gamma^{-3}), \qquad \Delta_{\mathrm{adapt},V} = -\frac{3\nu^4}{(1+\nu)^4}\frac{1}{\gamma^2} + O(\gamma^{-3}).$$

Finally let $\nu \to 0$ with $\gamma > 0$ fixed. Since $-\gamma/\nu \to -\infty$, the large-$|z|$ expansion of the Stieltjes transform and the first Marchenko–Pastur moments give

$$\gamma m_\gamma\Big(-\frac{\gamma}{\nu}\Big) = \nu - \frac{\nu^2}{\gamma} + \frac{(1+\gamma)\nu^3}{\gamma^2} - \frac{(1+3\gamma+\gamma^2)\nu^4}{\gamma^3} + O(\nu^5),$$
$$\frac{\nu\gamma(1+\nu)}{\gamma(1+\nu)+\nu} = \nu - \frac{\nu^2}{\gamma} + \frac{(1+\gamma)\nu^3}{\gamma^2} - \frac{(1+\gamma)^2\nu^4}{\gamma^3} + O(\nu^5).$$

The terms through order ν^3 cancel, and the order-ν^4 coefficient is $\{1+3\gamma+\gamma^2-(1+\gamma)^2\}/\gamma^3 = 1/\gamma^2$, so $\Delta_{\mathrm{adapt}}(\gamma,\nu) = \nu^4/\gamma^2 + O(\nu^5)$. The corresponding component expansions are

$$B^2_{\mathrm{scalar}} = \nu - \frac{2\nu^2}{\gamma} + \frac{3(1+\gamma)\nu^3}{\gamma^2} - \left(\frac{4}{\gamma} + \frac{8}{\gamma^2} + \frac{4}{\gamma^3}\right)\nu^4 + O(\nu^5),$$
$$B^2_{\mathrm{ridge}} = \nu - \frac{2\nu^2}{\gamma} + \frac{3(1+\gamma)\nu^3}{\gamma^2} - \left(\frac{4}{\gamma} + \frac{12}{\gamma^2} + \frac{4}{\gamma^3}\right)\nu^4 + O(\nu^5),$$

and

$$V_{\mathrm{scalar}} = \frac{\nu^2}{\gamma} - \frac{2(1+\gamma)\nu^3}{\gamma^2} + \frac{3(1+\gamma)^2\nu^4}{\gamma^3} + O(\nu^5),$$
$$V_{\mathrm{ridge}} = \frac{\nu^2}{\gamma} - \frac{2(1+\gamma)\nu^3}{\gamma^2} + \left(\frac{3}{\gamma} + \frac{9}{\gamma^2} + \frac{3}{\gamma^3}\right)\nu^4 + O(\nu^5).$$

Thus

$$\Delta_{\mathrm{adapt},B^2} = \frac{4\nu^4}{\gamma^2} + O(\nu^5), \qquad \Delta_{\mathrm{adapt},V} = -\frac{3\nu^4}{\gamma^2} + O(\nu^5).$$

□

D.5 Global Bounds and Peak Behavior of the Covariance-Adaptation Gap

Theorem 6.3 says where the scalar-to-ridge gap vanishes. This subsection says what the two benchmarks converge to in those limits, which simple rules share the closeness and which do not, and where between the limits the gap is largest. The reference point is the null predictor.

Lemma D.3 (The null predictor as a reference)**.** *The rule $\hat{w} = 0$ has normalized risk $\mathcal{R}_{\mathrm{null}} := 1+\nu$ at every (d, n), not only in the limit. In the proportional limit, for every $\gamma, \nu > 0$,*

$$\mathcal{R}_{\mathrm{ridge}} < \mathcal{R}_{\mathrm{scalar}} < \mathcal{R}_{\mathrm{null}}, \qquad \mathcal{R}_{\mathrm{null}} - \mathcal{R}_{\mathrm{scalar}} = \frac{\nu^2}{\gamma(1+\nu)+\nu},$$

whereas $\mathcal{R}_{\mathrm{iso}} \geq \mathcal{R}_{\mathrm{null}}$ if and only if $\gamma(1+\nu) \geq \nu$. Moreover, as $\gamma \to \infty$ with ν fixed,

$$\mathcal{R}_{\mathrm{null}} - \mathcal{R}_{\mathrm{scalar}} = \frac{\nu^2}{(1+\nu)\gamma} + O(\gamma^{-2}), \qquad \mathcal{R}_{\mathrm{null}} - \mathcal{R}_{\mathrm{ridge}} = \frac{\nu^2}{(1+\nu)\gamma} + O(\gamma^{-2}),$$

and, as $\nu \to 0$ with γ fixed,

$$\mathcal{R}_{\mathrm{null}} - \mathcal{R}_{\mathrm{scalar}} = \frac{\nu^2}{\gamma} + O(\nu^3), \qquad \mathcal{R}_{\mathrm{null}} - \mathcal{R}_{\mathrm{ridge}} = \frac{\nu^2}{\gamma} + O(\nu^3).$$

Proof. For $\hat{w} = 0$ the risk is $d^{-1}\mathbb{E}\|w\|^2 + \mathbb{E}[\sigma^2] = \tau^2 + \mathbb{E}[\sigma^2]$, which is $(1+\nu)\mathbb{E}[\sigma^2]$ at every (d, n). From Theorem 6.1, $\mathcal{R}_{\mathrm{scalar}} - 1 = \nu\gamma(1+\nu)/R_\gamma < \nu$ because $\gamma(1+\nu) < R_\gamma$, and $\nu - \nu\gamma(1+\nu)/R_\gamma = \nu\{R_\gamma - \gamma(1+\nu)\}/R_\gamma = \nu^2/R_\gamma$; the first strict inequality is Theorem 6.1. $\mathcal{R}_{\mathrm{iso}} - \mathcal{R}_{\mathrm{null}} = \gamma(1+\nu) - \nu$ gives the statement about the isotropic plug-in. The expansions of $\mathcal{R}_{\mathrm{null}} - \mathcal{R}_{\mathrm{scalar}} = \nu^2/R_\gamma$ are immediate, and those of $\mathcal{R}_{\mathrm{null}} - \mathcal{R}_{\mathrm{ridge}} = \nu - \gamma m_\gamma(-\gamma/\nu)$ follow from the expansions of $\gamma m_\gamma(-\gamma/\nu)$ in the proof of Theorem 6.3. □

Proposition D.4 (Attribution of the small gap)**.** *Let $\Delta_{\mathrm{reg}} = \gamma^2(1+\nu)^2/R_\gamma$ and Δ_{adapt} be the two adjacent gaps of Section 6.3. Then*

$$\frac{\Delta_{\mathrm{reg}}}{\Delta_{\mathrm{adapt}}} \to \infty \quad (\gamma \to \infty), \qquad \frac{\Delta_{\mathrm{reg}}}{\Delta_{\mathrm{adapt}}} \to \infty \quad (\nu \to 0), \qquad \frac{\Delta_{\mathrm{reg}}}{\Delta_{\mathrm{adapt}}} \to 0 \quad (\gamma \to 0),$$

with $\Delta_{\mathrm{reg}} \sim (1+\nu)\gamma$ as $\gamma \to \infty$, $\Delta_{\mathrm{reg}} \to \gamma$ as $\nu \to 0$, and $\Delta_{\mathrm{reg}} \sim (1+\nu)^2\gamma^2/\nu$ as $\gamma \to 0$.

Proof. The three asymptotic forms of Δ_{reg} follow from $R_\gamma \sim (1+\nu)\gamma$ as $\gamma \to \infty$, $R_\gamma \to \gamma$ as $\nu \to 0$, and $R_\gamma \to \nu$ as $\gamma \to 0$. Theorem 6.3 gives $\Delta_{\mathrm{adapt}} \sim \nu^4(1+\nu)^{-3}\gamma^{-2}$, $\Delta_{\mathrm{adapt}} \sim \nu^4/\gamma^2$, and $\Delta_{\mathrm{adapt}} \sim \nu\gamma$ in the three regimes, and the ratios follow. □

Reading the three limits. Lemma D.3 makes precise the objection that a small scalar-to-ridge gap in the regimes $\gamma \to \infty$ and $\nu \to 0$ says little on its own. In both, the two benchmarks converge to the null predictor, at rate $1/\gamma$ in the first case and ν^2 in the second, while their difference is of the smaller order γ^{-2}, respectively ν^4. In these two limits "nearly ridge-optimal" is a statement about an additive metric in a regime where the improvement of either rule over the null predictor vanishes; at every finite (γ, ν) both improvements are strictly positive, and the main text should be read with that qualification. Proposition D.4 qualifies the objection in one direction only: the closeness is not a generic property of simple rules in a difficult regime. The isotropic plug-in $\hat{w}_{\mathrm{iso}} = dG_{xy}$

differs from the optimal scalar rule only in its tuning, yet its distance to ridge diverges as $\gamma \to \infty$ and stays at the fixed value γ as $\nu \to 0$, and by Lemma D.3 it is worse than the null predictor as soon as $\gamma(1+\nu) > \nu$. What closes the gap to ridge is the optimized scalar tuning, and Δ_{reg} measures how much of the work it does. The data-rich limit $\gamma \to 0$ is different in kind. There all three rules approach the noise floor, the scalar tuning removes only an $O(\gamma^2)$ amount, and covariance adaptation removes the first-order amount $\nu\gamma$: the gap is small in total risk because everything is, and Appendix D.5.1 shows that relative to the excess risk it does not vanish at all. Two relative metrics make the same point. The fraction of ridge's improvement over the null predictor that the scalar rule retains, $(\mathcal{R}_{\mathrm{null}} - \mathcal{R}_{\mathrm{scalar}})/(\mathcal{R}_{\mathrm{null}} - \mathcal{R}_{\mathrm{ridge}})$, tends to one in all three limits, by Lemma D.3 in the first two and because both differences tend to ν as $\gamma \to 0$, so it does not discriminate between them. The share of the scalar rule's excess risk that ridge removes, $\Delta/(\mathcal{R}_{\mathrm{scalar}} - 1)$, tends to $\nu/(1+\nu)$ as $\gamma \to 0$ and to zero in the other two limits (Appendix D.5.1); it is the metric that separates the data-rich limit from the degenerate ones.

Where the gap is largest. Between these limits the gap peaks at an intermediate aspect ratio, and both the location and the height of the peak can be quantified. Write $\Delta(\gamma,\nu) := \Delta_{\mathrm{adapt}}(\gamma,\nu)$, $\Delta_{\max}(\nu) := \sup_{\gamma>0} \Delta(\gamma,\nu)$, and let $\gamma^*(\nu)$ denote a maximizer.

Proposition D.5 (Global bounds and location of the peak)**.** *For the proportional-limit risks, the following statements hold.*

1. Global quadratic bound. *For every $\nu > 0$,*

$$\frac{4\nu^2}{27(1+5\nu/9)} \le \Delta_{\max}(\nu) \le \frac{4\nu^2}{27}.$$

In particular, $\Delta(\gamma,\nu) \le 4\nu^2/27$ uniformly over all $\gamma > 0$.

2. Weak signal. *As $\nu \to 0$,*

$$\Delta_{\max}(\nu) = \frac{4\nu^2}{27} + O(\nu^3), \qquad \frac{\gamma^*(\nu)}{\nu} \to \frac{1}{2}$$

for every choice of maximizer. For every fixed $c > 0$,

$$\Delta(c\nu,\nu) = \frac{c}{(1+c)^3}\,\nu^2 + O(\nu^3).$$

At the maximizing aspect ratio, ridge removes a share asymptotic to $4\nu/9$ of the scalar rule's excess risk.

3. Strong signal. *For every $\nu > 0$,*

$$\frac{\nu}{2} - \sqrt{\nu} \;\le\; \Delta_{\max}(\nu) \;\le\; \frac{\nu}{2} + \frac{1}{4},$$

so that $\Delta_{\max}(\nu)/\nu \to 1/2$ as $\nu \to \infty$; every maximizer satisfies $\gamma^*(\nu) \to 1$, and the share $\Delta_{\max}(\nu)/\{\mathcal{R}_{\mathrm{scalar}}(\gamma^*(\nu),\nu) - 1\}$ of the scalar rule's excess risk removed at the peak tends to one.

Proof. Item 1. Write $r := \mathcal{R}_{\mathrm{ridge}} - 1 = \gamma m_\gamma(-\gamma/\nu)$. The explicit Marchenko–Pastur transform gives

$$r^2 + (1 - \nu + \nu/\gamma)r - \nu = 0.$$

Its positive root lies in $(0,\nu)$. Setting $u := r/\nu \in (0,1)$ yields

$$\gamma = \frac{\nu u}{(1-u)(1+\nu u)}.$$

For each fixed $\nu > 0$, this is a strictly increasing bijection from $(0,1)$ to $(0,\infty)$: its derivative has positive numerator $\nu(1+\nu u^2)$. Substitution into $\mathcal{R}_{\mathrm{scalar}} - 1 = \nu\gamma(1+\nu)/\{\gamma(1+\nu)+\nu\}$ gives the exact identity

$$\Delta(\gamma,\nu) = \nu^2 f_\nu(u), \qquad f_\nu(u) := \frac{u(1-u)^2}{1+\nu u(2-u)}.$$

The continuous extension of f_ν to $[0,1]$ vanishes at both endpoints and is positive inside, so a global maximizer exists in $(0,1)$. Since the denominator is at least one and $\max_{u\in[0,1]} u(1-u)^2 = 4/27$, the upper bound follows. Evaluating at $u = 1/3$, corresponding to $\gamma = \nu/\{2(1+\nu/3)\}$, gives the lower bound.

Item 2. The bounds in item 1 give $\Delta_{\max}(\nu) = 4\nu^2/27 + O(\nu^3)$. Moreover, $f_\nu(u) \to u(1-u)^2$ uniformly on $[0,1]$, and the limiting function has the unique maximizer $u = 1/3$. Thus every maximizing $u^*(\nu)$ tends to $1/3$, and the bijection above gives $\gamma^*(\nu)/\nu \to 1/2$. For fixed $c > 0$ and $\gamma = c\nu$, the same relation gives $u = c/(1+c) + O(\nu)$, so the exact gap identity yields $\Delta(c\nu, \nu) = c\nu^2/(1+c)^3 + O(\nu^3)$. Finally, at $\gamma = \gamma^*(\nu)$ the scalar excess risk is $\nu/3 + o(\nu)$; dividing the maximum gap by this quantity gives the share $4\nu/9 + o(\nu)$.

Item 3. For the upper bound, first let $\gamma \le 1$. Since $\Delta \le \mathcal{R}_{\text{scalar}} - 1$ and $\mathcal{R}_{\text{scalar}} - 1 = \nu\gamma(1+\nu)/R_\gamma$ is increasing in γ,

$$\Delta(\gamma,\nu) \le \frac{\nu(1+\nu)}{1+2\nu} = \frac{\nu}{2} + \frac{\nu}{2(1+2\nu)} \le \frac{\nu}{2} + \frac{1}{4}.$$

Next let $\gamma > 1$. The atom of μ_γ at zero has mass $1 - 1/\gamma$, and the integrand of the ridge squared bias equals ν there, so $\mathcal{R}_{\text{ridge}} - 1 \ge B^2_{\text{ridge}} \ge \nu(1 - 1/\gamma)$ and

$$\Delta(\gamma,\nu) \le f(\gamma) := \frac{\nu\gamma(1+\nu)}{\gamma(1+\nu)+\nu} - \nu\Big(1 - \frac{1}{\gamma}\Big).$$

Since $f'(\gamma) = \nu^2(1+\nu)/\{\gamma(1+\nu)+\nu\}^2 - \nu/\gamma^2 < 0$ for $\gamma \ge 1$, because $\{\gamma(1+\nu)+\nu\}^2 > \gamma^2(1+\nu)^2 \ge \nu(1+\nu)\gamma^2$, we get $\Delta \le f(1) = \nu(1+\nu)/(1+2\nu) \le \nu/2 + 1/4$ again. For the lower bound, evaluate at $\gamma = 1$: $\mathcal{R}_{\text{scalar}}(1,\nu) - 1 = \nu(1+\nu)/(1+2\nu) \ge \nu/2$, and the explicit Stieltjes transform at $\gamma = 1$ gives $\mathcal{R}_{\text{ridge}}(1,\nu) - 1 = m_1(-1/\nu) = \{\sqrt{1+4\nu} - 1\}/2 \le \sqrt{\nu}$, so $\Delta_{\max}(\nu) \ge \Delta(1,\nu) \ge \nu/2 - \sqrt{\nu}$. For the location, fix $\delta \in (0,1)$. For $\gamma \le 1 - \delta$, monotonicity gives $\Delta \le \mathcal{R}_{\text{scalar}}(1-\delta,\nu) - 1 \le (1-\delta)(1+\nu)/(2-\delta)$, and for $\gamma \ge 1 + \delta$, $\Delta \le f(1+\delta) \le (1+\delta)(1+\nu)/(2+\delta) - \nu\delta/(1+\delta)$; both bounds are of the form $c_\delta\nu + O(1)$ with $c_\delta < 1/2$, so for all large ν they fall below the lower bound $\nu/2 - \sqrt{\nu}$, and every maximizer lies in $(1-\delta, 1+\delta)$. Finally, since $\gamma^*(\nu) \to 1$, $\mathcal{R}_{\text{scalar}}(\gamma^*(\nu),\nu) - 1 = \nu/2 + o(\nu)$, and the share at the peak is at least $(\nu/2 - \sqrt{\nu})/(\nu/2 + o(\nu)) \to 1$. □

Table 3 confirms both regimes: at $\nu = 0.01$ the maximizer is $0.496\,\nu$ and the height is $0.147\,\nu^2$, against the predicted $0.5\,\nu$ and $0.148\,\nu^2$; at $\nu = 10^4$ the maximizer is 0.989 and the height is $0.491\,\nu$. At $\nu = 1$ the peak is at $\gamma^* = 0.29$ with height 0.098, where ridge removes 27% of the scalar rule's excess risk. The two relative metrics peak elsewhere. The total-risk ratio $M_{\text{tot}} = \Delta/\mathcal{R}_{\text{ridge}}$ of Appendix D.5.1 is maximized slightly below γ^* (at $\gamma \approx 0.23$ for $\nu = 1$, with value 7.9%, and at $\gamma \approx 0.39$ for $\nu = 10$, with value 153%), and the excess share $\Delta/(\mathcal{R}_{\text{scalar}} - 1)$ is, numerically, decreasing in γ for every ν we checked, from its data-rich limit $\nu/(1+\nu)$ down to zero. So the covariance-adaptation gap matters in absolute terms for strong signal at aspect ratios just below one, and in relative terms throughout the data-rich regime; it matters little for weak signal at any aspect ratio, where even at the peak ridge removes only a share of about $4\nu/9$ of the scalar rule's excess risk. This is the regime in which a comparison with a trained transformer is informative, and Appendix G.5 measures the recovered share of the gap on a grid that brackets it. All of this concerns isotropic features; Appendix F.3 finds a substantial relative gap under uncentered ReLU random features over the tested range up to $\gamma = 8$ at fixed $n = 16$, without establishing a nonvanishing asymptotic gap.

D.5.1 Relative Gaps in Total and Excess Risk

The additive gap $\Delta = \mathcal{R}_{\text{scalar}} - \mathcal{R}_{\text{ridge}}$ does not by itself distinguish two notions of closeness: closeness relative to the full prediction risk, which includes the common irreducible-noise floor, and closeness relative to the reducible excess risk. Define

$$M_{\text{tot}} := \frac{\mathcal{R}_{\text{scalar}}}{\mathcal{R}_{\text{ridge}}} - 1 = \frac{\Delta}{\mathcal{R}_{\text{ridge}}}, \qquad M_{\text{ex}} := \frac{\mathcal{R}_{\text{scalar}} - 1}{\mathcal{R}_{\text{ridge}} - 1} - 1 = \frac{\Delta}{\mathcal{R}_{\text{ridge}} - 1}.$$

The first is the usual multiplicative risk gap. The second removes the irreducible-noise term before taking the ratio.

Table 3: The covariance-adaptation gap at its peak over the aspect ratio, in the proportional limit and normalized by $\mathbb{E}[\sigma^2]$: the maximizer $\gamma^*(\nu)$, the height $\Delta_{\max}(\nu)$, the share of the scalar rule's excess risk that ridge removes at the peak, and the predictions of Proposition D.5, namely the weak-signal leading terms $\nu/2$ and $4\nu^2/27$ and the strong-signal bounds $[\nu/2 - \sqrt{\nu},\ \nu/2 + 1/4]$ on $\Delta_{\max}$.

ν	$\gamma^*(\nu)$	$\nu/2$	$\Delta_{\max}(\nu)$	$4\nu^2/27$, or bounds on $\Delta_{\max}$	share of scalar excess
0.01	0.0050	0.005	1.47×10^{-5}	1.48×10^{-5}	0.4%
0.1	0.046	0.05	0.0014	0.0015	4%
$10^{-1/2}$	0.128	0.16	0.0127	0.0148	12%
1	0.289	0.5	0.098	0.148; [−0.50, 0.75]	27%
$10^{1/2}$	0.499	—	0.60	[−0.20, 1.83]	48%
10	0.683	—	2.87	[1.84, 5.25]	67%
100	0.890	—	42.0	[40.0, 50.25]	89%
10^4	0.989	—	4914	[4900, 5000.25]	99%

Corollary D.6 (Multiplicative adaptation-gap rates)**.** *Under Assumption 3.1, in the proportional regime, the expansions in Theorem 6.3 imply*

$$M_{\text{tot}} = \begin{cases} \nu\gamma + O(\gamma^2), & \gamma \to 0, \\ \dfrac{\nu^4}{(1+\nu)^4}\gamma^{-2} + O(\gamma^{-3}), & \gamma \to \infty, \\ \dfrac{\nu^4}{\gamma^2} + O(\nu^5), & \nu \to 0, \end{cases}$$

and

$$M_{\text{ex}} = \begin{cases} \nu + O(\gamma), & \gamma \to 0, \\ \dfrac{\nu^3}{(1+\nu)^3}\gamma^{-2} + O(\gamma^{-3}), & \gamma \to \infty, \\ \dfrac{\nu^3}{\gamma^2} + O(\nu^4), & \nu \to 0. \end{cases}$$

Each limit holds with the other parameter fixed, as in Theorem 6.3.

Thus the multiplicative gap in total risk vanishes in all three regimes, but the multiplicative gap in excess risk does not vanish as $\gamma \to 0$ at a fixed signal-to-noise ratio. Indeed,

$$\frac{\Delta}{\mathcal{R}_{\text{scalar}} - 1} = \frac{M_{\text{ex}}}{1 + M_{\text{ex}}} \longrightarrow \frac{\nu}{1+\nu} \qquad (\gamma \to 0).$$

For example, at $\nu = 1$, ridge removes asymptotically one half of the scalar rule's reducible error even though their total risks are multiplicatively close because both approach the same noise floor. By contrast, when $\gamma \to \infty$ or $\nu \to 0$, both the total-risk and excess-risk multiplicative gaps vanish, for the reason given by Lemma D.3. The phrase *nearly ridge-optimal* is therefore metric-dependent in the data-rich limit and unambiguous, if also uninformative, in the latter two limits.

Proof of Corollary D.6. The proof of Theorem 6.3 gives the ridge excess-risk expansions

$$\mathcal{R}_{\text{ridge}} - 1 = \begin{cases} \gamma + O(\gamma^2), & \gamma \to 0, \\ \nu + O(\gamma^{-1}), & \gamma \to \infty, \\ \nu + O(\nu^2), & \nu \to 0. \end{cases}$$

Combining these with the corresponding expansions of Δ from Theorem 6.3 and dividing by $\mathcal{R}_{\text{ridge}} - 1$ yields the three expansions for M_{ex}. For M_{tot}, use $\mathcal{R}_{\text{ridge}} = 1 + O(\gamma)$ as $\gamma \to 0$, $\mathcal{R}_{\text{ridge}} = 1 + \nu + O(\gamma^{-1})$ as $\gamma \to \infty$, and $\mathcal{R}_{\text{ridge}} = 1 + O(\nu)$ as $\nu \to 0$, and divide the same expansions of Δ by these denominators. Finally, $\Delta/(\mathcal{R}_{\text{scalar}} - 1) = M_{\text{ex}}/(1 + M_{\text{ex}})$, so the $\gamma \to 0$ limit equals $\nu/(1+\nu)$. □

D.6 OLS and Ridgeless Benchmarks at the Interpolation Threshold

Proposition D.7 (OLS, ridgeless, and interpolation threshold)**.** *In the proportional regime of Section 6, under Assumption 3.1, and assuming in addition that $x^{(i)} \sim \mathcal{N}(0, I_d/d)$, the following limits hold. This is the one result in the paper whose proof needs Gaussian features: it evaluates inverse-Wishart moments of G_x at finite (d, n), and the spectral functional $x \mapsto 1/x$ is unbounded at the origin, so the universality argument of Appendix E does not reach it. The limits themselves are universal.*

For OLS, $\hat{w}_{\mathrm{OLS}} = G_x^{-1} G_{xy}$ with $n > d + 1$, and $0 < \gamma < 1$,

$$\frac{\mathrm{Bias}^2(\hat{w}_{\mathrm{OLS}})}{\mathbb{E}[\sigma^2]} \to 0, \qquad \frac{\mathrm{Var}(\hat{w}_{\mathrm{OLS}})}{\mathbb{E}[\sigma^2]} \to \frac{\gamma}{1-\gamma}, \qquad \frac{L_{\mathrm{OLS}}}{\mathbb{E}[\sigma^2]} \to \frac{1}{1-\gamma}.$$

For ridgeless regression, $\hat{w}_{\mathrm{RL}} = G_x^{\dagger} G_{xy}$ with $d > n + 1$, and $\gamma > 1$,

$$\frac{\mathrm{Bias}^2(\hat{w}_{\mathrm{RL}})}{\mathbb{E}[\sigma^2]} \to \nu\left(1 - \frac{1}{\gamma}\right), \qquad \frac{\mathrm{Var}(\hat{w}_{\mathrm{RL}})}{\mathbb{E}[\sigma^2]} \to \frac{1}{\gamma - 1}, \qquad \frac{L_{\mathrm{RL}}}{\mathbb{E}[\sigma^2]} \to \mathcal{R}_{\mathrm{RL}}(\gamma, \nu) := 1 + \frac{1}{\gamma - 1} + \nu\left(1 - \frac{1}{\gamma}\right).$$

Moreover, for $\gamma > 1$,

$$\mathcal{R}_{\mathrm{RL}} - \mathcal{R}_{\mathrm{scalar}} = \frac{\gamma^2 + 2\gamma\nu - (\gamma - 1)\nu^2}{\gamma(\gamma - 1)\{\gamma(1 + \nu) + \nu\}},$$

so that, with $\nu^*_{\mathrm{RL}} := \gamma(1 + \sqrt{\gamma})/(\gamma - 1)$, the scalar rule has lower risk than ridgeless regression for $\nu < \nu^*_{\mathrm{RL}}$, the two have equal risk at $\nu = \nu^*_{\mathrm{RL}}$, and ridgeless regression has lower risk for $\nu > \nu^*_{\mathrm{RL}}$.

Proof. The conditioning argument used in the proof of Proposition D.1 gives, for any estimator $\hat{w} = A(G_x) G_{xy}$ with $A(G_x)$ measurable with respect to G_x,

$$\mathrm{Bias}^2(\hat{w}) = \frac{1}{d}\mathbb{E}\left[\|(A(G_x)G_x - I_d)w\|^2\right], \qquad \mathrm{Var}(\hat{w}) = \frac{\mathbb{E}[\sigma^2]}{dn}\mathbb{E}\left[\mathrm{Tr}(A(G_x)G_x A(G_x)^{\top})\right].$$

For OLS, $A(G_x) = G_x^{-1}$ and hence the bias is zero. The inverse-Wishart moment gives, for $n > d + 1$, $\mathbb{E}[\mathrm{Tr}(G_x^{-1})] = d^2 n/(n - d - 1)$, so $\mathrm{Var}(\hat{w}_{\mathrm{OLS}}) = d\,\mathbb{E}[\sigma^2]/(n - d - 1)$, which yields the stated limit when $d/n \to \gamma \in (0, 1)$.

For ridgeless regression, $A(G_x) = G_x^{\dagger}$ and $G_x^{\dagger} G_x$ is the orthogonal projection onto the column space of the design matrix, which has rank n with probability one in the underdetermined case. Thus $\mathrm{Bias}^2(\hat{w}_{\mathrm{RL}}) = \tau^2(1 - n/d)$. The nonzero eigenvalues of G_x are the same as those of the dual sample covariance up to the same normalization, and the inverse-Wishart moment gives, for $d > n + 1$, $\mathbb{E}[\mathrm{Tr}(G_x^{\dagger})] = dn^2/(d - n - 1)$. Hence $\mathrm{Var}(\hat{w}_{\mathrm{RL}}) = n\,\mathbb{E}[\sigma^2]/(d - n - 1)$, which yields the stated limit when $d/n \to \gamma > 1$.

It remains to compare ridgeless regression with the optimal scalar rule. Direct subtraction gives the displayed expression for $\mathcal{R}_{\mathrm{RL}} - \mathcal{R}_{\mathrm{scalar}}$. The denominator is positive for $\gamma > 1$ and $\nu > 0$. The numerator is a quadratic in ν with a unique positive root $\nu^*_{\mathrm{RL}} = \gamma(1 + \sqrt{\gamma})/(\gamma - 1)$, which gives the stated threshold. □

Remark D.8 (Ridge versus ridgeless). For $\gamma > 1$, optimally tuned ridge remains strictly better than ridgeless regression. Indeed, in the spectral-risk representation used in Theorem 6.1, ridgeless regression uses $h(x) = 1/x$ on the positive part of the Marchenko–Pastur spectrum, whereas the unique pointwise minimizer is the ridge profile $h_{\mathrm{ridge}}(x) = 1/(x + \gamma/\nu)$. The positive part of the spectrum is nondegenerate, so the inequality is strict.

E Beyond Gaussianity: Assumptions, Universality, and Numerical Checks

Assumption 3.1 asks only for moments and independence. This appendix records which item each result actually uses, proves the two facts that let the Gaussian case be dropped, and marks the exceptions.

E.1 Assumptions by Result

The isotropic development starts from Assumption 3.1, with additional hypotheses where stated; Appendix F replaces the feature assumptions. Table 4 separates fixed-coefficient identities, optimal tuning, and proportional limits.

Table 4: Assumptions used by each result. The anisotropic propositions state their replacement assumptions explicitly.

Result	Conditions used
Fixed-coefficient dispersion identities (Appendix C.1)	(1)–(4); feature isotropy suffices, without task isotropy
Proposition 5.1 at θ^* and its limit	(1)–(5); optimal scaling uses task isotropy and $\psi_{d,n}$, and the limit uses Lemma B.1
Proposition 5.2	(1)–(4), finite risk and the stated regularity for differentiation
Proposition 4.1; Proposition B.2 (consistency)	(1)–(5); feature isotropy and finite fourth moment suffice for the finite-sample formulas
Proposition B.2 (central limit theorems)	Additionally $\mathbb{E}\|w\|^4 < \infty$ and $\mathbb{E}\|x\|^8 < \infty$
Proposition 3.2	(1)–(5), vanishing first and third task moments; the feature reduction needs $\mathbb{E}[\|x\|^2 xx^\top] \propto I_d$
Proposition C.1; Propositions C.4 and C.5	(1)–(5); additionally Assumption C.3 for the local expansions
Proposition C.6, Corollary C.7	(1)–(5), $n > d$, full column rank, and Gaussian demonstration noise
Theorems 6.1, 6.2, 6.3; Proposition D.1 and the gap/noise corollaries	(1)–(5), the Marchenko–Pastur limit and scalar moment limit; the local ridge result also uses Assumption C.3
Proposition D.7	(1)–(5), and Gaussian features
Propositions F.1 and F.2	Gaussian feature/task laws in the first statement; central symmetry, independence, and the stated finite moments in the general statement
Proposition F.4	Feature second moment C, $\operatorname{Tr} C = 1$, finite fourth moment, task second-moment isotropy, and independence; no feature symmetry required
Proposition F.5	The scalar-risk assumptions and finite noise-level variance; its limit additionally requires convergence of $m_{2,d}, m_{3,d}, \chi_d$

Assumption 3.1(3) does two different jobs. For scalar optimality, isotropic feature and task second moments make the linear term in the fixed map depend only on its trace, while $\mathbb{E}[G_x^2] \propto I_d$ makes the quadratic term isotropic. The latter is a fourth-moment property: covariance isotropy alone does not imply it. Independent identically distributed standardized coordinates supply both properties even when their law is skewed, as for centered exponential coordinates. Sign-and-permutation invariance is an alternative sufficient condition for the fourth-moment property, not a necessary symmetry of all admitted feature laws.

Separately, the i.i.d.-coordinate assumption delivers the classical Marchenko–Pastur limit. Dependent coordinates require separate spectral conditions, such as the quadratic-form concentration discussed in Appendix E.3; symmetry alone does not suffice. For example, let $x = Ru$ with u uniform on the unit sphere, independent R, $\mathbb{E}R^2 = 1$, and $\mathbb{E}R^4 < \infty$. The scalar structure and finite-sample risk formulas still hold under their task/noise assumptions, but for nonconstant R^2, $\psi_{d,n} \to \mathbb{E}R^4 + 1/\gamma \neq 1 + 1/\gamma$, and the spectral law also changes. The unit-sphere design itself has $\psi_{d,n} = 1 + (n-1)/d$ and the needed spectral concentration, so it supports the isotropic scalar and positive-penalty ridge results under the remaining task/noise hypotheses. The Gaussian-feature OLS statement is not asserted for that design; the residual plug-in result still requires Gaussian noise and full rank.

The anisotropic results in Appendix F cover non-isotropic second moments. Their symmetry argument removes directional terms, while existence of a finite-risk optimum and its explicit formula use the moment and independence conditions stated there. For a fixed scalar coefficient, the noise-dispersion slope uses only the feature second moment; evaluating it at the optimal coefficient intro-

duces the task second moment and feature fourth moment. Finally, the isotropic structural result uses vanishing first and third task moments, rather than Gaussianity of the task prior. Feature skewness is permitted under the stated coordinate and moment conditions.

E.2 Universality of Scalar-Risk Limits

Lemma B.1 shows that the single constant $\psi_{d,n}$ through which the feature law enters the finite-sample formulas converges to $1 + 1/\gamma$ whatever the coordinate kurtosis. So the fourth moment of the features, which does move the finite-sample constants (Appendix E.4), leaves every proportional-limit statement in the paper untouched, the dispersion peak at $\gamma = \nu/(1+\nu)$ included. The same is true of the results of Section 6.

Theorem E.1 (Universality of the Section 6 results)**.** *Under Assumption 3.1, Proposition D.1 and Theorems 6.1, 6.2, and 6.3 hold as stated, with the same constants, and so do Lemma D.3, Propositions D.4 and C.8, and Corollaries D.6 and C.9.*

Proof. These proofs use the feature law in exactly two ways. First, through the limiting spectral distribution of $S_d = dG_x$, which is μ_γ by Appendix E.3 below; the functionals involved, $x \mapsto (x+\gamma\lambda)^{-2}$ and $x \mapsto x(x+\gamma\lambda)^{-2}$ for fixed $\lambda > 0$, and $x \mapsto x h_{\mathrm{ridge}}(x)$ and $x \mapsto x h_{\mathrm{ridge}}(x)^2$, are bounded and continuous on $[0, \infty)$, so weak convergence suffices and, being bounded, the limits also hold in expectation. Second, through the second spectral moment $\int x^2 d\mu_\gamma(x) = 1 + \gamma$, which enters the scalar-rule risk; this is Lemma B.1 in spectral form and needs only the fourth moment of Assumption 3.1(4). Every remaining step uses $\mathbb{E}[ww^\top] = \tau^2 I_d$, $\mathbb{E}[G_{x\epsilon} \mid \bar{X}] = 0$, and $\mathrm{Var}(G_{x\epsilon} \mid \bar{X}) = G_x/n$, which follow from Assumption 3.1(1)–(3) and (5). The statements added in Appendix D.5 are algebraic consequences of the closed forms and inherit their universality. □

E.3 Universality of Spectral Limits

Write $S_d = dG_x = n^{-1}\sum_{i=1}^n z^{(i)} z^{(i)\top}$ with $z^{(i)} = \sqrt{d}\, x^{(i)}$. Under Assumption 3.1(2)–(3), the $d \times n$ matrix with columns $z^{(i)}$ has i.i.d. entries of mean zero and unit variance, and the Marchenko–Pastur theorem states that the empirical spectral distribution of S_d converges almost surely to μ_γ whenever $d/n \to \gamma \in (0, \infty)$ (Marchenko & Pastur, 1967; Silverstein & Bai, 1995; Bai & Silverstein, 2010). Finite variance is the only moment condition; Gaussianity plays no role.

Two boundaries of that statement are worth recording. First, the i.i.d.-coordinate condition in Assumption 3.1(3) is stronger than isotropy alone, $\mathbb{E}[xx^\top] = I_d/d$. Features drawn uniformly from a sphere satisfy isotropy but have dependent coordinates, so they fall outside this coordinate assumption and require a separate spectral universality argument (Bai & Zhou, 2008). Second, weak convergence controls only bounded continuous functionals of the spectrum. That is enough for every ridge quantity at a fixed $\lambda > 0$, and for the scalar rule once Assumption 3.1(4) supplies the second spectral moment, but it is not enough at $\lambda = 0$, where $x \mapsto 1/x$ is unbounded at the origin and the smallest eigenvalue must be controlled (Bai & Yin, 1993). This is why Proposition D.7, whose finite-sample step evaluates inverse-Wishart moments, is stated for Gaussian features.

E.4 Numerical Check of Scalar Risk (Proposition 4.1)

Assumption 3.1 asks for moments, not for a parametric family, and the proof above shows that the two components of the quadratic coefficient of $L(\theta)$ have different sensitivities to the design: the variance part $1/(nd)$ uses only $\mathbb{E}[\mathrm{Tr}\, G_x] = 1$, whereas the bias part $\nu\,\mathbb{E}[\mathrm{Tr}(G_x^2)]/d = \nu\psi_{d,n}/(nd)$ depends on the fourth moment. This gives a sharper test than asking whether the Gaussian formula "roughly works": under a change of design one term should stay fixed and the other should move by exactly the amount that $\psi_{d,n}$ predicts.

Protocol. We fix $d = 10$, $n = 20$, $\tau^2 = 1$, and a constant noise variance $\sigma^2 = 1$, so $\nu = 1$, and draw features from five designs normalized so that $\mathbb{E}[xx^\top] = I_d/d$ in every case: Gaussian; Rademacher coordinates; the uniform law on the unit sphere; Laplace coordinates; and Student t_5 coordinates. The last two have coordinate kurtosis $\kappa = 6$ and $\kappa = 9$, the Rademacher design has $\kappa = 1$, and the sphere has dependent coordinates with $\|x\|^2 = 1$ exactly. No model is trained: the estimator is $\hat{w} = \theta G_{xy}$ in closed form, and the normalized excess risk $\mathbb{E}\|w - \theta G_{xy}\|^2/(d\,\mathbb{E}[\sigma^2])$

is estimated on 200,000 tasks for each of three seeds on a grid of 40 values of θ. Because $\hat{w}$ is linear in θ, the risk is exactly quadratic in θ, and a parabola fitted to the Monte Carlo curve recovers its coefficients rather than approximating them. The quantity $\mathbb{E}[\mathrm{Tr}(G_x^2)]$ is estimated on the same draws. The variance coefficient in the component plot is supplied analytically as $1/(nd)$, rather than measured separately; the risk curve and its fitted minimum are empirical, but this plot does not independently verify both components.

Table 5: The moment-only risk formula on five feature designs at $d = 10$, $n = 20$, $\nu = 1$. "Predicted" values use $\psi_{d,n} = \mathbb{E}\|x\|^4 + (n-1)/d$ with $\mathbb{E}\|x\|^4 = 1 + (\kappa - 1)/d$ for i.i.d. coordinates of kurtosis κ and $\mathbb{E}\|x\|^4 = 1$ for the sphere; the Gaussian value is $\psi_{d,n} = 3.1$. "Measured" values are Monte Carlo estimates over $3 \times 200{,}000$ tasks; the empirical minimizer is the vertex of the fitted parabola. The last column is the relative change of the bias part of the θ^2 coefficient relative to the Gaussian design.

design	κ	$\mathbb{E}\|x\|^4$	$\psi_{d,n}$	$\mathbb{E}\,\mathrm{Tr}(G_x^2)$ pred.	measured	θ^* pred.	minimizer	bias part vs. Gaussian
Gaussian	3	1.20	3.10	0.1550	0.1549	4.878	4.879	$\pm 0\%$
Rademacher	1	1.00	2.90	0.1450	0.1450	5.128	5.127	-6.5%
sphere	—	1.00	2.90	0.1450	0.1450	5.128	5.129	-6.5%
Laplace	6	1.50	3.40	0.1700	0.1699	4.545	4.543	$+9.6\%$
Student t_5	9	1.80	3.70	0.1850	0.1838	4.255	4.274	$+18.5\%$

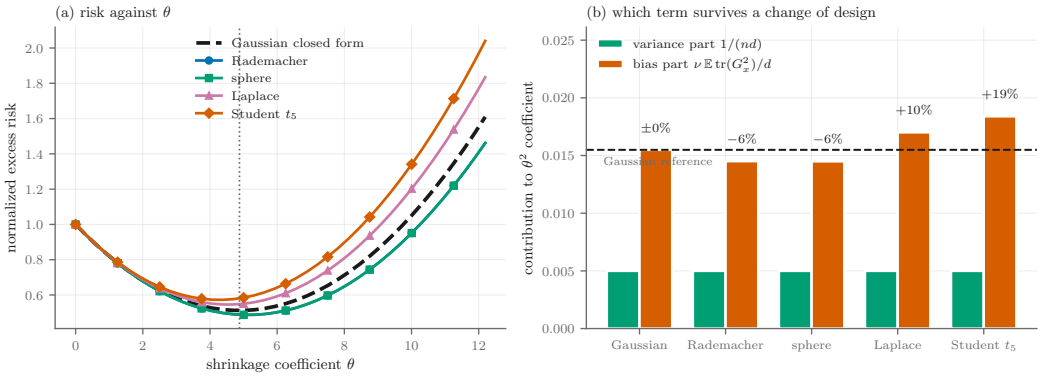

Figure 4: The moment-only formula under a change of design ($d = 10$, $n = 20$, $\nu = 1$). Left: normalized excess risk against θ for the five designs of Table 5; the dashed curve is the Gaussian closed form and the dotted vertical line its minimizer. Right: the two parts of the θ^2 coefficient. The variance part $1/(nd)$ is identical across designs, as Proposition 4.1 requires, while the bias part $\nu\,\mathbb{E}\,\mathrm{Tr}(G_x^2)/d$ moves by -6% to $+19\%$, in each case by the amount that the design's fourth moment predicts through $\psi_{d,n}$.

Reading. Table 5 and Figure 4 show three things. First, the displayed variance part of the quadratic coefficient is the theoretical reference $1/(nd) = 0.005$ for every design, not a separate Monte Carlo measurement; the separate measurement of both components is reported below (Table 6). Second, the bias part moves, and it moves by the predicted amount: the measured $\mathbb{E}[\mathrm{Tr}(G_x^2)]$ agrees with $\psi_{d,n}/n$ to three decimals for the four light-tailed designs and to within the Monte Carlo standard deviation (0.0016 over seeds) for the heavy-tailed t_5 design. Third, the empirical minimizer tracks the design-specific $\theta^* = \nu n/(1 + \nu\psi_{d,n})$, so a scalar rule tuned with the Gaussian value of $\psi_{d,n}$ on a Student t_5 design would use a coefficient about 15% too large at this (d, n), that is, shrink too little; in the proportional limit the discrepancy disappears by Lemma B.1. The sphere design, whose coordinates are dependent, behaves exactly as Remark A.1 predicts.

Both components measured separately. The plot above inserts the theoretical variance coefficient; the following measurement does not. For each task we form $u = \theta G_x w - w$ and $v = \theta X^\top \epsilon/n$ and average $u^\top C u$ and $\xi\, v^\top C v$ separately ($C = I_d/d$), on five replicates of 200,000 tasks, at the design's own θ^*; the empirical minimizer $\hat{\theta}_{\mathrm{emp}} = \sum_t w_t^\top C g_t / \sum_t g_t^\top C g_t$ is computed on a sepa-

rate tuning sample of 200,000 tasks, and the direct noisy excess risk (single noise draw, no pairing) is recorded as a check. Table 6: the measured squared bias agrees with $\tau^2(1-2\theta/d+\theta^2\psi_{d,n}/(dn))$ to within 0.12% and the measured variance with $\theta^2\bar{\xi}/(dn)$ to within 0.10% on all five light-tailed designs, the direct excess equals the sum of the two components to within 0.04%, and the empirical minimizer is within 0.08% of θ^*. The Gaussian value of $\psi_{d,n}$ would put the squared bias off by +7.4% (Rademacher and sphere), -7.0% (Laplace) and -11.2% (centred exponential) at this d, which is the kurtosis correction $(\kappa-3)/d$ of Table 5 seen component by component. The centred-exponential design is included here because the identity needs only $\mathbb{E}\|x\|^4<\infty$; it does not appear in Table 5. Student t_5 is listed separately as a stress test: its variance component matches (0.0905 against 0.0905 at every sample size), but its squared-bias estimate sits -1.4%, -1.3% and -0.9% below the reference at 50,000, 200,000 and 800,000 tasks and drifts upward, and the replicate spread shrinks more slowly than $m^{-1/2}$, as expected when $\mathbb{E}\|x\|^8=\infty$ makes the estimator of $\mathbb{E}[u^\top Cu]$ heavy-tailed; the usual finite-variance standard errors do not apply to that row.

Table 6: Squared bias and variance of the scalar rule measured separately ($d=10$, $n=20$, $\nu=1$, five replicates of 200,000 tasks, at the design's own θ^*), against the exact moment formulas and against the formula with the Gaussian $\psi_{d,n}$; last column, the empirical minimizer from an independent tuning sample. Student t_5 rows: three sample sizes, standard errors over replicates shown but not calibrated (infinite eighth moment).

design	$\psi_{d,n}$	Bias^2 MC	exact	Gaussian-ψ formula	Var MC	exact	$\hat{\theta}_{\mathrm{emp}}/\theta^*$
Gaussian	3.10	0.3931 ± 0.0002	0.3932	0.3932 (+0.0%)	0.1189 ± 0.0001	0.1190	0.9995
Rademacher	2.90	0.3561 ± 0.0001	0.3557	0.3820 (+7.4%)	0.1314 ± 0.0001	0.1315	1.0008
Laplace	3.40	0.4418 ± 0.0004	0.4421	0.4112 (−7.0%)	0.1034 ± 0.0000	0.1033	0.9994
centred exp.	3.70	0.4836 ± 0.0004	0.4839	0.4296 (−11.2%)	0.0906 ± 0.0001	0.0905	0.9999
sphere	2.90	0.3556 ± 0.0002	0.3557	0.3820 (+7.4%)	0.1315 ± 0.0001	0.1315	0.9996
Student t_5, $m=50$k	3.70	0.4769 ± 0.0012	0.4839	0.4296	0.0903 ± 0.0001	0.0905	0.9784
Student t_5, $m=200$k	3.70	0.4778 ± 0.0008	0.4839	0.4296	0.0905 ± 0.0000	0.0905	0.9784
Student t_5, $m=800$k	3.70	0.4795 ± 0.0006	0.4839	0.4296	0.0905 ± 0.0000	0.0905	0.9784

The standardized Student t_5 setup satisfies the assumptions of Propositions 3.2 and 4.1, including the finite fourth-moment requirement. Its larger numerical discrepancy is consistent with heavy-tailed Monte Carlo fluctuations: the eighth feature moment is infinite, so the empirical average of $\mathrm{Tr}(G_x^2)$ has infinite sampling variance and the usual finite-variance Monte Carlo error analysis does not apply. This does not invalidate the population risk formula, but it excludes the additional eighth-moment hypothesis of the finite-task central limit theorem; the standard deviation across three seeds reported above should therefore be read as descriptive rather than as a calibrated uncertainty estimate.

E.5 Numerical Check of Scalar Optimality (Proposition 3.2)

We test Proposition 3.2 by training an unrestricted fixed map $\hat{w}=BG_{xy}$ and the complete reduced model, with $p, q, \varphi, \bar{Q}$ all free, against the population-optimal scalar rule $\theta^* G_{xy}$.

Protocol. We use $n=40$, $d\in\{10,40,80\}$, and i.i.d. Gaussian, Rademacher, Laplace, or centred exponential feature coordinates, with kurtoses $3,1,6,9$ respectively. Tasks have $w\sim\mathcal{N}(0,I_d)$ and $\xi\sim\mathrm{Unif}(0,2)$, so $\nu=1$; the reference $\theta^*=\nu n/(1+\nu\psi_{d,n})$ uses each family's kurtosis. Both classes minimize squared prediction loss using Adam without weight decay, fresh batches of 1,024 tasks, and learning rate 10^{-3}, selected from $\{10^{-4}, 3\cdot10^{-4}, 10^{-3}\}$ at $d=40$. For each configuration we use five seeds and two initializations: small random parameters, or the scalar optimum perturbed by 1% noise, with no parameters frozen. Perturbed starts run for 10,000 steps; random starts run for 40,000 steps at $d\le 40$ and 160,000 at $d=80$. Checkpoints are at 5k and 10k for perturbed starts, 10k, 20k and 40k for random starts, with additional checkpoints through 160k at $d=80$. A single-seed schedule comparison at $d=80$ is followed by cosine-decay runs from 10^{-3} to 10^{-4} for all random-start seeds there. The reported set comprises 240 constant-rate runs and 40 decayed random-start runs.

Training, validation (20,000 tasks), and evaluation (50,000 tasks shared across rules within a configuration) use separate streams; the learning-rate sweep has its own evaluation seed. All results use

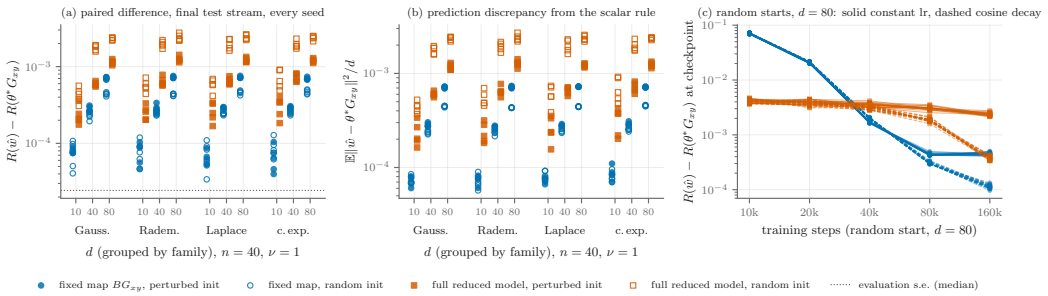

Figure 5: Proposition 3.2 under non-Gaussian designs, every seed shown, common evaluation stream. (a) Paired excess-risk difference between the learned class and the theoretical scalar rule on the same 50,000 tasks; positive means the learned class did not beat the scalar. (b) Prediction discrepancy $\mathbb{E}\|\hat{w} - \theta^* G_{xy}\|^2/d$; the two panels estimate quantities that agree in population by the orthogonality identity, and their Monte Carlo estimates need not coincide point by point. (c) Random starts at $d = 80$: paired difference at the checkpoints, every seed, constant learning rate (solid) and cosine decay (dashed); blue fixed map, orange full model.

the revised evaluation stream, which is monitored at every checkpoint, so schedule and convergence comparisons are exploratory rather than a once-only independent final test. The table reports checkpoint improvement divided by the final-reference evaluation SE as a descriptive scale, not an SE or significance test for the checkpoint change.

Results and convergence. All 280 reported runs have positive paired excess-risk differences from the scalar reference, ranging from 3.4×10^{-5} to 0.0028 (at most 0.4% of the reference excess risk); four are within three evaluation SEs of zero. Across the 48 constant-rate (family, dimension, class, start) cells, the seed-mean fitted scalar $\hat{\theta}/\theta^*$ is within 0.5% of one in 37 cells and within 2% in all cells. For perturbed starts, non-scalarness ranges from 0.010 to 0.061 across dimensions, and the directional terms carry 10^{-4}–8×10^{-3} of the scalar prediction energy. Table 7 gives the full-model results by family and dimension; the four feature laws show similar residuals despite their different kurtoses.

At $d = 80$, constant-rate random starts end with paired risk differences of 0.00043–0.00047 for the fixed map and 0.0023–0.0024 for the full model. Cosine decay reduces these to 0.00011–0.00012 and 0.00035–0.00038, respectively, with fitted scalars within 0.3% of the reference and non-scalarness 0.024 and 0.033. These residuals are below those of the constant-rate perturbed starts and continue to decrease near the end of the schedule (Figure 5c), consistent with an optimization contribution to the remaining error.

Interpretation. Neither learned class outperforms the scalar reference on the common evaluation stream, and the smaller residuals under longer training and learning-rate decay support the structural prediction without establishing convergence to the exact optimum. Training-seed standard deviations (Table 7) are distinct from evaluation SEs of paired differences (1–5×10^{-5}). The Student-t_5 features and Rademacher task prior proposed in the design document are not tested here.

E.6 Numerical Check of Task-Noise Dispersion (Proposition 5.1)

Proposition 5.1 states that the dispersion of excess conditional risk across task-noise levels, normalized by the noise-level variance, converges to the universal curve $K_\infty(\gamma, \nu) = [\nu^2\gamma/\{\gamma(1+\nu) + \nu\}^2]^2$. At finite size the same argument gives the exact identity $\mathrm{Var}_\xi[L(\theta^* \mid \xi) - \xi]/\mathrm{Var}(\xi) = K_{d,n} := (\theta^{*2}/(dn))^2$ under Assumption 3.1, because $L(\theta \mid \xi) - \xi$ is affine in ξ with slope $\theta^2/(dn)$ whatever the law of the i.i.d. coordinates; the coordinate law enters only through θ^*, via $\psi_{d,n} = (d + n + \kappa - 2)/d$ with κ the coordinate kurtosis. This experiment tests three things: the finite-sample identity at the design-specific θ^*, the collapse across noise-law shapes with matched noise variance, and the convergence to K_∞.

Table 7: Experiment 1, second pass, full reduced model, five seeds (mean ± s.d. across seeds). Fitted scalar from the perturbed and random starts, non-scalarness of B_{eff} (perturbed), the paired excess-risk difference ΔR on the common evaluation stream from the perturbed start and from the random start at the constant rate, the change of the random-start ΔR between its last two checkpoints divided by the final-reference evaluation SE (positive: a lower final value, not a significance test), and ΔR of the random start with cosine decay (run at $d = 80$).

family	d	$\hat{\theta}/\theta^*$ (full, pert.)	$\hat{\theta}/\theta^*$ (full, rand.)	$\|B_{\text{eff}} - \bar{b}I\|_F/\|B_{\text{eff}}\|_F$	ΔR (full, pert.)	ΔR (full, rand.)	improvement / final-reference SE (rand.)	ΔR (full, rand., cosine)
Gauss.	10	1.0004 ± 0.0045	0.9986 ± 0.0023	0.010 ± 0.001	0.00024 ± 0.00007	0.00043 ± 0.00010	4.0 ± 6.4	$--$
Gauss.	40	1.0010 ± 0.0035	0.9935 ± 0.0042	0.029 ± 0.001	0.00059 ± 0.00004	0.00177 ± 0.00015	3.8 ± 6.9	$--$
Gauss.	80	0.9913 ± 0.0024	0.9912 ± 0.0073	0.060 ± 0.000	0.00114 ± 0.00008	0.00231 ± 0.00009	17.9 ± 8.9	0.00035 ± 0.00001
Radem.	10	1.0011 ± 0.0054	1.0031 ± 0.0031	0.010 ± 0.001	0.00025 ± 0.00005	0.00057 ± 0.00010	0.8 ± 4.8	$--$
Radem.	40	0.9968 ± 0.0052	0.9949 ± 0.0014	0.028 ± 0.001	0.00064 ± 0.00009	0.00197 ± 0.00019	-0.9 ± 6.0	$--$
Radem.	80	0.9971 ± 0.0067	0.9858 ± 0.0041	0.059 ± 0.001	0.00125 ± 0.00011	0.00239 ± 0.00024	15.0 ± 9.9	0.00037 ± 0.00003
Laplace	10	1.0018 ± 0.0068	1.0035 ± 0.0067	0.011 ± 0.000	0.00026 ± 0.00007	0.00059 ± 0.00011	-2.2 ± 3.3	$--$
Laplace	40	0.9947 ± 0.0053	0.9926 ± 0.0016	0.029 ± 0.000	0.00065 ± 0.00007	0.00196 ± 0.00030	0.8 ± 7.3	$--$
Laplace	80	0.9931 ± 0.0110	0.9924 ± 0.0031	0.061 ± 0.000	0.00122 ± 0.00008	0.00241 ± 0.00018	13.9 ± 6.9	0.00038 ± 0.00004
c. exp.	10	1.0046 ± 0.0073	0.9962 ± 0.0059	0.012 ± 0.001	0.00030 ± 0.00009	0.00070 ± 0.00020	-3.9 ± 3.7	$--$
c. exp.	40	1.0005 ± 0.0079	0.9955 ± 0.0044	0.030 ± 0.000	0.00072 ± 0.00008	0.00193 ± 0.00023	-0.2 ± 4.6	$--$
c. exp.	80	0.9913 ± 0.0105	0.9883 ± 0.0025	0.061 ± 0.000	0.00120 ± 0.00007	0.00238 ± 0.00015	15.8 ± 4.1	0.00038 ± 0.00003

Setup. Four coordinate families with kurtosis $\kappa = 3, 1, 6, 9$ (Gaussian, Rademacher, Laplace, centred exponential), Gaussian w, $\nu \in \{0.25, 1, 4\}$ with $\bar{\xi} = 1/\nu$, $n \in \{40, 80, 160\}$, and 14 aspect ratios per (ν, n): the base grid $\gamma \in \{0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 1, 1.5, 2, 4\}$ plus four points around the predicted peak $\nu/(1+\nu)$, with d rounded to an integer and the realized d/n recorded; 504 cells in all. The scalar is fixed at the population coefficient θ^* of the family (no tuning at individual noise levels). The dispersion is never computed as the variance of individual task losses. For two-point noise $\xi = \bar{\xi}(1 \pm a)$ the normalized dispersion equals the squared slope of $r(\xi) = L(\theta^* \mid \xi) - \xi$, and the slope is estimated from the simulated errors as the mean of $\|\theta^* X^\top \epsilon/n\|^2/d$ on two independent batches of exactly 5,000 tasks each, whose product is the (unbiased) estimate $\hat{K}$; the theoretical slope is not inserted. For Gamma noise (shape a^{-2}, scale $\bar{\xi}a^2$, same mean and variance as the two-point law) 120 noise levels are drawn, two independent conditional-risk estimates $\hat{r}_A(\xi_j), \hat{r}_B(\xi_j)$ are formed from 200 tasks each, and $\hat{K}$ is their sample cross-covariance divided by the sample variance of ξ_j; the independence of the arms removes the inflation from inner Monte Carlo noise. Batches are drawn in memory-bounded chunks and concatenated, so every arm has its nominal size (an earlier pass capped the arm size in 36 large-d cells; the realized size is now stored with each cell). Everything is replicated five times with the same replicate seeds across γ, the per-replicate estimates are stored, and standard errors are across replicates, so a 2-standard-error interval has nominal coverage 88% (t_4 scale).

Finite-sample identity. Panel (b) of Figure 6: over all 504 cells the two-point estimate has $\hat{K}/K_{d,n}$ with median 0.9999 and range $[0.978, 1.024]$; 90% of cells are within 0.55% of $K_{d,n}$, the median relative standard error is 0.22%, and 87% (95%) of cells lie within two (three) standard errors, against the t_4 nominal values 88% (96%). The largest deviations are at $\gamma = 0.05$ ($d = 2$), where a single replicate carries little information. The estimated slope of $r(\xi)$ agrees with $\theta^{*2}/(dn)$ to a median ratio of 0.9999 (range $[0.989, 1.012]$), and panel (d) shows the affine dependence directly: both arms scatter around one line whose slope was not fitted. The identity holds at each family's own θ^*; because $\psi_{d,n}$ contains $(\kappa - 3)/d$, the finite-sample curves differ across families by $O(1/d)$, and the direction is as predicted (larger kurtosis, smaller θ^*, smaller dispersion): at the peak and $n = 40$, $K_{d,n}/K_\infty = 1.052, 0.952, 0.823, 0.715$ for $\kappa = 1, 3, 6, 9$, with the measured $\hat{K}/K_\infty$ within 0.006 of each.

Collapse across noise laws. Gamma noise at $a = 0.5$ was run in the 336 cells with $n \leq 80$; the ratio $\hat{K}/K_{d,n}$ has median 0.9997, median relative standard error 1.3% (the cross-covariance estimator is noisier than the two-point product), 90% of cells within 3.9%, and 86% within two standard errors. At the grid point nearest the peak, $a \in \{0.25, 0.5, 0.9\}$ were all run: for every family and ν the three Gamma estimates and the two-point estimate agree with $K_{d,n}$ within two standard errors (for example, Gaussian, $\nu = 1$, $n = 40$, $\gamma = 0.5$: 0.970 ± 0.029, 0.984 ± 0.013, 1.007 ± 0.005 and 1.0005). The normalized dispersion depends on the noise law only through $\mathrm{Var}(\xi)$, as the affine form implies.

Convergence to the universal curve. Panel (c): at the peak the ratio $\hat{K}/K_\infty$ moves toward 1 with a gap that halves as n doubles, from either side. Gaussian: 0.952, 0.977, 0.987 at $n = 40, 80, 160$

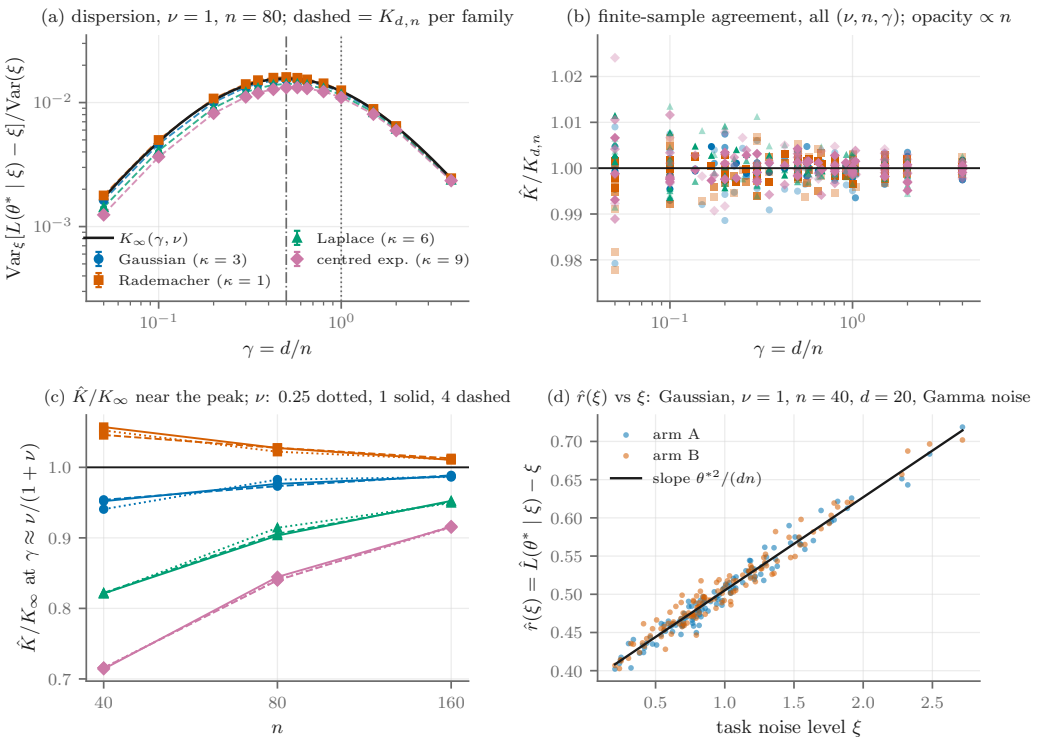

Figure 6: Dispersion of excess conditional risk at the population scaling, isotropic non-Gaussian features. (a) Normalized dispersion against γ at $\nu = 1$, $n = 80$: two-point estimates with standard errors (markers), the family-specific finite-sample value $K_{d,n}$ (dashed) and the universal limit K_∞ of Proposition 5.1 (solid); vertical lines at $\nu/(1+\nu)$ and at 1. (b) $\hat{K}/K_{d,n}$ for all 504 cells. (c) $\hat{K}/K_\infty$ at the grid point nearest the predicted peak, against n. (d) Conditional excess risk against the noise level for Gamma noise at one configuration, both independent arms shown, with the line of slope $\theta^{*2}/(dn)$ through the arm mean.

($\nu = 1$); Rademacher: 1.057, 1.027, 1.011; centred exponential: 0.714, 0.845, 0.916, the slowest because its $(\kappa - 3)/d$ correction is the largest. The rate is $O(1/n)$ at fixed γ, as $\psi_{d,n} \to 1 + 1/\gamma$ implies. The empirical maximum over the grid sits at $\nu/(1+\nu)$ for the Gaussian and Rademacher families at every n; for Laplace and centred exponential at $n = 40$ it sits one or two grid points to the right (0.575 at $\nu = 1$, 0.925 at $\nu = 4$), at or one grid point from the maximum of the finite-sample $K_{d,n}$, and it reaches $\nu/(1+\nu)$ by $n = 160$. The peak location is quoted with its replicate spread rather than as a single argmax: the argmax of each replicate's own curve (grid spacing 0.025–0.075 near the peak) ranges over $[0.500, 0.650]$ for Laplace and $[0.575, 0.650]$ for centred exponential at $\nu = 1$, $n = 40$ (mean-curve argmax 0.575), over $[0.800, 1.000]$ and $[0.925, 1.000]$ at $\nu = 4$, $n = 40$, and by $n = 160$ over at most two adjacent grid points that include $\nu/(1+\nu)$ for every family and ν (for Gaussian and Rademacher all replicates agree on $\nu/(1+\nu)$ in 10 of 18 cases: 3 of 9 Gaussian and 7 of 9 Rademacher cases). Beyond its maximum $\hat{K}$ decreases monotonically in 35 of the 36 (family, ν, n) curves, the exception being two adjacent points near the peak for Gaussian $\nu = 0.25$, $n = 40$ that differ by less than a standard error; the decline is throughout the overparameterized regime, with $\hat{K}(\gamma = 1)/\hat{K}(\gamma = 4) = 9.3, 5.0, 3.1$ for $\nu = 0.25, 1, 4$ against K_∞ ratios $9.4, 5.1, 3.2$.

Scope. The scalar is evaluated at the theoretical θ^* of each family, which is the object Proposition 5.1 is about; the experiment does not test that θ^* is the minimizer (that is Appendix E.5). Only isotropic features with i.i.d. coordinates of finite kurtosis were run; the heavy-tailed Student-t_5 case listed in the design as a separate numerical-stability check was not run for dispersion (its component-level behaviour is in Appendix E.4), and anisotropic features are treated separately in Appendices F.8 and F.9.

E.7 Scope of the Non-Gaussian Extensions

Feature kurtosis enters the isotropic finite-sample formulas through the single scalar $\psi_{d,n}$ of Proposition 4.1, and that dependence washes out of the proportional limits by Lemma B.1. The Marchenko–Pastur limit is a universality statement requiring only i.i.d. entries of finite variance, so it does not require Gaussian features. The structural scalar result uses the vanishing first and third moments of w, which is a moment/symmetry condition rather than a Gaussianity assumption.

Within the isotropic development, two results assume a Gaussian law explicitly, each for a stated technical reason: Proposition D.7 needs Gaussian features to evaluate inverse-Wishart moments, and Proposition C.6 needs Gaussian demonstration noise for the independence of the residual from the cross-moment. The anisotropic fixed-map result is first stated for Gaussian feature covariance, but Proposition F.2 extends its structural conclusion to centrally symmetric feature and task laws, so Gaussianity is not essential there. The description of ridge as *Bayes-optimal* in Section 6 is a statement about the task prior on w and not about the features; the inequalities of Theorem 6.1 themselves need only $\mathbb{E}[ww^\top] = \tau^2 I_d$.

F Beyond Isotropy: Theory and Numerical Checks

This section collects everything in the paper that leaves the isotropic setting $\mathbb{E}[xx^\top] = I_d/d$: the structure of the reduced-class optimum under a general covariance (Gaussian, and then merely centrally symmetric), the scalar rule's risk and noise sensitivity as functions of the spectral moments of the covariance, the behaviour of the scalar-to-ridge gap under ReLU random features, and the numerical checks.

F.1 Fixed-Map Optimality under Anisotropy

Isotropy is responsible for exactly one step of Proposition 3.2: forcing the optimal linear map to be a scalar. The confinement of the reduced class to a single fixed linear map of the cross-moment survives for every Gaussian feature covariance.

Proposition F.1 (Structure of the reduced-class optimum under anisotropy)**.** *Let Assumption 3.1 hold with item 3 replaced by $x^{(i)} \sim \mathcal{N}(0, \Sigma/d)$ for an arbitrary positive definite $\Sigma \in \mathbb{R}^{d\times d}$, let $w \sim \mathcal{N}(0, \tau^2 I_d)$ be independent of σ, and let σ satisfy Assumption 3.1(4). Then the minimum population risk over the reduced parameterization $(p, q, \varphi, \bar{Q})$ is attained by an estimator of the form $\hat{w} = BG_{xy}$ for some $B \in \mathbb{R}^{d\times d}$, realized with $p = q = 0_d$, $\varphi = 1$, and $\bar{Q} = B^\top$. In particular, the directional terms $\bar{Q}^\top G_x p$ and $\varphi q G_y$ never help. When $\Sigma = I_d$ the optimal B is a scalar multiple of I_d (Proposition 3.2); for general Σ it is not claimed to be.*

Proof of Proposition F.1. As in the proof of Proposition 3.2, every estimator in the reduced class has the form $\hat{w}_{A,b,B,c}(Z) = AG_x b + BG_{xy} + cG_y$ with $A = \bar{Q}^\top$, $b = p$, $B = \varphi\bar{Q}^\top + qp^\top$, and $c = \varphi q$, and it suffices to dominate every element of the larger class $\mathcal{F} = \{\hat{w}_{A,b,B,c}\}$ by an attainable rule of the form BG_{xy}.

Consider the sign map that sends $w \mapsto -w$ and $x^{(i)} \mapsto -x^{(i)}$ for every $i = 1, \dots, n+1$, while leaving σ and the $\epsilon^{(i)}$ unchanged. Each output $y^{(i)} = w^\top x^{(i)} + \sigma\epsilon^{(i)}$ is unchanged. Because w and $x^{(i)}$ are symmetric about the origin (centring alone would not suffice) and independent of (σ, ϵ), the joint law of $(w, \sigma, \{x^{(i)}, \epsilon^{(i)}\}_{i\le n+1})$ is invariant under this map for every positive definite Σ. Under the map, G_x and G_y are unchanged, whereas G_{xy} and the query feature $x^{(n+1)}$ change sign. Hence the transformed prediction is

$$\hat{w}_{A,b,B,c}(Z_-)^\top(-x^{(n+1)}) = \big(-AG_x b + BG_{xy} - cG_y\big)^\top x^{(n+1)} = \hat{w}_{-A,b,B,-c}(Z)^\top x^{(n+1)},$$

where Z_- denotes the transformed demonstration matrix. Invariance of the joint law therefore gives $L(A, b, B, c) = L(-A, b, B, -c)$, with L the population prediction risk as a function of the parameters.

The average of the two predictors is $\frac{1}{2}\{\hat{w}_{A,b,B,c} + \hat{w}_{-A,b,B,-c}\} = BG_{xy}$. By convexity of the squared loss in the prediction,

$$\mathbb{E}\big[(y^{(n+1)} - (BG_{xy})^\top x^{(n+1)})^2\big] \le \tfrac{1}{2}\{L(A, b, B, c) + L(-A, b, B, -c)\} = L(A, b, B, c).$$

Thus every element of $\mathcal{F}$ is weakly dominated by BG_{xy} with the same matrix B. The rule BG_{xy} is realized in the reduced parameterization by $p = q = 0_d$, $\varphi = 1$, and $\bar{Q} = B^\top$, since then $\hat{w} = \varphi \bar{Q}^\top G_{xy} = BG_{xy}$. Under isotropy, Step 1 of the proof of Proposition 3.2 forces the optimal B to be a scalar multiple of I_d; for general Σ the matrix $\mathbb{E}[G_{xy}G_{xy}^\top]$ is no longer a multiple of the identity, and no such reduction is claimed. □

Because ridge, $(G_x + \lambda I_d/n)^{-1}G_{xy}$, applies a map that depends on the realized demonstrations, it lies outside this class for every Σ; the comparison of Section 6 therefore measures the value of an escape from a genuine representational restriction of one linear-attention layer, and not merely of the imposed isotropy. Appendix F.4 measures, for a linear-attention layer trained within the class $\{BG_{xy}\}$, how far the learned map B departs from a scalar as Σ becomes ill-conditioned.

Gaussianity is not what drives Proposition F.1. The proof uses two properties only: the joint law is invariant under the simultaneous sign change $(w, x^{(1)}, \dots, x^{(n+1)}) \mapsto -(w, x^{(1)}, \dots, x^{(n+1)})$, and the squared loss is convex in the prediction. The first holds for any centrally symmetric feature and task laws; nothing else about their shape is used. The following statement records the extension, together with the explicit optimal map, which depends on the feature law only through its second and fourth moments.

Proposition F.2 (Fixed-map reduction under central symmetry)**.** *Let the demonstration and query features $x^{(1)}, \dots, x^{(n+1)}$ be i.i.d. and centrally symmetric ($x \overset{d}{=} -x$) with $C = \mathbb{E}[xx^\top] \succ 0$ and $\mathbb{E}\|x\|^4 < \infty$. Let the task vector w be centrally symmetric, independent of the features and of $(\sigma, \{\epsilon^{(i)}\})$, with $\mathbb{E}[ww^\top] = \tau^2 I_d$ and $\mathbb{E}\|w\|^4 < \infty$, where $\tau^2 > 0$. Let the standardized noises $\epsilon^{(i)}$ be i.i.d., centred, with unit variance and $\mathbb{E}[\epsilon^4] < \infty$, independent of everything else, and let σ be independent of all features, with $\xi = \sigma^2$ satisfying $\mathbb{E}[\xi^2] < \infty$. Then the population risk of every rule in the reduced parameterization $(p, q, \varphi, \bar{Q})$ is finite, and the minimum over that class is attained by $\hat{w} = B^* G_{xy}$, realized with $p = q = 0_d$, $\varphi = 1$, $\bar{Q} = B^{*\top}$, where*

$$B^* = \tau^2 C M^{-1}, \qquad M = \tau^2 \mathbb{E}[G_x^2] + \frac{\bar{\xi}}{n} C,$$
$$\mathbb{E}[G_x^2] = \Big(1 - \frac{1}{n}\Big) C^2 + \frac{1}{n} F, \qquad F := \mathbb{E}[\|x\|^2 xx^\top]. \tag{8}$$

For $x = C^{1/2}z$ with i.i.d. symmetric standardized coordinates of kurtosis κ, $F = \operatorname{Tr}(C)\, C + 2C^2 + (\kappa - 3)\, C^{1/2} \operatorname{diag}(C_{11}, \dots, C_{dd}) C^{1/2}$; for the radial mixture $x = RC^{1/2}z$ of Appendix F.2, $F = r\{\operatorname{Tr}(C)\, C + 2C^2\}$ with $r = \mathbb{E}[R^4]$. The Gaussian case is $\kappa = 3$, recovering Proposition F.1.

Proof. Finite risk. Each prediction in the reduced class is a linear combination of $(G_x b)^\top x^{(n+1)}$, $(BG_{xy})^\top x^{(n+1)}$ and $G_y\, c^\top x^{(n+1)}$. Because the query feature is independent of (w, σ) and the demonstrations, it suffices to bound the second moments of these coefficient vectors; multiplication by any fixed A or B preserves finiteness. We have

$$\mathbb{E}\|G_x b\|^2 \le \|b\|^2 \mathbb{E}\|x\|^4, \qquad \mathbb{E}\|G_{xy}\|^2 \le \mathbb{E}\|x\|^4 \mathbb{E}\|w\|^2 + \bar{\xi}\, \operatorname{Tr}(C)/n.$$

Jensen's inequality gives $\mathbb{E}[G_y^2] \le \mathbb{E}[y^4]$. Expanding $y = w^\top x + \sigma\epsilon$, the odd cross terms vanish by independence, centering of ϵ, and symmetry of w, so

$$\mathbb{E}[G_y^2] \le \mathbb{E}\|x\|^4 \mathbb{E}\|w\|^4 + 6\tau^2 \bar{\xi}\, \operatorname{Tr}(C) + \mathbb{E}[\xi^2]\mathbb{E}[\epsilon^4] < \infty.$$

The target has variance $\tau^2 \operatorname{Tr}(C) + \bar{\xi}$. Thus every rule in the reduced parameterization has finite risk.

Reduction. The sign map of the proof of Proposition F.1 leaves $(y^{(i)})_{i \le n+1}$, G_x and G_y unchanged and flips G_{xy} and $x^{(n+1)}$; invariance of the joint law now follows from central symmetry of x and w and their independence of (σ, ϵ). The identity $L(A, b, B, c) = L(-A, b, B, -c)$ and the convexity step are unchanged, so every rule is weakly dominated by BG_{xy} with the same B.

The optimal map. Write $G_{xy} = G_x w + X^\top \sigma\epsilon/n$. Conditioning on the demonstrations and averaging first over the centred noise and then over w, using only $\mathbb{E}[ww^\top] = \tau^2 I_d$ and independence,

$$\mathbb{E}\big[(BG_{xy} - w)^\top C (BG_{xy} - w)\big] = \tau^2 \big[\operatorname{Tr} C - 2 \operatorname{Tr}(CB\, \mathbb{E}G_x) + \operatorname{Tr}(B^\top CB\, \mathbb{E}G_x^2)\big] + \frac{\bar{\xi}}{n} \operatorname{Tr}(B^\top CB\, \mathbb{E}G_x),$$

with $\mathbb{E}G_x = C$ and, by independence across demonstrations, $\mathbb{E}G_x^2 = n^{-2}\sum_{i,j}\mathbb{E}[x^{(i)}x^{(i)\top}x^{(j)}x^{(j)\top}] = (1-1/n)C^2 + F/n$. This is a strictly convex quadratic in B whose gradient $-2\tau^2C^2 + 2CBM$ vanishes only at $B^* = \tau^2CM^{-1}$, because $M \succ 0$ and C is invertible. For independent standardized coordinates, $\mathbb{E}[z_iz_jz_kz_l] = \delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} + (\kappa-3)\delta_{ijkl}$ gives $\mathbb{E}[(z^\top Cz)zz^\top] = \mathrm{Tr}(C)I_d + 2C + (\kappa-3)\,\mathrm{diag}(C)$, and conjugating by $C^{1/2}$ gives the stated F; for the radial mixture $\|x\|^2xx^\top = R^4\|C^{1/2}z\|^2(C^{1/2}z)(C^{1/2}z)^\top$ with R independent of z. □

Centred exponential coordinates are not centrally symmetric and fall outside this statement; whether a skewed feature law still admits a fixed-map optimum is not claimed. For independent coordinates with $\kappa \neq 3$ the map B^* differs from its Gaussian counterpart at the same C through the diagonal correction in F, so the Gaussian B^* is not the optimum for a non-Gaussian design and must not be reused as its reference. Appendix F.5 tests this proposition by training the full reduced class on Rademacher and Laplace features.

Remark F.3 (Which isotropy does what). Three separate isotropy conditions enter Proposition 3.2, and they fail in different ways. Isotropy of the task prior, $\mathbb{E}[ww^\top] = \tau^2I_d$ (Assumption 3.1(5)), and of the feature second moment, $\mathbb{E}[x^{(i)}x^{(i)\top}] = I_d/d$, are what make the linear term in B depend only on $\mathrm{Tr}\,B$; if the feature covariance is not isotropic the optimal map is not even diagonal in general (Proposition F.1 and Appendix F.4). Isotropy of $\mathbb{E}[G_x^2]$ is a condition on a *fourth* moment, and $\mathbb{E}[x^{(i)}x^{(i)\top}] = I_d/d$ alone does not imply it: features with isotropic covariance but coordinate-dependent kurtosis give an optimal map that is diagonal but not scalar, because the diagonal entries of $\mathbb{E}[\|x\|^2xx^\top]$ then differ. This is why Assumption 3.1(3) is stated through identically distributed coordinates rather than through the covariance alone. The vanishing odd moments of w play no part in any of this; they are what removes the directional terms G_xb and G_y, and a mean-zero but skewed task prior already breaks the conclusion.

F.2 Scalar Risk and Task-Noise Dispersion under Anisotropy

This subsection extends the scalar risk calculation of Proposition 4.1 and the excess-risk dispersion result of Proposition 5.1 beyond isotropy without assuming Gaussian features. The optimum considered here is within the scalar class $\{\theta G_{xy}\}$; it is not a characterization of the optimum over the full reduced transformer class under these general assumptions. Even under Gaussian anisotropic features, the optimal fixed map in Proposition F.1 need not be scalar. Write $C = \mathbb{E}[xx^\top]$ for the feature second-moment matrix and normalize $\mathrm{Tr}(C) = 1$, so that the signal variance is τ^2 when the task vector is centered. Prediction excess risk is weighted by C: it is $\mathbb{E}[(\hat{w}-w)^\top C(\hat{w}-w)]$, rather than $\mathbb{E}\|\hat{w}-w\|^2/d$.

Proposition F.4 (Anisotropic scalar risk decomposition and optimal scaling)**.** *Let demonstration and query features be i.i.d., with $C = \mathbb{E}[xx^\top] \succ 0$, $\mathrm{Tr}(C) = 1$, and $\mathbb{E}\|x\|^4 < \infty$. Let (w,ξ) be independent of all features, with w independent of ξ and $\mathbb{E}[ww^\top] = \tau^2I_d$. Let the standardized observation noises be i.i.d., independent of the features and (w,ξ), with mean zero and variance one. Assume $\tau^2 > 0$, $\xi \geq 0$, and $\bar{\xi} = \mathbb{E}[\xi] \in (0,\infty)$. Define*

$$a_C := \mathrm{Tr}(C^2), \qquad h_C := \mathbb{E}[\|x\|^2x^\top Cx],$$
$$b_{C,n} := \left(1 - \frac{1}{n}\right)\mathrm{Tr}(C^3) + \frac{h_C}{n}.$$

For the scalar rule $\hat{w} = \theta G_{xy}$, the conditional prediction risk is

$$L_C(\theta \mid \xi) = \underbrace{\tau^2(1 - 2a_C\theta + b_{C,n}\theta^2)}_{\mathrm{Bias}_C^2(\theta)} + \underbrace{\frac{a_C\theta^2}{n}\xi}_{\mathrm{Var}_C(\theta|\xi)} + \xi. \tag{9}$$

Here squared bias averages demonstration noise conditional on the design and task vector before taking the covariance-weighted squared error, as in Proposition 4.1. The unique population-optimal

scalar coefficient and its risk are

$$\theta^*_{\text{aniso}} := \arg\min_\theta \mathbb{E}_\xi[L_C(\theta \mid \xi)] = \frac{\tau^2 a_C}{\tau^2 b_{C,n} + \bar{\xi} a_C / n}, \tag{10}$$

$$L_C(\theta^*_{\text{aniso}}) = \tau^2 + \bar{\xi} - \frac{\tau^4 a_C^2}{\tau^2 b_{C,n} + \bar{\xi} a_C / n}. \tag{11}$$

For Gaussian features $x \sim \mathcal{N}(0, C)$, these expressions specialize to

$$h_C = \operatorname{Tr}(C^2) + 2\operatorname{Tr}(C^3), \qquad b_{C,n} = \left(1 + \frac{1}{n}\right)\operatorname{Tr}(C^3) + \frac{\operatorname{Tr}(C^2)}{n}.$$

Proof. Independence across demonstrations gives

$$\mathbb{E}[G_x] = C, \qquad \mathbb{E}[G_x^2] = \left(1 - \frac{1}{n}\right) C^2 + \frac{1}{n}\mathbb{E}[\|x\|^2 x x^\top].$$

Thus $\operatorname{Tr}(C\mathbb{E}[G_x^2]) = b_{C,n}$. Averaging over demonstration noise first yields

$$\begin{aligned}
\operatorname{Bias}_C^2(\theta) &= \tau^2 \mathbb{E}\operatorname{Tr}\left[(\theta G_x - I_d)^\top C (\theta G_x - I_d)\right] \\
&= \tau^2 \left[\operatorname{Tr}(C) - 2\theta \operatorname{Tr}(C\mathbb{E}G_x) + \theta^2 \operatorname{Tr}(C\mathbb{E}G_x^2)\right], \\
\operatorname{Var}_C(\theta \mid \xi) &= \frac{\xi\theta^2}{n}\mathbb{E}\operatorname{Tr}(CG_x) = \frac{\xi\theta^2}{n}\operatorname{Tr}(C^2).
\end{aligned}$$

The independent query noise adds ξ, proving (9). Minimizing the strictly convex population quadratic gives (10) and (11). For Gaussian features, the fourth-moment identity $\mathbb{E}[\|x\|^2 x x^\top] = \operatorname{Tr}(C)C + 2C^2$ gives the stated specialization. □

Proposition F.5 (Anisotropic excess-risk dispersion and its proportional limit)**.** *Under Proposition F.4, additionally assume $\operatorname{Var}(\xi) < \infty$. For every deterministic scalar coefficient θ, including θ^*_{aniso},*

$$\operatorname{Var}_\xi[L_C(\theta \mid \xi) - \xi] = \left(\frac{a_C\theta^2}{n}\right)^2 \operatorname{Var}(\xi), \tag{12}$$

$$\operatorname{Var}_\xi[L_C(\theta \mid \xi)] = \left(1 + \frac{a_C\theta^2}{n}\right)^2 \operatorname{Var}(\xi). \tag{13}$$

The coefficient for excess-risk dispersion is the square of the normalized estimation variance, since $\mathbb{E}_\xi[\operatorname{Var}_C(\theta \mid \xi)]/\bar{\xi} = a_C\theta^2/n$.

Now consider $C_d = \Sigma_d/d$ with $\operatorname{Tr}(\Sigma_d) = d$ and

$$\frac{d}{n} \to \gamma \in (0, \infty), \qquad m_{k,d} := \frac{1}{d}\operatorname{Tr}(\Sigma_d^k) \to m_k < \infty, \quad k = 2, 3,$$
$$\chi_d := d\,\mathbb{E}[\|x\|^2 x^\top C_d x] \to \chi < \infty.$$

Keep $\nu = \tau^2/\bar{\xi} > 0$ fixed and assume $\operatorname{Var}(\xi) > 0$. Then

$$\frac{\theta^*_{\text{aniso}}}{d} \longrightarrow \frac{\nu m_2}{\nu m_3 + \gamma(\nu\chi + m_2)}, \tag{14}$$

$$\frac{\operatorname{Var}_\xi[L_{C_d}(\theta^*_{\text{aniso}} \mid \xi) - \xi]}{\operatorname{Var}(\xi)} \longrightarrow K(\gamma, \nu; m_2, m_3, \chi) := \left[\frac{\nu^2\gamma m_2^3}{\{\nu m_3 + \gamma(\nu\chi + m_2)\}^2}\right]^2. \tag{15}$$

For fixed (ν, m_2, m_3, χ), this limiting coefficient has a unique maximum at

$$\gamma_{\text{peak}} = \frac{\nu m_3}{\nu\chi + m_2},$$

is strictly increasing below that point and strictly decreasing above it, and has rates $\Theta(\gamma^2)$ as $\gamma \to 0$ and $\Theta(\gamma^{-2})$ as $\gamma \to \infty$. The peak lies above one if and only if $m_3 > \chi + m_2/\nu$.

Proof. The exact identities follow from the affine dependence of (9) on ξ. For the limit, substitute $a_{C_d} = m_{2,d}/d$ and

$$d^2 b_{C_d,n} = \left(1 - \frac{1}{n}\right) m_{3,d} + \frac{d}{n}\chi_d$$

into (10) and (12). At finite (d, n), writing $\gamma = d/n$, the normalized dispersion coefficient is

$$K_{d,n} = \left[\frac{\nu^2 \gamma m_{2,d}^3}{\{\nu(1-1/n)m_{3,d} + \gamma(\nu\chi_d + m_{2,d})\}^2}\right]^2.$$

Taking limits gives the displayed expressions. The positive square root of K is proportional to $\gamma/(A + B\gamma)^2$, where $A = \nu m_3 > 0$ and $B = \nu\chi + m_2 > 0$. Its derivative has the sign of $A - B\gamma$, proving the unique maximum and monotonicity; expansion at zero and infinity gives the rates. □

Gaussian features and recovery of the isotropic result. For Gaussian features, $\chi_d = m_{2,d} + 2m_{3,d}/d \to m_2$. In this case the exact finite-sample coefficient is

$$K_{d,n} = \left[\frac{\nu^2 \gamma m_{2,d}^3}{\{\nu(1+1/n)m_{3,d} + \gamma(1+\nu)m_{2,d}\}^2}\right]^2,$$

and the limiting peak is $\gamma_{\text{peak}} = \nu m_3/\{(1+\nu)m_2\}$. When $\Sigma_d = I_d$, $m_2 = m_3 = 1$, recovering Proposition 5.1. For nonnegative eigenvalues of mean one, $m_3 \geq m_2^2 \geq m_2$, so in this Gaussian case the peak cannot move to the left of its isotropic location at a fixed SNR. A spectrum with equal proportions of $1 - \delta$ and $1 + \delta$ has $m_2 = 1 + \delta^2$ and $m_3 = 1 + 3\delta^2$. At $\delta = 0.8$ and $\nu = 4$, the limiting peak is $292/205 \approx 1.424$, whereas its isotropic counterpart is 0.8. Consequently, dispersion can initially increase within the overparameterized regime. These comparisons hold the limiting feature moments fixed as γ varies.

Independent non-Gaussian coordinates give the same limit. Suppose $x = C_d^{1/2} z$, with the symmetric positive square root, where the coordinates of z are i.i.d., have mean zero, variance one, and a common finite fourth moment κ independent of d. The quadratic-form fourth-moment identity gives

$$h_{C_d} = \operatorname{Tr}(C_d^2) + 2\operatorname{Tr}(C_d^3) + (\kappa - 3)\sum_{j=1}^{d} (C_d)_{jj}(C_d^2)_{jj}.$$

If $\sup_d \|\Sigma_d\|_{\text{op}} < \infty$, then

$$0 \leq \sum_j (C_d)_{jj}(C_d^2)_{jj} \leq \|C_d\|_{\text{op}} \operatorname{Tr}(C_d^2) = O(d^{-2}),$$

so $\chi_d \to m_2$ whenever the spectral moments converge as above. Thus the Gaussian proportional-limit scaling and dispersion curve also hold for transformed Rademacher, Laplace, centered exponential, and standardized Student t_5 coordinates under these spectral conditions. Their finite-sample coefficients may differ through κ and the orientation of C_d in the independent-coordinate basis. No eighth-moment assumption is needed for these population identities or their limits; its absence for t_5 affects Monte Carlo precision as discussed in Appendix E.4.

Radial fluctuations can change the limit. Covariance alone does not determine the non-Gaussian limit. For each feature sample, let $x = RC_d^{1/2} z$, with $z \sim \mathcal{N}(0, I_d)$, independent R, $\mathbb{E}[R^2] = 1$, and $\mathbb{E}[R^4] = r < \infty$; draw independent copies of (R, z) across samples. Then $\mathbb{E}[xx^\top] = C_d$, but

$$h_{C_d} = r\{\operatorname{Tr}(C_d^2) + 2\operatorname{Tr}(C_d^3)\}, \qquad \chi = rm_2,$$

so the limiting peak is

$$\gamma_{\text{peak}} = \frac{\nu}{1 + \nu r}\frac{m_3}{m_2}.$$

When R^2 is nonconstant, $r > 1$: persistent radial fluctuations shift the peak relative to a Gaussian design with the same covariance. Finally, a covariance with a fixed fraction of its trace in one eigenvalue can have divergent $m_{2,d}$ and $m_{3,d}$, so the finite-moment proportional limit above does not cover such sequences; the finite-sample identities remain valid.

F.3 Scalar-to-Ridge Gaps under ReLU Random Features

Theorem 6.3 and Lemma D.3 explain why the scalar-to-ridge gap vanishes as $\gamma \to \infty$ under isotropic features: both rules approach the null predictor. The isotropic argument does not directly extend to feature laws with a few dominant second-moment eigenvalues. Ridge can scale these directions separately, whereas a scalar rule cannot. We compare uncentered ReLU random features with isotropic Gaussian features of the same dimension; this changes the feature law, not just its second moment.

Protocol. Raw inputs are $x \sim \mathcal{N}(0, I_{d_0})$ with $d_0 = 8$, and the features are $\phi(x) = \sqrt{2/p}\,\mathrm{ReLU}(Wx) \in \mathbb{R}^p$ with the rows of W uniform on the unit sphere, so that $\mathbb{E}\|\phi\|^2 = 1$ exactly; W is drawn once per (p, seed). The comparison arm uses $\phi \sim \mathcal{N}(0, I_p/p)$. In both arms, $y = \beta^\top \phi + \sigma\epsilon$, with independent $\beta \sim \mathcal{N}(0, \tau^2 I_p)$ and standard Gaussian observation noises, and $\tau^2 = 1$. The unconditional signal second moment is τ^2 in both arms, but the feature means, dependence, fourth moments, and second-moment matrix $\Sigma_\phi = \mathbb{E}[\phi\phi^\top]$ differ. For ReLU features, Σ_ϕ is not the centered covariance. Noise is constant within an arm, with $\sigma^2 = \tau^2/\nu$, $\nu \in \{1, 3\}$; $n = 16$ and $p \in \{2, 4, \ldots, 128\}$, so $\gamma = p/n$ ranges from $1/8$ to 8.

Three rules act on the features. The scalar rule $\hat{\theta} G_{xy}$ estimates the population target $\theta^* = \mathbb{E}[y_q q]/\mathbb{E}[q^2]$, $q = \phi_q^\top G_{xy}$, on 2,000 tuning tasks. The fixed-map population target is $B^* = \tau^2 \Sigma_\phi \{\tau^2 \mathbb{E}[G_x^2] + \sigma^2 \Sigma_\phi / n\}^{-1}$. Its moments are exact in the Gaussian arm, where $B^* = \theta^* I_p$; in the ReLU arm the evaluated map $\hat{B}$ uses 20,000 feature samples, for $p \le 64$. Ridge uses the exact Bayes penalty $\sigma^2/(n\tau^2)$, justified by the Gaussian task prior and Gaussian likelihood conditional on the features, and is computed in dual form. All rules are scored on a common draw of 2,000 evaluation tasks by conditional excess risk divided by σ^2, so the null predictor scores ν. We repeat over three seeds (three draws of W in the ReLU arm); the observed spread includes feature-map, tuning/moment-estimation, and evaluation variability.

Why the fixed-map reduction applies here. The ReLU features do not meet the feature-symmetry hypothesis of Proposition F.2, but the Gaussian task and noise laws supply a different symmetry. Conditional on W, simultaneously replace β and every observation noise, including the query noise, by their negatives while keeping all features fixed. The joint law is unchanged, y and G_{xy} change sign, and G_x, G_y do not. Writing the coefficient vector as $BG_{xy} + u$ with $u = AG_x b + cG_y$, symmetry of squared loss gives $L(BG_{xy} + u) = L(BG_{xy} - u)$. Convex averaging therefore shows that BG_{xy} weakly dominates the original rule. All relevant moments are finite; when $\Sigma_\phi \succ 0$, the fixed-map quadratic calculation in Proposition F.2 then gives the displayed B^* without requiring centered features. This identifies a theoretical population optimum; evaluating its moment-based estimate $\hat{B}$ does not by itself certify its exact risk.

Table 8: The random-feature comparison of Figure 7 in numbers ($n = 16$, three draws of W, seed mean $\pm$ standard deviation of the normalized excess risk). "Top share" is the largest eigenvalue of the ReLU feature second-moment matrix Σ_ϕ divided by its trace, and "eff. rank" is $\mathrm{Tr}(\Sigma_\phi)^2/\mathrm{Tr}(\Sigma_\phi^2)$, both estimated from 20,000 feature samples. The relative gap is $(R_{\mathrm{scalar}} - R_{\mathrm{ridge}})/R_{\mathrm{scalar}}$, given for the Gaussian arm at both signal levels and for the ReLU arm.

		ReLU second moment		Gaussian rel. gap		ReLU, $\nu = 1$				ReLU, $\nu = 3$			
p	γ	top share	eff. rank	$\nu = 1$	$\nu = 3$	scalar	fixed map	ridge	rel. gap	scalar	fixed map	ridge	rel. gap
2	0.125	0.60	1.9	0.45	0.74	0.306 ± 0.016	0.294 ± 0.022	0.144 ± 0.005	0.53	0.870 ± 0.057	0.854 ± 0.062	0.185 ± 0.009	0.79
4	0.25	0.50	2.8	0.33	0.62	0.436 ± 0.016	0.398 ± 0.019	0.254 ± 0.007	0.42	1.276 ± 0.046	1.056 ± 0.018	0.340 ± 0.009	0.73
8	0.5	0.44	3.8	0.19	0.46	0.542 ± 0.024	0.477 ± 0.014	0.360 ± 0.017	0.34	1.427 ± 0.091	1.241 ± 0.076	0.587 ± 0.032	0.59
16	1	0.38	5.0	0.08	0.23	0.566 ± 0.033	0.517 ± 0.028	0.431 ± 0.025	0.24	1.554 ± 0.049	1.347 ± 0.020	0.829 ± 0.008	0.47
32	2	0.37	5.6	0.02	0.08	0.576 ± 0.012	0.527 ± 0.015	0.464 ± 0.009	0.19	1.622 ± 0.019	1.383 ± 0.039	0.985 ± 0.002	0.39
64	4	0.35	6.3	0.01	0.02	0.624 ± 0.014	0.572 ± 0.010	0.522 ± 0.013	0.16	1.687 ± 0.053	1.492 ± 0.054	1.140 ± 0.050	0.32
128	8	—	—	0.00	0.01	0.611 ± 0.019	—	0.515 ± 0.017	0.16	1.732 ± 0.013	—	1.174 ± 0.003	0.32

Results and scope. The Gaussian scalar-to-ridge gap, divided by scalar excess risk, falls to approximately 0% ($\nu = 1$) and 1% ($\nu = 3$) at $\gamma = 8$. For ReLU features it remains about 16% and 32% at $\gamma = 4, 8$; the feature second moment has top-eigenvalue share 0.35–0.60 and effective rank rising only from 1.9 to 6.3 over $p = 2$ to 64 (Table 8). At $\gamma = 4$, the estimated fixed-map-to-ridge gap is 8.0% and 20.9% of scalar excess risk, respectively; no fixed-map value was measured at $\gamma = 8$. Thus the estimated fixed map narrows but does not close the observed gap. This comparison

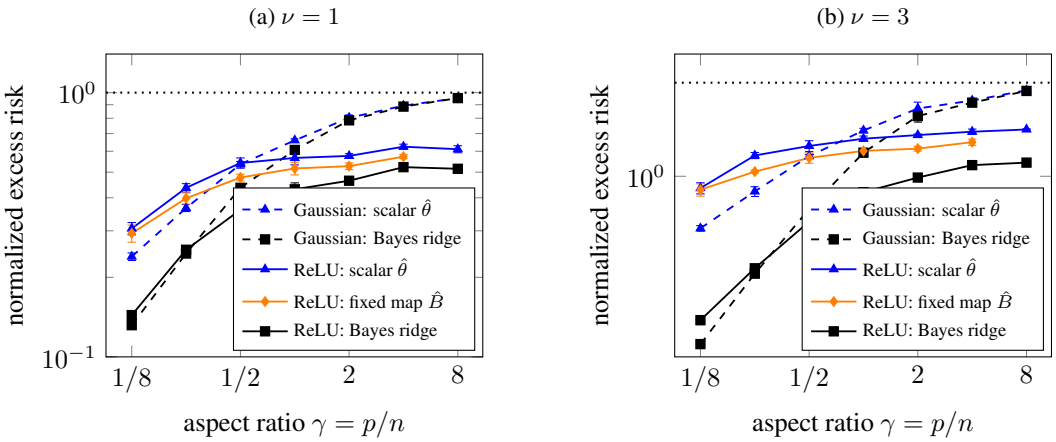

Figure 7: Normalized excess risk (conditional excess risk divided by σ^2) against the aspect ratio $\gamma = p/n$ at $n = 16$, for isotropic Gaussian features (dashed) and ReLU random features (solid) of the same dimension p, at $\nu = 1$ (left) and $\nu = 3$ (right). Curves: the empirically tuned scalar rule $\hat{\theta}G_{xy}$, the estimated fixed map $\hat{B}G_{xy}$ (ReLU arm, $p \leq 64$), and Bayes-optimal ridge; the corresponding population fixed-map optimum is justified above. The dotted line is the null predictor, whose normalized excess risk is ν. Error bars are standard deviations over three seeds, including feature-map and Monte Carlo variability; each point averages 2,000 evaluation tasks scored on a common draw. Over the tested range, the Gaussian curves approach the null predictor and each other, whereas the ReLU curves remain below the null predictor with a substantial scalar-to-ridge gap.

involves no transformer training, keeps $n = 16$ fixed, and establishes neither a nonvanishing asymptotic gap nor a covariance-only effect: ReLU also changes feature means and higher moments, and no covariance-matched Gaussian or centered-ReLU control was run. The scalar and ReLU fixed-map comparators are estimated, so their measured risks are not exact lower bounds for the one-layer class.

F.4 Numerical Check of Non-Scalar Fixed Maps

We train within the fixed-map class BG_{xy} to test how the learned map departs from a scalar under anisotropy. This does not test elimination of directional terms in the complete reduced class; Appendix F.5 addresses that question separately.

Protocol. With $d = 10$, $n = 20$, $\mathbb{E}[\sigma^2] = 1$ and $\tau^2 = 1$, features are drawn as $x \sim \mathcal{N}(0, \Sigma/d)$ with three spectra for Σ, each normalized to $\mathrm{Tr}(\Sigma) = d$ so that $\mathbb{E}\|x\|^2 = 1$ as in Assumption 3.1: isotropic; spiked, with two eigenvalues ten times the remaining eight ($\kappa(\Sigma) = 10$); and power-law, with eigenvalues proportional to $j^{-3/2}$ ($\kappa(\Sigma) \approx 32$). The fixed-map linear-attention model is trained for 4,000 AdamW steps at learning rate 2×10^{-3} with mini-batches of 1,000 tasks, three seeds each. Its effective map B is recovered by least-squares regression of $\hat{w}$ on G_{xy} over 4,000 contexts, which is exact because the model is linear in the cross-moment. We report the symmetric part $S = (B + B^\top)/2$, its normalized departure $\|S - \bar{b}I_d\|_F / \|S\|_F$ with $\bar{b} = \mathrm{Tr}(S)/d$, and its eigenvalues. This departure does not measure the antisymmetric part of B and differs from the full-map diagnostic in Appendix F.5.

Result. The symmetric-map departure is 0.013 ± 0.001, 0.398 ± 0.006, and 0.572 ± 0.003 for the isotropic, spiked, and power-law spectra, respectively. The learned symmetric part is therefore nearly scalar under isotropy and distinctly non-scalar in the two anisotropic cases; these three spectra do not establish a general monotone relation with covariance conditioning.

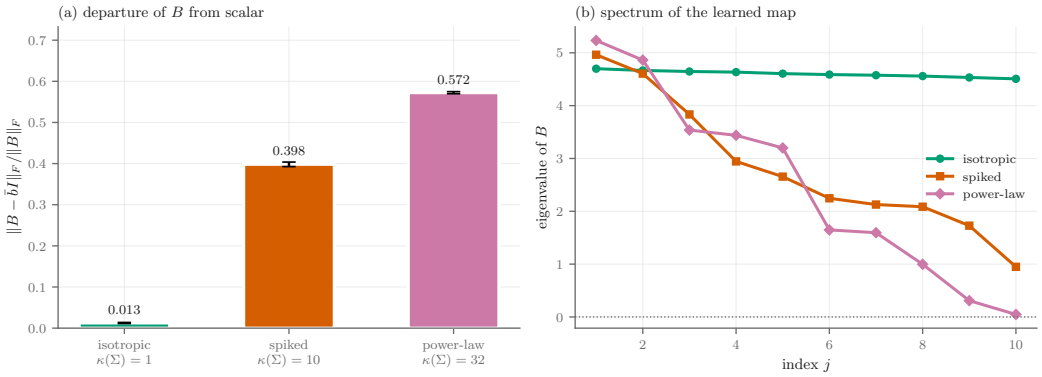

Figure 8: How far the learned linear-attention map departs from a scalar under anisotropy, at $d = 10$, $n = 20$, $\mathbb{E}[\sigma^2] = 1$, three seeds. Left: the normalized departure $\|S - \bar{b}I\|_F / \|S\|_F$ of $S = (B + B^\top)/2$ (the matrix denoted B in the plot's metric label). Right: the eigenvalues of the symmetric part of B, averaged over seeds. Under isotropy the spectrum is flat and the departure is 0.013; it rises with the conditioning of Σ to 0.398 for the spiked spectrum and 0.572 for the power-law spectrum.

F.5 Numerical Check of Fixed-Map Optimality (Proposition F.2)

We test whether the unrestricted fixed-map and complete reduced classes attain the analytical $B^* G_{xy}$ of Proposition F.2 under centrally symmetric non-Gaussian features. The scalar optimum is a restrictive benchmark; the design follows Appendix E.5 with covariance-weighted errors and family-specific fourth moments.

Setup. $n = 40$, two-level spectra $C = \mathrm{diag}(s)/d$ with $\delta \in \{0.5, 0.8\}$ and $d \in \{10, 40\}$, features $x = C^{1/2} z$ with i.i.d. Rademacher ($\kappa = 1$) or Laplace ($\kappa = 6$) coordinates, Gaussian controls at $(d, \delta) = (10, 0.5)$ and $(40, 0.8)$; $w \sim \mathcal{N}(0, I_d)$, $\xi \sim \mathrm{Unif}(0, 2)$. References are computed from (8) with each family's own F: the fixed map B^* and, within the scalar class, θ^*_{gen} of Proposition F.4. Two learned classes, the unrestricted fixed map $\hat{w} = B G_{xy}$ and the complete reduced model $\hat{w} = \bar{Q}^\top G_x p + (\varphi \bar{Q}^\top + q p^\top) G_{xy} + \varphi q G_y$ with $p, q, \varphi, \bar{Q}$ all free; squared query-prediction loss on fresh batches of 1,024 tasks, Adam, no weight decay, learning rate 10^{-3} for both classes (selected on a validation sweep over $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}\}$ at $d = 40$, $\delta = 0.8$, Laplace, where it was best or tied for every class and initialisation). Two initialisations: a perturbation of the theoretical optimum ($p = q = 0$, $\varphi = 1$, $\bar{Q} = B^{*\top}$ plus 1% noise, nothing frozen), and small random. As in Appendix E.5, training, validation (20,000 tasks, query-prediction loss at every checkpoint) and evaluation (50,000 tasks, shared by every rule of a configuration, never used by the learning-rate sweep) are separate random streams. The evaluation stream is reused at every checkpoint, so the convergence and schedule comparisons are exploratory rather than a once-only final test; budgets are 10,000 steps from the perturbed start, 40,000 from the random start at $d = 10$ and 160,000 at $d = 40$, with checkpoints, a constant-versus-cosine comparison on one seed (Laplace, $d = 40$, $\delta = 0.8$), and a re-run of the $d = 40$ random starts with cosine decay for all seeds. Five seeds per configuration: 200 constant-rate runs and 50 decayed random-start runs. (The first pass, whose evaluation seed was shared with the learning-rate sweep, was superseded; all numbers below are from the second pass.) All rules of a configuration are evaluated with the covariance-weighted excess risk $R(\hat{w}) = \mathbb{E}[(\hat{w} - w)^\top C (\hat{w} - w)]$; we report paired differences from $B^* G_{xy}$, the prediction discrepancy $\mathbb{E}[(\hat{w} - B^* G_{xy})^\top C (\hat{w} - B^* G_{xy})]$, the relative matrix error $\|B_{\mathrm{eff}} - B^*\|_F / \|B^*\|_F$ of the effective map $B_{\mathrm{eff}} = \varphi \bar{Q}^\top + q p^\top$, its non-scalarness $\|B_{\mathrm{eff}} - \bar{b} I\|_F / \|B_{\mathrm{eff}}\|_F$ next to that of B^*, and the energy of the directional term $h = \bar{Q}^\top G_x p + \varphi q G_y$ relative to the fixed-map term.

Results and convergence. All 250 reported paired risk differences from $B^* G_{xy}$ are positive, ranging from 2.8×10^{-5} to 0.0016; six lie within three evaluation SEs of zero (Figure 9, Table 9). From perturbed starts, the effective maps are within 1.1–1.3% of B^* at $d = 10$ and 3.3–4.2% at $d = 40$ in relative Frobenius norm, with non-scalarness within 0.003 of the reference and directional energy 1.3×10^{-4}–1.4×10^{-3} of the fixed-map term. The full-model residual uses 0.2–2% of the scalar-to-

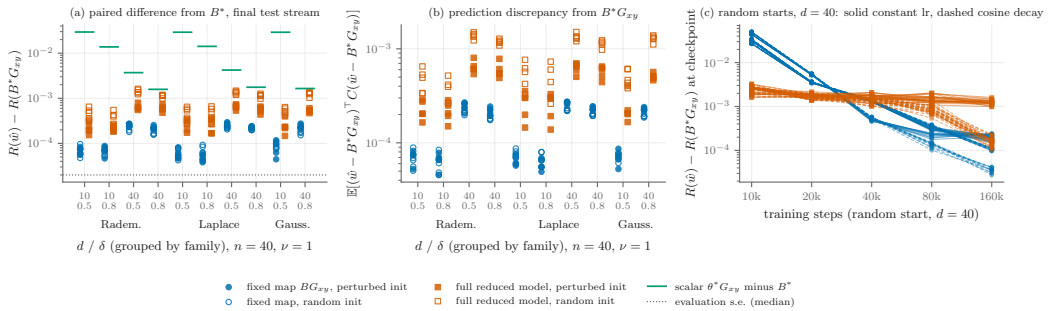

Figure 9: Fixed-map reduction under non-Gaussian anisotropic features, $n = 40$, $\nu = 1$, every seed shown, common evaluation stream. (a) Paired covariance-weighted excess risk of the learned rule minus that of the analytical fixed map B^*G_{xy} of (8); the green tick is the excess of the best scalar $\theta^*_{\text{gen}}G_{xy}$ over B^*, the gap the learned classes are meant to close; dotted line, median evaluation standard error. (b) Prediction discrepancy from B^*G_{xy}. Filled markers, perturbed initialisation (10,000 steps); hollow, random initialisation (40,000 steps at $d = 10$, 160,000 at $d = 40$). (c) Random starts at $d = 40$: paired excess over B^* at the checkpoints, constant rate (solid) and cosine decay (dashed); blue fixed map, orange full model.

Table 9: Full reduced model against the analytical fixed map, second pass, non-Gaussian anisotropic features, $n = 40$, $\nu = 1$; mean $\pm$ s.d. over five seeds. Matrix error of B_{eff} from the perturbed and the random (constant-rate) start, non-scalarness of B_{eff} (perturbed) next to that of B^*, directional energy, the paired covariance-weighted excess ΔR over B^*G_{xy} on the common evaluation stream from the perturbed start and from the random start at the constant rate, the change of the latter between its last two checkpoints divided by the final-reference SE (a descriptive scale, not an SE for the improvement), and ΔR of the random start with cosine decay (run at $d = 40$).

family	d	δ	$\|B_{\text{eff}} - B^*\|_F / \|B^*\|_F$ (pert. / rand.)	non-scalarness B_{eff} / B^*	dir. energy (rel.)	ΔR vs B^* (pert.)	ΔR vs B^* (rand., const.)	improvement / final-reference SE	ΔR (rand., cosine)
Radem.	10	0.5	0.011 ± 0.001 / 0.016 ± 0.002	0.314 ± 0.001 / 0.314	$2.0e-04 \pm 5.0e-05$	0.00022 ± 0.00006	0.00047 ± 0.00013	1.6 ± 6.9	– –
Radem.	10	0.8	0.013 ± 0.001 / 0.016 ± 0.002	0.467 ± 0.000 / 0.467	$1.8e-04 \pm 3.6e-05$	0.00022 ± 0.00003	0.00038 ± 0.00013	2.0 ± 7.6	– –
Radem.	40	0.5	0.033 ± 0.001 / 0.047 ± 0.001	0.164 ± 0.001 / 0.162	$1.0e-03 \pm 2.5e-04$	0.00064 ± 0.00009	0.00142 ± 0.00015	3.2 ± 4.2	0.00017 ± 0.00003
Radem.	40	0.8	0.041 ± 0.001 / 0.060 ± 0.002	0.255 ± 0.001 / 0.254	$8.3e-04 \pm 2.5e-04$	0.00060 ± 0.00009	0.00113 ± 0.00008	4.6 ± 4.4	0.00014 ± 0.00002
Laplace	10	0.5	0.012 ± 0.003 / 0.018 ± 0.001	0.325 ± 0.001 / 0.325	$2.7e-04 \pm 9.9e-05$	0.00026 ± 0.00007	0.00048 ± 0.00016	-0.0 ± 3.9	– –
Laplace	10	0.8	0.013 ± 0.003 / 0.018 ± 0.003	0.481 ± 0.001 / 0.482	$2.6e-04 \pm 1.1e-04$	0.00026 ± 0.00009	0.00047 ± 0.00016	-0.6 ± 4.6	– –
Laplace	40	0.5	0.034 ± 0.001 / 0.047 ± 0.001	0.176 ± 0.001 / 0.175	$9.9e-04 \pm 2.4e-04$	0.00065 ± 0.00008	0.00134 ± 0.00011	9.0 ± 7.7	0.00018 ± 0.00003
Laplace	40	0.8	0.042 ± 0.001 / 0.061 ± 0.001	0.274 ± 0.001 / 0.273	$7.9e-04 \pm 2.1e-04$	0.00056 ± 0.00009	0.00114 ± 0.00010	8.7 ± 9.9	0.00015 ± 0.00004
Gauss.	10	0.5	0.011 ± 0.002 / 0.016 ± 0.001	0.319 ± 0.001 / 0.319	$2.3e-04 \pm 1.0e-04$	0.00024 ± 0.00009	0.00053 ± 0.00009	1.7 ± 4.7	– –
Gauss.	40	0.8	0.041 ± 0.001 / 0.061 ± 0.001	0.263 ± 0.001 / 0.262	$6.3e-04 \pm 7.6e-05$	0.00053 ± 0.00005	0.00125 ± 0.00013	3.2 ± 4.1	0.00015 ± 0.00002

B^* gap at $d = 10$ and 15–40% at $d = 40$, where that gap is smaller. Family-specific fourth moments matter: using the Gaussian B^* on Laplace features costs 0.0022 at $(d, \delta) = (10, 0.8)$ and 0.0005 at $(40, 0.8)$, or 16% and 29% of the scalar-to-B^* gap.

At $d = 40$, constant-rate random starts end 0.00018–0.00022 above B^* for the fixed map and 0.0011–0.0014 for the full model. Cosine decay reduces these residuals to 0.00003–0.00011 and 0.00014–0.00018, below the constant-rate perturbed-start values, while reproducing the non-scalarness of B^*. The checkpoint improvement / final-reference SE in the table is descriptive, not an SE for the checkpoint difference. The schedule dependence supports an optimization contribution to the residual; it does not prove convergence or global optimality.

Scope. No trained rule outperforms B^* on the common evaluation stream, and its Monte Carlo risk agrees with the closed form within 0.0022 in every configuration. The task prior is Gaussian throughout. Rademacher and Laplace features satisfy the proposition's central symmetry; centred exponential features are outside its scope and were not tested. The proposed Rademacher task prior and radial-mixture extensions were not run.

F.6 Numerical Check of Scalar Risk under Gaussian Anisotropy (Proposition F.4)

We check the Gaussian specialization of Proposition F.4 using independently measured bias and variance, then assess the isotropic formulas outside isotropy. The latter comparison separates coefficient mistuning from the restriction to scalar rather than fixed-map prediction.

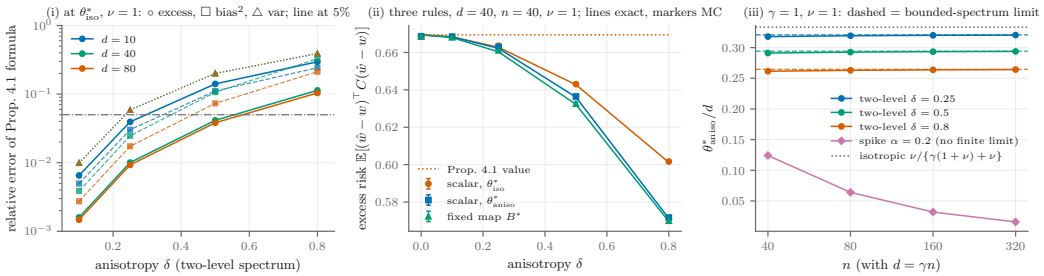

Figure 10: Anisotropic Gaussian features. (i) Relative error of the isotropic formulas of Proposition 4.1, evaluated at θ^*_{iso}, against the exact values of Proposition F.4; the variance error is exactly $\delta^2/(1+\delta^2)$ at every d and ν. (ii) Excess risk of the three rules at $d = 40$, $\nu = 1$: lines are the exact formulas of Proposition F.4 and of the fixed map, markers are Monte Carlo with standard errors, the dotted line is the value Proposition 4.1 would predict. The gap between the two scalar rules is the cost of the misspecified coefficient; the gap to B^* is the cost of the scalar restriction. (iii) $\theta^*_{\text{aniso}}/d$ from (10) at $\gamma = 1$ converges to the bounded-spectrum limit at rate $1/n$ and stays below the isotropic value; for the spike it halves as n doubles.

Setup. Features are Gaussian, $x \sim \mathcal{N}(0, C)$ with $C = U\,\mathrm{diag}(s)U^\top/d$ and $d^{-1}\sum_j s_j = 1$, in two families: the two-level spectrum $\{1-\delta, 1+\delta\}$ in equal proportion with $\delta \in \{0, 0.1, 0.25, 0.5, 0.8\}$, and a fixed-trace spike with eigenvalues $\{\alpha, (1-\alpha)/(d-1), \dots\}$ for $\alpha \in \{0.2, 0.5\}$. The grid is $n = 40$, $d \in \{10, 40, 80\}$, $\nu \in \{0.1, 1, 10\}$, Gaussian w and constant $\xi = 1/\nu$. Three rules are evaluated on the same 50,000 tasks: the scalar rule at the isotropic coefficient θ^*_{iso} of Proposition 4.1; the scalar rule at θ^*_{aniso} of (10); and the optimal fixed map $B^* = \tau^2 C M^{-1}$, $M = \tau^2 \mathbb{E}[G_x^2] + \bar{\xi} C/n$, of Appendix F.1. Squared bias and variance are estimated from independent components, $u = \theta G_x w - w$ and $v = \theta X^\top \epsilon/n$, averaging $u^\top C u$ and $\bar{\xi}\, v^\top C v$; no theoretical coefficient is inserted into a measured quantity.

Verification and scope. Across 63 configurations and both scalar coefficients, measured squared bias, variance, and excess risk agree with the anisotropic formulas within 0.7%, with median SE 0.1%; the fixed-map risk agrees to the same accuracy. The analytical optimum is inserted rather than fitted, so this checks risk at that coefficient, not its minimizing property. Appendix F.7 independently tests optimal scaling and extends the check beyond Gaussian features.

Accuracy of the isotropic approximation. For two-level spectra, the isotropic variance formula's relative error is exactly $\delta^2/(1+\delta^2)$, independent of d, ν. Bias errors partly offset it: at $\nu = 1$ the excess-risk approximation is within 5% for $\delta \le 0.25$ at every tested d, and for $\delta = 0.5$ at $d \ge 40$, but errs by 10–30% at $\delta = 0.8$. At $\nu = 1$, $\delta = 0.5$, coefficient mistuning costs 0.0466, 0.0066, 0.0016 at $d = 10, 40, 80$, while scalar restriction costs 0.0296, 0.0038, 0.0009. These decrease on the fixed-n grid without establishing asymptotic rates; the restriction cost is also non-monotone in δ, since low-eigenvalue directions receive little risk weight. For fixed-trace spikes, the isotropic coefficient overshoots by 2.7–13 at $d \ge 40$, formula errors reach 55–99%, and the mistuned rule can be worse than the null predictor (Table 10).

The separate proportional sequence at $\gamma = 1$ evaluates the closed-form coefficient, not additional Monte Carlo measurements. For two-level spectra, $\theta^*_{\text{aniso}}/d$ approaches its anisotropic limit with a discrepancy approximately halving as n doubles; that limit is 4%, 12%, 21% below the isotropic value at $\delta = 0.25, 0.5, 0.8$. For a fixed trace-fraction spike α, $a_C \to \alpha^2$ and $b_{C,n} \to \alpha^3$, giving $\theta^*_{\text{aniso}} \to 1/\alpha$ and $\theta^*_{\text{aniso}}/d \to 0$. This lies outside the bounded-spectral-moment limit, but the finite-sample formula remains valid.

Table 10: Anisotropic Gaussian features at $\nu = 1$, $n = 40$, two-level spectrum. Relative error of the isotropic excess-risk formula of Proposition 4.1 at θ^*_{iso} against the exact value of Proposition F.4; cost of the misspecified coefficient $R(\theta^*_{\text{iso}}) - R(\theta^*_{\text{aniso}})$; cost of the scalar restriction $R(\theta^*_{\text{aniso}}) - R(B^*)$; non-scalarness $\|B^* - \bar{b}I\|_F / \|B^*\|_F$. Costs are paired Monte Carlo means on 50,000 tasks; standard errors are at most 0.0012.

d	δ	rel. err. (excess)	coefficient cost	restriction cost	non-scalarness of B^*
10	0.1	0.007	0.0001	0.0029	0.067
10	0.25	0.039	0.0041	0.0144	0.166
10	0.5	0.141	0.0466	0.0296	0.319
10	0.8	0.296	0.1924	0.0138	0.474
40	0.1	0.002	0.0000	0.0004	0.034
40	0.25	0.010	0.0006	0.0019	0.084
40	0.5	0.041	0.0066	0.0038	0.167
40	0.8	0.113	0.0300	0.0016	0.262
80	0.1	0.001	0.0000	0.0001	0.020
80	0.25	0.009	0.0001	0.0004	0.051
80	0.5	0.038	0.0016	0.0009	0.101
80	0.8	0.104	0.0071	0.0004	0.161

F.7 Numerical check of Proposition F.4 beyond Gaussian features: the fourth moment h_C at work

We vary the feature law at fixed C to test the dependence on h_C in Proposition F.4, independently estimate the minimizing scalar, and quantify the cost of substituting the Gaussian fourth moment. The scalar form θG_{xy} is enforced; fixed-map optimality is tested in Appendix F.5.

Setup. $n = 40$, two-level spectra $C = \text{diag}(s)/d$ with $\delta \in \{0, 0.5, 0.8\}$, $d \in \{10, 40, 80\}$, $\nu = 1$; features $x = C^{1/2}z$ with i.i.d. Gaussian, Rademacher, Laplace or centred-exponential coordinates ($\kappa = 3, 1, 6, 9$), for which $h_C = \text{Tr}(C^2) + 2\,\text{Tr}(C^3) + (\kappa - 3)\sum_j C_{jj}(C^2)_{jj}$, and the radial Gaussian mixture $x = RC^{1/2}z$ with $R^2 \in \{1/2, 3\}$ with probabilities $4/5, 1/5$ ($\mathbb{E}R^2 = 1$, $r = \mathbb{E}R^4 = 2$), for which $h_C = r\{\text{Tr}(C^2) + 2\,\text{Tr}(C^3)\}$. At $(d, \delta) = (10, 0.8)$ and $(40, 0.5)$ the grid is extended to $\nu \in \{0.1, 10\}$, to a fixed random rotation $C = U\,\text{diag}(s)U^\top/d$, to $\xi \sim \text{Unif}(0, 2/\nu)$ in place of constant $\xi = 1/\nu$, and to the radial mixture; 72 cells in all. Three scalars are compared on the same tasks: θ^*_{gen} from (10) with the family's own h_C; θ^*_{gauss}, the same formula with the Gaussian h_C at the same C; and the empirical minimizer $\hat{\theta}_{\text{emp}} = \sum_t w_t^\top C g_t / \sum_t g_t^\top C g_t$, $g_t = G_{xy,t}$, computed on 50,000 independent tuning tasks, which uses no theoretical quantity. At the two extended configurations a fourth scalar is trained by Adam on the squared query-prediction loss (fresh batches of 4,096 tasks, 20,000 steps, step size $10^{-2}d/\sqrt{s}$) from three initial values $0.1d, 0.5d, 1.5d$, with convergence judged by the spread of the three end points and by the validation loss on 100,000 independent tasks. Evaluation: five replicates of 50,000 held-out tasks; squared bias and variance from the independent components $u^\top Cu$ and $\xi v^\top Cv$, excess risk as their sum, and paired differences between scalars on the same tasks. The general formulas are stored as references; nothing theoretical enters a measurement.

Risk formulas. Across 72 cells, the general formula agrees with measured squared bias, variance, and excess risk to maximum relative errors 0.31%, 0.41%, and 0.26%, respectively, including rotated covariances and the radial mixture. Constant and uniform task noise agree within 0.1%, consistent with dependence on $\bar{\xi}$ alone. At $(d, \delta) = (10, 0.8)$, the Gaussian formula's signed squared-bias errors are $+15.8\%$, -13.7%, -21.3%, and -24.0% for Rademacher, Laplace, centred exponential, and radial features. For bounded spectra, the independent-coordinate correction to h_C is relatively $O(d^{-1})$ and the measured errors decrease with d; rotation changes this correction without compromising the general formula. For the radial mixture, $h_C/h_C^{\text{gauss}} = r = 2$ persists at every dimension.

Optimal scaling. The independent empirical minimizer is within 0.54% of θ^*_{gen} in every cell (median 0.09%), with held-out paired excess risk at most 10^{-5}. Across the ten trained-scalar con-

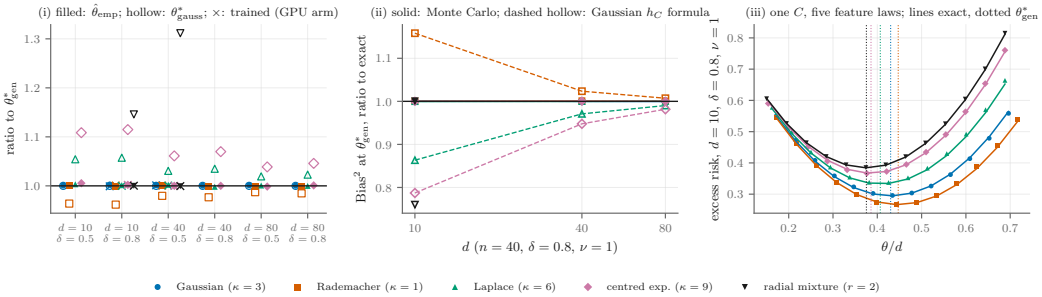

Figure 11: Proposition F.4 beyond Gaussian features, $n = 40$. (i) Optimal scaling: the empirical minimizer $\hat{\theta}_{\text{emp}}$ from an independent tuning sample (filled), the Gaussian-h_C coefficient θ^*_{gauss} (hollow) and the trained scalar (crosses), all as ratios to θ^*_{gen}, grouped by (d, δ), $\nu = 1$. (ii) Squared bias at θ^*_{gen}, $\delta = 0.8$: Monte Carlo over exact (solid) and the Gaussian-h_C formula over exact (dashed hollow), against d. (iii) Excess risk against θ/d for the five feature laws at one covariance ($d = 10$, $\delta = 0.8$): markers Monte Carlo on shared tasks, lines the exact general curves, dotted verticals θ^*_{gen}.

figurations, the fitted coefficients are 0.9991–1.0021 times the theoretical optimum and the three initializations end within 0.2–0.5% of one another. A separate final test stream of five replicates of $50{,}000$ common tasks, unused for training, tuning, or the earlier evaluation, gives trained-scalar excess at most 5×10^{-6} (SE at most 4×10^{-6}) and empirical-minimizer excess at most 6×10^{-6}. On that stream the Gaussian coefficient instead costs 0.00016–0.0272 for non-Gaussian families.

Gaussian shortcut and scope. Table 11 reports the mistuning cost: for independent coordinates it decreases over the tested dimensions and increases with signal strength; for the radial mixture it is 0.0131 at $(10, 0.8)$ and 0.0270 at $(40, 0.5)$. These radial cells change both dimension and anisotropy and establish no dimension-dependent rate. A persistent fourth-moment correction changes the proportional-limit coefficient, but does not imply a nonvanishing fixed-n risk penalty: with bounded spectra the scalar coefficients remain $O(1)$ and the mistuning cost is $O(d^{-1})$. Only two-level spectra were tested, with training and radial-mixture checks limited to the two extended configurations. The scalar-risk proposition requires no feature symmetry, so centred exponential features are covered here even though they are excluded from the centrally symmetric fixed-map proposition.

F.8 Numerical Check of Task-Noise Dispersion under Gaussian Anisotropy (Proposition F.5)

We test the Gaussian finite-sample dispersion formula of Proposition F.5, its limiting peak $\gamma_{\text{peak}} = \nu m_3/\{(1+\nu)m_2\}$, and optimal scaling. The spectral ratio m_3/m_2 shifts the peak relative to the isotropic value and can move it above one.

Setup. Gaussian features with the two-level spectrum $\{1-\delta, 1+\delta\}$, $\delta \in \{0, 0.25, 0.5, 0.8\}$ ($m_3/m_2 = 1,\ 1.12,\ 1.40,\ 1.78$), $\nu \in \{1, 4\}$, $n \in \{40, 80, 160\}$, aspect ratios at 0.1 spacing on $[0.1, 4]$, 0.05 spacing on $[0.8, 1.6]$ ($n \le 80$) and within 0.3 of the predicted peak; the grid maps to $1{,}245$ distinct (δ, ν, n, d) cells. Three scalars on the same tasks: θ^*_{aniso} of (10); the isotropic θ^* of Proposition 4.1; and the empirical minimizer $\hat{\theta}_{\text{emp}}$ of the covariance-weighted risk on an independent tuning batch of exactly $8{,}000$ tasks, without inserting the theoretical optimum. This empirical quadratic uses opposite-sign noise pairing, so its objective is the empirical bias component plus the mean-noise-weighted empirical variance component. Two-point noise $\xi = \bar{\xi}(1 \pm 0.5)$; the two-point product estimator of Appendix E.6 on two arms of exactly $3{,}000$ tasks, five replicates (3 at $n = 160$) with the same replicate seeds across γ, slope measured from simulated $v^\top C v$; batches are drawn in memory-bounded chunks so that every arm and tuning batch has its nominal size (an earlier pass capped 168 arm and 409 tuning batches), and the per-replicate estimates and realized sizes are stored with each cell.

Table 11: General against Gaussian fourth moment at the same covariance, $n = 40$, $\nu = 1$, constant ξ. $\theta^*_{\text{gauss}}/\theta^*_{\text{gen}}$; empirical minimizer over θ^*_{gen} (independent tuning sample); absolute relative error of the Gaussian-h_C squared-bias formula at θ^*_{gen}; Monte Carlo squared bias over the exact general value; paired excess risk of the Gaussian coefficient over the general one on $5 \times 50{,}000$ held-out tasks ($\pm$ standard error over replicates).

family	d	δ	$\theta^*_{\text{gauss}}/\theta^*_{\text{gen}}$	$\hat{\theta}_{\text{emp}}/\theta^*_{\text{gen}}$	Gauss. formula, abs. rel. err. Bias^2	MC Bias^2/exact	cost of Gaussian coef.
Gauss.	10	0.5	1.000	1.0008	0.000	1.0001	0.0000 ± 0.0000
Gauss.	10	0.8	1.000	1.0025	0.000	1.0011	0.0000 ± 0.0000
Gauss.	40	0.5	1.000	1.0014	0.000	0.9995	0.0000 ± 0.0000
Gauss.	40	0.8	1.000	1.0004	0.000	0.9998	-0.0000 ± 0.0000
Gauss.	80	0.5	1.000	1.0012	0.000	0.9995	0.0000 ± 0.0000
Gauss.	80	0.8	1.000	1.0005	0.000	0.9995	0.0000 ± 0.0000
Radem.	10	0.5	0.964	1.0013	0.105	0.9997	0.0009 ± 0.0000
Radem.	10	0.8	0.962	0.9994	0.158	1.0004	0.0011 ± 0.0000
Radem.	40	0.5	0.980	1.0005	0.015	1.0004	0.0002 ± 0.0000
Radem.	40	0.8	0.977	0.9991	0.024	1.0012	0.0002 ± 0.0000
Radem.	80	0.5	0.987	0.9989	0.004	0.9995	0.0000 ± 0.0000
Radem.	80	0.8	0.985	0.9993	0.007	1.0005	0.0001 ± 0.0000
Laplace	10	0.5	1.054	1.0032	0.101	0.9989	0.0018 ± 0.0000
Laplace	10	0.8	1.057	1.0014	0.137	0.9994	0.0021 ± 0.0001
Laplace	40	0.5	1.031	1.0022	0.019	0.9992	0.0003 ± 0.0000
Laplace	40	0.8	1.035	0.9981	0.029	0.9996	0.0005 ± 0.0000
Laplace	80	0.5	1.019	1.0007	0.006	1.0000	0.0001 ± 0.0000
Laplace	80	0.8	1.023	0.9999	0.010	1.0001	0.0001 ± 0.0000
c. exp.	10	0.5	1.109	1.0054	0.163	0.9988	0.0067 ± 0.0001
c. exp.	10	0.8	1.115	1.0021	0.213	1.0001	0.0084 ± 0.0001
c. exp.	40	0.5	1.061	1.0001	0.035	0.9993	0.0013 ± 0.0000
c. exp.	40	0.8	1.070	1.0001	0.052	1.0004	0.0019 ± 0.0000
c. exp.	80	0.5	1.039	0.9985	0.012	1.0002	0.0003 ± 0.0000
c. exp.	80	0.8	1.046	1.0009	0.018	1.0000	0.0006 ± 0.0000
radial	10	0.8	1.146	1.0000	0.240	1.0012	0.0131 ± 0.0001
radial	40	0.5	1.311	0.9992	0.100	1.0000	0.0270 ± 0.0001

Identity and optimal scaling. Across 1,245 cells, $\hat{K}/K_{d,n}$ has median 1.0000, range $[0.973, 1.021]$, and 90% of values within 0.53% of one. Two-/three-SE coverage is $89\%/96\%$ for five-replicate cells and $79\%/89\%$ for three-replicate cells, compared with the corresponding t_4 and t_2 nominal rates $88\%/96\%$ and $82\%/90\%$. The empirical minimizer is within 1.1% of θ^*_{aniso} at $n = 40$ and 0.4% at $n \geq 80$; its dispersion is within $[0.98, 1.04]$ of the theoretical-coefficient rule.

Peak and convergence. At $n = 160$, mean-curve peaks are $0.50, 0.60, 0.70, 0.89$ for $\nu = 1$ and $0.80, 0.95, 1.12, 1.40$ for $\nu = 4$ as δ increases; each matches the finite-sample grid maximum or an adjacent point. The peak is broad, and Figure 12 reports replicate argmax ranges rather than an exact empirical location. At $(\delta, \nu) = (0.8, 4)$, the paired contrast $\hat{K}(1.4) - \hat{K}(0.8)$ is 0.0053, 0.0053, 0.0055 for $n = 40, 80, 160$, with SEs 0.0001, 0.0001, 0.0002 and positive signs in every replicate, supporting the predicted shift above one. The isotropically tuned rule instead gives a negative contrast (-0.014), masking the shift and inflating dispersion by median factors $1.15, 1.56, 2.31$ at $\delta = 0.25, 0.5, 0.8$ (maximum 8.3). At the predicted peak, $\hat{K}/K_\infty$ rises from 0.94–0.96 at $n = 40$ to 0.984–0.992 at $n = 160$, consistent with the finite-size correction. These tests cover Gaussian two-level spectra and scalar rules only; non-Gaussian dispersion is tested next, and no dispersion claim for the fixed-map optimum follows.

F.9 Numerical Check of Task-Noise Dispersion under Non-Gaussian Anisotropy (Proposition F.5)

At the same covariance, Proposition F.5 predicts different dispersion curves through $\chi = \lim d h_{C_d}$. For the two-level spectrum with $\delta = 0.8$ and $\nu = 4$, independent-coordinate families have $\chi = m_2$ and limiting peak 1.424, whereas a radial Gaussian mixture with $r = 2$ has $\chi = 2m_2$ and peak 0.791. We test the finite-sample identity and convergence to peaks on opposite sides of one.

Setup. $C = \text{diag}(s)/d$ with $\delta = 0.8$, $\nu = 4$ ($\bar{\xi} = 1/4$), $n \in \{40, 80, 160\}$, even d with the realized d/n recorded; the aspect ratios $\{0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 2, 3, 4\}$ plus 0.05 spacing on $[0.6, 1.7]$ at $n \leq 80$ (28 points) and, at $n = 160$, the same refinement kept within 0.12 of

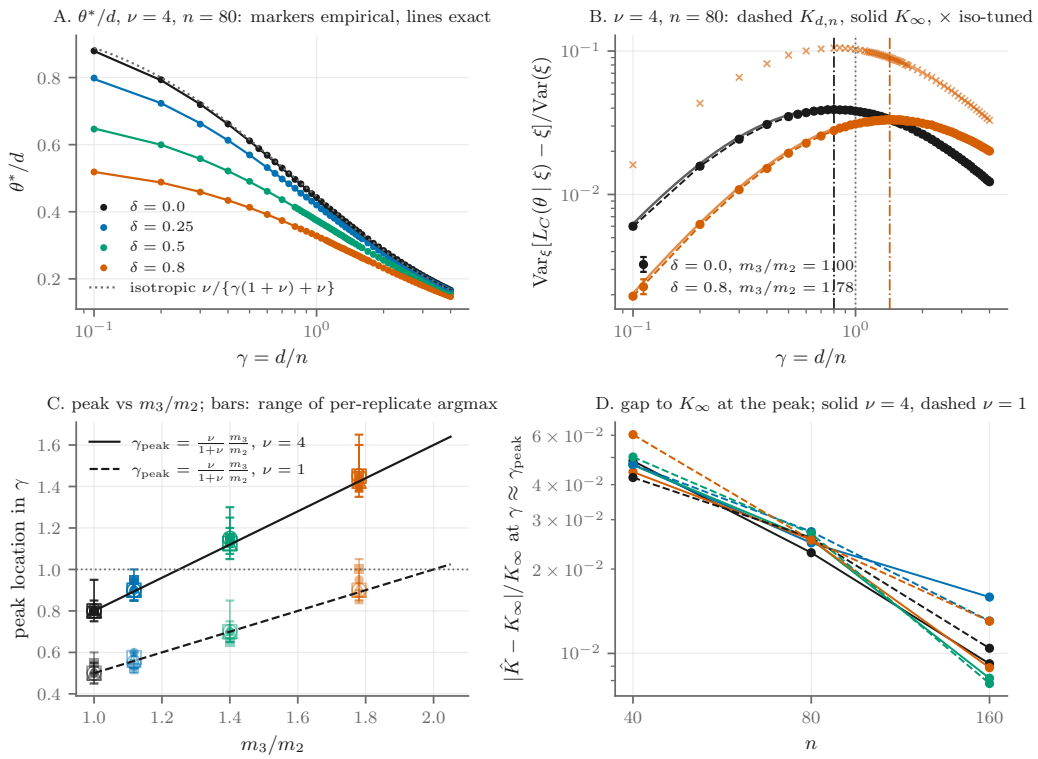

Figure 12: Anisotropic Gaussian dispersion, two-level spectra. A. Optimal scaling at $\nu = 4$, $n = 80$: the empirical minimizer from an independent tuning batch (markers) against the exact $\theta^*_{\mathrm{aniso}}/d$ (lines) and the isotropic value (dotted). B. Normalized dispersion at $\nu = 4$, $n = 80$ for $\delta = 0$ and 0.8: two-point estimates with standard errors at $\theta^*_{\mathrm{aniso}}$, the exact $K_{d,n}$ (dashed), K_∞ (solid), the isotropically tuned scalar ($\times$); vertical lines at the predicted peaks and at 1. C. Peak location against m_3/m_2 for every (δ, ν, n): grid argmax of $\hat{K}$ (filled) and of $K_{d,n}$ (hollow) with the lines $\gamma_{\mathrm{peak}} = \nu m_3/\{(1+\nu)m_2\}$. D. $|\hat{K} - K_\infty|/K_\infty$ at the grid point nearest the predicted peak against n.

the two predicted peaks (18 points); 296 cells. Four feature laws at every cell: Gaussian, Laplace ($\kappa = 6$), centred exponential ($\kappa = 9$) and the radial mixture $x = RC^{1/2}z$ with $R^2 \in \{1/2, 3\}$ with probabilities $4/5, 1/5$, R drawn independently for every feature sample. Each family is evaluated at its own θ^*_{gen} from (10) with its own h_C (a Gaussian coefficient would confound the distributional effect with mistuning; the Gaussian-coefficient rule is evaluated on the same tasks as a control). Two-point noise $\xi = \bar{\xi}(1 \pm 0.5)$; the estimator is the product of the slope estimates $\overline{v^\top C v}$ on two independent arms of exactly 5,000 tasks, replicated five times (3 at $n = 160$) with the same replicate seeds across γ, never inserting the theoretical slope $a_C\theta^2/n$; Gamma noise with the cross-covariance estimator of Appendix E.6 at the 34 cells within 0.11 of a predicted peak with $n \le 80$. Batches are drawn in memory-bounded chunks, so every arm has its nominal size (an earlier pass capped it in 36 large-d cells); the per-replicate estimates and realized sizes are stored with each cell. References: the exact finite-sample $K_{d,n} = [\nu^2\gamma m^3_{2,d}/\{\nu(1-1/n)m_{3,d} + \gamma(\nu\chi_d + m_{2,d})\}^2]^2$ with each family's χ_d, and the two limits K_∞ with $\chi = m_2$ and $\chi = 2m_2$.

Identity and peak shift. Across 296 cells, $\hat{K}/K_{d,n}$ has median 1.0000, range $[0.988, 1.013]$, and 90% of values within 0.52% of one. Two-/three-SE coverage is 88%/96% for five replicates and 83%/90% for three, close to the respective nominal rates; Gamma-noise checks give median ratio 1.002 across 34 cells. For Gaussian, Laplace, and centred-exponential features, the replicate-paired contrast $\hat{K}(1.4) - \hat{K}(0.8)$ is $+0.0048$ to $+0.0054$ (SE 0.5–1.5×10^{-4}), positive in all 13 replicates per family. For the radial mixture it is -0.0010, -0.0012, -0.0015 at $n = 40, 80, 160$ (SE 0.3–

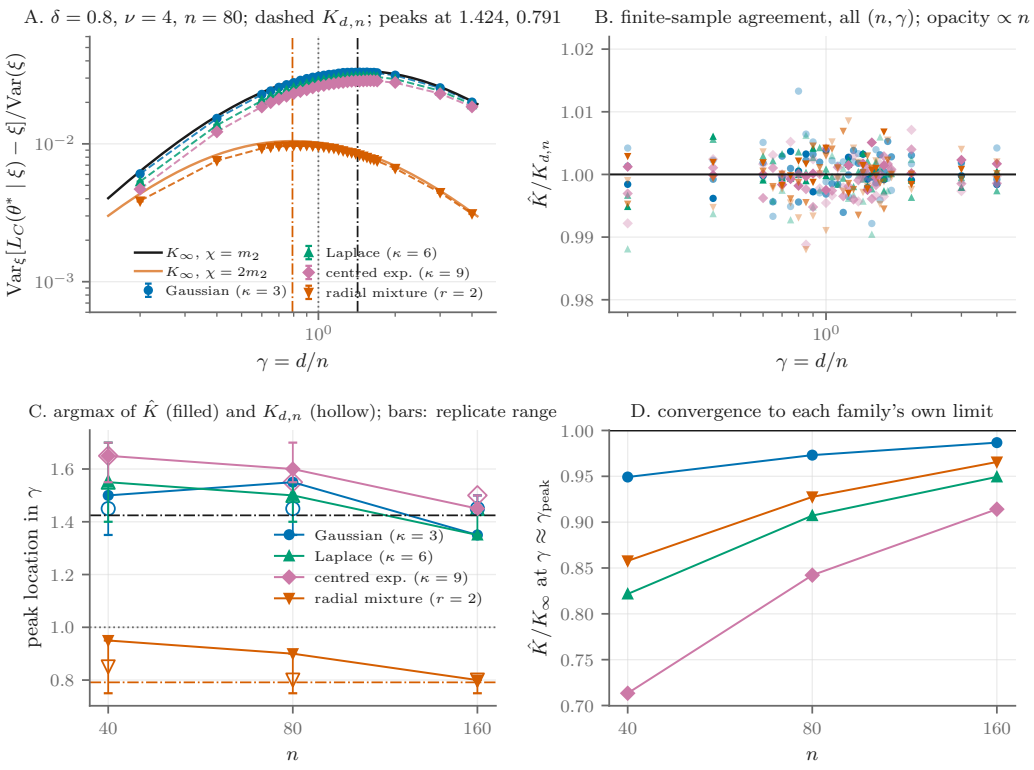

Figure 13: Dispersion under anisotropic non-Gaussian features at one covariance ($\delta = 0.8$, $\nu = 4$). A. Normalized dispersion against γ at $n = 80$: two-point estimates with standard errors (markers), the family-specific exact $K_{d,n}$ (dashed), the two limiting curves of Proposition F.5 for $\chi = m_2$ (black) and $\chi = 2m_2$ (orange); vertical dash-dotted lines at the predicted peaks 1.424 and 0.791, dotted at 1. B. $\hat{K}/K_{d,n}$ for all 296 cells. C. Grid argmax of $\hat{K}$ (filled) and of $K_{d,n}$ (hollow) against n; horizontal lines, the two limiting peaks and 1. D. $\hat{K}/K_\infty$ at the grid point nearest each family's own limiting peak.

0.4×10^{-4}), negative in every replicate. Thus identical covariance can yield dispersion peaks on opposite sides of one. At $n = 160$, mean-curve argmaxes are $1.35, 1.35, 1.45, 0.80$, all within 0.1 of their limits; replicate argmax ranges are shown in Figure 13. The broad peak limits grid resolution: for the radial mixture at $n = 40$, the mean-curve maximum is 0.95 versus theoretical 0.85, while replicate maxima span $[0.75, 0.95]$.

Convergence and mistuning. At each family's peak, $\hat{K}/K_\infty$ at $n = 160$ is 0.987 (Gaussian), 0.949 (Laplace), 0.914 (centred exponential), and 0.966 (radial), with discrepancies approximately halving as n doubles. Independent-coordinate families approach the same limit; the radial mixture approaches a different one. Using the Gaussian coefficient at $n = 80$ inflates radial dispersion near the peaks by 2.9–3.9 and excess risk by 0.041–0.054, versus 8–20% and 0.0002–0.0013 for Laplace and centred exponential. The dispersion ratios match $(\theta^*_{\mathrm{gauss}}/\theta^*_{\mathrm{gen}})^4$ to three decimals.

Scope. The test uses one spectrum, one SNR, and one radial law, with each scalar set to its theoretical optimum; independent tuning is checked in Appendix F.7. Rademacher features were not run. For this diagonal trace-one design they would give $\chi_d = \chi = m_2$ exactly, yet retain the leftward finite-size peak shift $(1 - 1/n)\nu m_3/\{(1 + \nu)m_2\}$. Appendix F.8 provides the broader Gaussian sweep.

G Multi-Layer Softmax Transformers: Experiments and Diagnostics

We report architectures, protocols, seeded performance comparisons, estimator probes, and task-noise adaptation for multi-layer softmax transformers. Shared data-generation and evaluation conventions are given first, in Appendix G.1.

G.1 Shared Data-Generation and Evaluation Conventions

The main-text and multi-layer experiments use synthetic in-context linear regression tasks from Section 3. Unless specified otherwise, features are Gaussian with second moment I_d/d, tasks satisfy $w \sim \mathcal{N}(0, \tau^2 I_d)$, demonstration noise is Gaussian, and σ^2 is shared within a task. The non-Gaussian and anisotropic experiments in Appendices E and F specify their own data-generating laws; each experiment's protocol takes precedence over these defaults.

The main-text figures and the sweeps in Appendix G.9 draw $\sigma^2 \sim \mathrm{Unif}(0, 2\tau^2/\nu)$ independently per task; Figure 1 (left) uses $\mathrm{Unif}(0, 2)$, with $\mathrm{Var}(\xi) = 1/3$. The seeded performance comparison and effective-estimator probe in Appendices G.5 and G.6 use constant noise at each setting. The noise-adaptation study and information 2×2 in Appendices G.7 and G.8 use a log-normal task-noise mixture.

The main-text figures report Monte Carlo estimates of population prediction risk and its squared-bias and variance components, normalized by $\mathbb{E}[\sigma^2]$, over $10{,}000$ independent tasks, with components estimated from 100 to $4{,}000$ resamples of demonstration noise conditional on the task. The seeded experiments report excess risk against the noiseless target $w^\top x^{(n+1)}$, excluding irreducible query noise and averaging over evaluation tasks. At a fixed noise level ξ, this estimates the conditional excess risk $L(\theta \mid \xi) - \xi$ for a scalar rule; evaluation under a noise mixture additionally averages over ξ.

G.2 Model Architectures

Three model classes appear:

1. *Reduced one-layer linear attention:* $T_{\varphi,\kappa}$, with two scalar parameters and $\hat{w} = \varphi\kappa G_{xy}$ by construction.
2. *Fixed-map linear attention:* $d \times d$ key–query and value blocks with $p = q = 0_d$, giving BG_{xy}; Appendix F.4 measures its departure from a scalar.
3. *Softmax attention:* L layers and H heads following Ahn et al. (2023), with full $(d+1) \times (d+1)$ matrices P_{ij}, Q_{ij} per head. For each column k, a head computes $a_{kl} = \mathrm{softmax}_l(z_k^\top Q_{ij} z_l)$ over all $n+1$ columns and outputs $n^{-1} \sum_{l \le n} a_{kl} P_{ij} z_l$. Thus the query enters the softmax denominator but its value is masked; normalization is not over demonstrations alone. Head outputs are summed into the residual stream, with no feedforward blocks, normalization layers, or positional encodings; prediction is entry $(d+1, n+1)$ after the last layer. Every head has width $d+1$, without splitting width across heads.

Figures 1 and 19 use 8 layers/4 heads, Figure 2 uses 6 layers/2 heads, and revision experiments use 4 layers/4 heads selected below. Main-text markers were not regenerated under the revision protocol; Appendix G.5 is a separate seeded comparison.

G.3 Protocol for the Main-Text Experiments

The finite markers of the main-text figures and of the sweep figures use

$$(n, d) \in \{(300, 3), (75, 3), (33, 3), (20, 5), (10, 10), (5, 20), (3, 33), (2, 50), (1, 100)\},$$

which spans $\gamma = d/n$ from 0.01 to 100. The last two points have two and one demonstrations; they test the finite-sample formulas of Proposition 4.1, which are exact at every (n, d), and not the proportional asymptotics. All models minimize the squared prediction loss with Adam (Kingma & Ba, 2015) at $\beta_1 = \beta_2 = 0.9$, with learning rates from 10^{-3} to 10^{-1} and mini-batch sizes from

4,000 to 10,000 tasks; reduced linear-attention models are trained for 1,000 to 10,000 steps and softmax-attention models for 100 to 2,000 steps, until the training loss has stabilized, and each marker averages the final 10 training steps. The learning rate of each run is encoded in the name of the corresponding figure file in `fig/`. Each marker is a single training run without repetition over seeds.

Figure 1. Both panels include practical-transformer markers from 8-layer, 4-head softmax-attention transformers. For the left panel, the task-risk dispersion is computed from 1,000 draws of the task-noise variance $\sigma^2 \sim \mathrm{Unif}(0, 2)$, so that $\mathbb{E}[\sigma^2] = 1$, and the theory curve is the asymptotic formula in Proposition 5.1.

Resolution of the dispersion peak. The grid in Figure 1 (left) has no point between $\gamma = 0.25$ and 1, whose predicted dispersion values at $\nu = 1$ are equal. It supports the non-monotone trend but cannot resolve the predicted peak at $\gamma = 0.5$ or distinguish it empirically from $\gamma = 1$.

Figure 2. The figure compares the closed forms of Proposition 4.1 (solid curves) with trained one-layer linear-attention transformers ($\times$), the optimal ridge benchmark of Proposition D.1 (red dashed overlays), and a 6-layer, 2-head softmax-attention transformer ($\blacklozenge$) trained under the common setup above with mini-batch size 10,000.

These settings are recorded as ranges because the runs predate the protocol of Appendix G.4, which fixes every learning rate, step count, batch size, seed, and Monte Carlo sample count exactly.

G.4 Protocol for the Seeded Experiments

Unless a subsection says otherwise, the revision experiments use $d = 10$ and $n = 20$, so $\gamma = 1/2$, with $\tau^2 = 1$. Every reported number is an average over the three training seeds $\{0, 1, 2\}$, and the quoted spread is the standard deviation across those seeds rather than across Monte Carlo draws. The random seed is set before the model is constructed, since initialization itself consumes the random stream.

Optimization. All models minimize the squared in-context prediction loss with AdamW (Loshchilov & Hutter, 2019) at $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, weight decay 0.01, and a constant learning rate, using a fresh mini-batch of 1,000 tasks at every step. The reduced model uses learning rate 2×10^{-3} for 4,000 steps. The softmax model uses learning rate 10^{-2}, with the entries of every P_{ij} and Q_{ij} initialized i.i.d. $\mathcal{N}(0, 0.02^2)$, for 12,000 steps in the probe and noise-adaptation experiments and for 6,000 steps in the seeded marker rerun.

Architecture selection. At $(n, d) = (20, 10)$ and $\nu = 1$, we compared learning rates $\{2 \times 10^{-3}, 5 \times 10^{-3}, 10^{-2}\}$ and 2-layer/2-head versus 4-layer/4-head architectures, two seeds each. Only the 4-layer/4-head model at 10^{-2} beat the reduced scalar's training loss (1.485 versus 1.515), motivating its use in the revision experiments.

G.5 Recovery of the Scalar-to-Ridge Risk Gap

This study uses one architecture and three seeds under Appendix G.4, with constant noise at each grid point. It is separate from the main-text experiments, whose architectures and heterogeneous noise differ: scalar/ridge population optima depend only on mean noise, but trained transformers need not behave identically.

Protocol. With $n = 20$, we vary $d \in \{2, 5, 10, 20, 40\}$ at $\nu = 1$, and $\mathbb{E}[\sigma^2] \in \{10, 3.16, 1, 0.316, 0.1\}$ at $d = 20$. Benchmarks are grid-tuned on 20,000 separate tasks: 400 values of $\theta \in [0.1, 20]$ and 120 log-spaced penalties in $[10^{-4}, 10^2]$, using $(G_x + \lambda I_d)^{-1} G_{xy}$. At $(n, d, \nu) = (20, 10, 1)$, tuning gives $\theta = 4.94$, $\lambda = 0.047$, versus exact values 4.88, 0.05. Each transformer trains for 6,000 steps; all rules share an evaluation draw of 8,000 tasks scored by conditional excess risk. Recovered share is $(R_{\mathrm{scalar}} - R_{\mathrm{tf}})/(R_{\mathrm{scalar}} - R_{\mathrm{ridge}})$, reported at the seed mean with its range over seeds. The references retain grid-tuning error, and seed ranges omit evaluation uncertainty; uncached per-task errors prevent retrospective paired uncertainty estimates, so small differences are descriptive.

Table 12: Seeded practical-transformer markers (4 layers, 4 heads, three seeds, $n = 20$) against the grid-tuned scalar and ridge references, all scored on a common evaluation draw of 8,000 tasks by conditional excess risk. The transformer column gives the seed mean and standard deviation; the last column gives the recovered share of the scalar-to-ridge gap with its range over seeds, or notes that the transformer lies above the scalar rule. The point $(\gamma, \nu) = (1, 1)$ appears in both sweeps.

sweep	(n, d)	γ	ν	scalar	ridge	transformer	recovered share
γ at $\nu = 1$	$(20, 2)$	0.1	1	0.206	0.107	0.134 ± 0.001	0.73 [0.72, 0.74]
	$(20, 5)$	0.25	1	0.370	0.247	0.286 ± 0.002	0.68 [0.67, 0.70]
	$(20, 10)$	0.5	1	0.533	0.436	0.489 ± 0.008	0.45 [0.37, 0.57]
	$(20, 20)$	1	1	0.669	0.613	0.660 ± 0.003	0.17 [0.11, 0.25]
	$(20, 40)$	2	1	0.832	0.811	0.839 ± 0.002	above scalar by 0.007
ν at $\gamma = 1$	$(20, 20)$	1	0.1	0.917	0.916	0.952 ± 0.002	above scalar by 0.036
	$(20, 20)$	1	0.316	0.805	0.795	0.827 ± 0.002	above scalar by 0.021
	$(20, 20)$	1	1	0.669	0.613	0.660 ± 0.003	0.17 [0.11, 0.25]
	$(20, 20)$	1	3.16	0.574	0.423	0.492 ± 0.004	0.54 [0.51, 0.57]
	$(20, 20)$	1	10	0.531	0.273	0.382 ± 0.008	0.58 [0.53, 0.60]

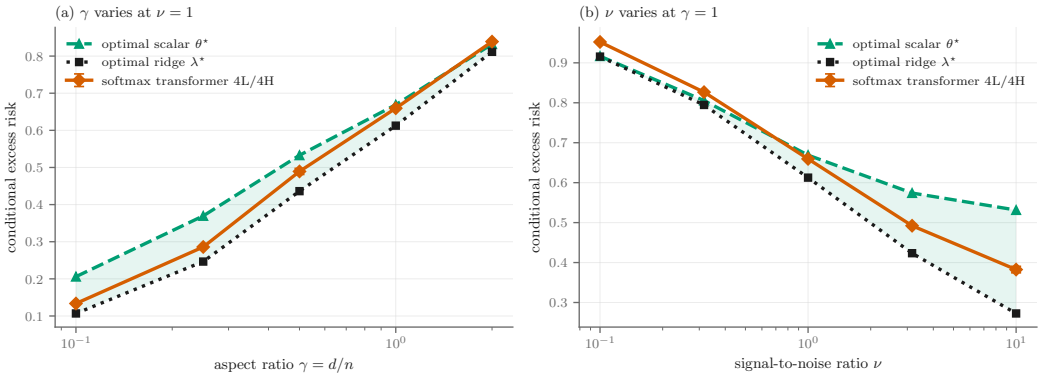

Figure 14: The seeded practical-transformer comparison of Table 12. Conditional excess risk of the grid-tuned scalar rule (dashed), grid-tuned ridge (dotted) and the 4-layer, 4-head softmax transformer (solid, error bars over three seeds), scored on a common evaluation draw. Left: γ varies at $\nu = 1$. Right: ν varies at $\gamma = 1$. The shaded band is the scalar-to-ridge gap. The transformer lies inside the band where the gap is large and above it at $\gamma = 2$ and at $\nu \leq 0.32$, where the gap is at most 0.02.

Results. Table 12 and Figure 14 show the transformer between the grid-tuned benchmarks at $\gamma \leq 1$, $\nu = 1$ and at $\gamma = 1$, $\nu \geq 1$, recovering 17–73% of the gap. At $\gamma = 2$ or $\nu \leq 0.32$, its point estimates exceed the scalar reference by 0.007–0.036, while the scalar-to-ridge gap is only 0.001–0.021. Seed spread is at most 0.008; these results support partial gap recovery where the benchmark separation is appreciable, without establishing every small ordering statistically. For seed 0, tripling training to 18,000 steps changes risk from 0.8390 to 0.8386 at $(\gamma, \nu) = (2, 1)$ and from 0.9495 to 0.9430 at $(1, 0.1)$, still above scalar values 0.8322 and 0.9167. This budget extension does not exclude other optimization or capacity explanations.

G.6 Effective-Estimator Diagnostics

We fit each trained model's predictions linearly in the query using 200 query draws for each of 40 fixed contexts. For $G_x = U \operatorname{diag}(s) U^\top$, write $a = U^\top \hat{w}$ and $b = U^\top G_{xy}$; pooled least squares compares $a = \theta b$ with $a = cb/(s + \lambda)$. We report the fraction of scalar-fit residual sum of squares removed by the ridge form, fitting (a, b) directly to avoid dividing by near-zero b_j. Controls recover a flat scalar profile and ridge penalties within 1% at $\lambda = 0.05, 0.5$; a noisy scalar control gives residual

reduction 0.003. Here s, λ use G_x units, so the Bayes penalty is $1/(\nu n) = 0.05$ at $(n, \nu) = (20, 1)$, with gain $c = 1$; $\gamma/\nu = 0.5$ instead uses dG_x units.

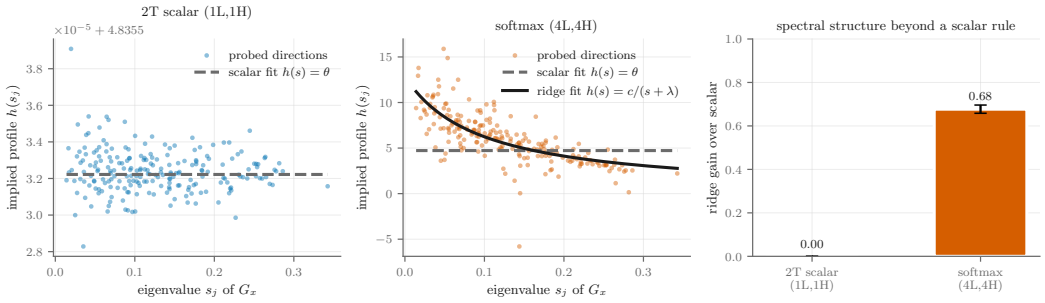

Figure 15: Spectral profile of the effective estimator recovered by the probe at $(n, d) = (20, 10)$, $\nu = 1$. Horizontal axis is the eigenvalue s_j of G_x; vertical axis is the implied profile $h(s_j)$. Left: the reduced model is flat to within 10^{-5}, as it must be, which is the negative control for the measurement. Middle: the trained softmax transformer decays with s_j and is tracked by the ridge form $c/(s + \lambda)$. Right: the fraction of the scalar fit's residual removed by the ridge form, with error bars over three seeds; the reduced model returns 0.000 and the softmax model 0.677 ± 0.019.

Results and limitations. The ridge form removes 0.677 ± 0.019 of the softmax model's scalar-fit residual, versus zero for the reduced model. Its fitted penalties are $0.094, 0.104, 0.091$ and gains $1.21, 1.24, 1.14$, indicating stronger low-eigenvalue shrinkage than Bayes ridge. This is an in-sample diagnostic of the query-linear approximation, not identification of an internal mechanism: neither fit uses held-out queries or contexts, and the ridge profile has two parameters against the scalar's one. A held-out comparison with a matched two-parameter profile and a multi-layer linear-attention control would help separate depth, softmax, and optimization effects. For example, zero-initialized least-squares gradient descent has profile $h_t(s) = \{1 - (1 - \eta s)^t\}/s$, $h_t(0) = \eta t$, giving the polynomial estimator $\eta \sum_{k=0}^{t-1} (I - \eta G_x)^k G_{xy}$.

G.7 Task-Noise Adaptation Diagnostics

Protocol. One model is trained on log-normal task noise with median 1 and unit log standard deviation, then evaluated at seven log-spaced levels from 0.1 to 10, using 8,000 tasks per level and conditional excess risk. A task-wise scalar reference is grid-tuned separately at each level on 20,000 tasks (400 values of $\theta \in [0.1, 20]$). The reference-level scalar is similarly tuned at $\sigma^2 = 1$, giving $\theta = 4.84$; the legacy figure label "population-tuned" denotes this median-level rule, not the mixture optimum $\theta^*(e^{1/2}) \approx 4.2$. The reduced model instead learns one scalar from the mixture by gradient descent.

Table 13: Conditional excess risk against the realized task-noise variance at $(n, d) = (20, 10)$, $\tau^2 = 1$, for models trained on the log-normal mixture (median 1, unit log standard deviation), with the two reference scalar rules. Model entries are the seed mean and standard deviation over three seeds.

σ^2	oracle scalar $\theta^*(\sigma^2)$	scalar tuned at $\sigma^2 = 1$	reduced model (mixture)	softmax 4L/4H (mixture)
0.10	0.376	0.417	0.433 ± 0.002	0.313 ± 0.025
0.22	0.403	0.420	0.466 ± 0.002	0.353 ± 0.027
0.46	0.442	0.456	0.491 ± 0.002	0.403 ± 0.021
1.0	0.509	0.514	0.526 ± 0.001	0.500 ± 0.013
2.2	0.604	0.644	0.615 ± 0.001	0.640 ± 0.002
4.6	0.756	0.936	0.869 ± 0.004	0.822 ± 0.013
10	0.848	1.497	1.344 ± 0.010	0.981 ± 0.034

Results and interpretation. The reduced model's risk grows approximately linearly with noise, consistent with a fixed scalar tuned to the mixture. Softmax point estimates fall below the empiri-

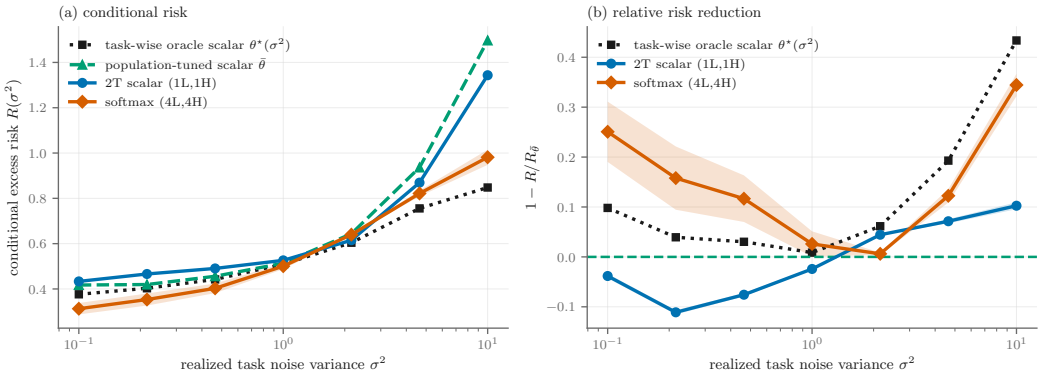

Figure 16: Adaptation to task-specific noise (Table 13). Left: conditional excess risk against the realized σ^2 for models trained on the log-normal mixture of noise levels. Right: risk reduction relative to the scalar rule tuned at $\sigma^2 = 1$ (the legacy legend's "population-tuned" curve denotes this reference-level scalar, not the mixture-optimal scalar), with the task-wise oracle shown for reference; normalizing by that rule's risk avoids the near-zero denominator that the difference between the two reference rules produces at $\sigma^2 \approx 1$. Shaded bands are ± 1 standard deviation over three training seeds, not intervals for Monte Carlo evaluation uncertainty.

Table 14: Reference rules at the seven noise levels of Table 13, at $(n, d) = (20, 10)$ and $\tau^2 = 1$: conditional excess risk against the noiseless target. The four scalar rules are exact, from (6) and, for the plug-in rule, Proposition C.6: the task-wise oracle $\theta^*(\sigma^2)$, the rule tuned at $\sigma^2 = 1$ ($\theta = 4.88$), the population optimum under the log-normal mixture $\theta^*(\bar{\xi})$ ($\theta = 4.21$), and the residual plug-in $\theta^*(\hat{\sigma}^2)$. The two ridge rules are Monte Carlo estimates over 20,000 tasks per level (standard error after $\pm$): population-tuned ridge with the exact mixture-optimal penalty $\bar{\xi}/(n\tau^2) = 0.082$, and task-wise ridge with the Bayes penalty $\sigma^2/(n\tau^2)$ at the realized level. The exact scalar values also calibrate the Monte Carlo error of Table 13, whose scalar columns differ from them by 0.01 to 0.09.

σ^2	oracle scalar	scalar tuned at 1	mixture scalar	plug-in scalar	population ridge	task-wise ridge
0.10	0.375	0.405	0.441	0.375	0.331 ± 0.004	0.089 ± 0.001
0.22	0.397	0.419	0.452	0.397	0.349 ± 0.004	0.166 ± 0.002
0.46	0.439	0.448	0.474	0.441	0.362 ± 0.004	0.268 ± 0.003
1.0	0.512	0.512	0.521	0.518	0.442 ± 0.005	0.422 ± 0.005
2.2	0.619	0.650	0.624	0.632	0.586 ± 0.007	0.578 ± 0.007
4.6	0.742	0.945	0.844	0.763	0.901 ± 0.010	0.722 ± 0.008
10	0.847	1.583	1.320	0.873	1.605 ± 0.019	0.846 ± 0.010

cally tuned oracle scalar at $\sigma^2 \leq 1$ and above it at $\sigma^2 \geq 2.2$; the roughly 2% difference at $\sigma^2 = 1$ is not statistically established. Seed deviations omit evaluation uncertainty, and Table 14 uses a separate draw, so cross-table comparisons are descriptive rather than paired tests.

A low-noise advantage over the scalar oracle need not imply noise inference: population-tuned ridge already achieves it through covariance adaptation. At $\sigma^2 = 0.1$, softmax risk 0.313 is near population-tuned ridge (0.331) and far above task-wise ridge (0.089). At $\sigma^2 = 4.6, 10$, softmax risks $0.822, 0.981$ are below population-tuned ridge's $0.901, 1.605$, but above task-wise ridge and the exact oracle scalar. This compares one fixed penalty and does not exclude all fixed-penalty rules; doing so would require held-out comparison to a fitted common penalty or direct diagnostics of context-dependent penalties. The residual plug-in scalar remains within about 0.03 of oracle-scalar risk; the largest transformer seed SD is 0.034 at $\sigma^2 = 10$. The results are consistent with covariance adaptation and partial noise adaptation, but do not establish explicit noise-variance inference.

G.8 The information 2×2: the two adaptations compound

We compare scalar versus ridge prediction and population versus task-wise noise tuning in one 2×2 experiment, evaluating all four rules on common tasks.

Protocol. We take $n = 20$, $d = 10$, $\tau^2 = 1$, and a log-normal task-noise law with median 1 and unit log standard deviation, so that $\bar{\xi} = e^{1/2} \approx 1.65$. The population-tuned scalar rule uses the single θ minimizing Monte Carlo risk under the mixture on a tuning draw of 40,000 tasks (400-point grid on $[0.1, 20]$; the minimizer is $\theta = 4.14$, against the closed form $\theta^*(\bar{\xi}) = 4.21$), and the population-tuned ridge rule uses the single penalty minimizing the same risk (200 log-spaced points on $[10^{-4}, 10^2]$; the minimizer is 0.084 in the convention $(G_x + \lambda I_d)^{-1} G_{xy}$, against the Bayes value $\bar{\xi}/(n\tau^2) = 0.082$). The task-wise oracle scalar rule uses $\theta^*(\xi)$ of (7) at the realized ξ, and the task-wise oracle ridge rule uses the Bayes penalty $\xi/(n\tau^2)$ at the realized ξ. All four rules are evaluated on an independent draw of 40,000 tasks by their conditional excess risk against the noiseless target.

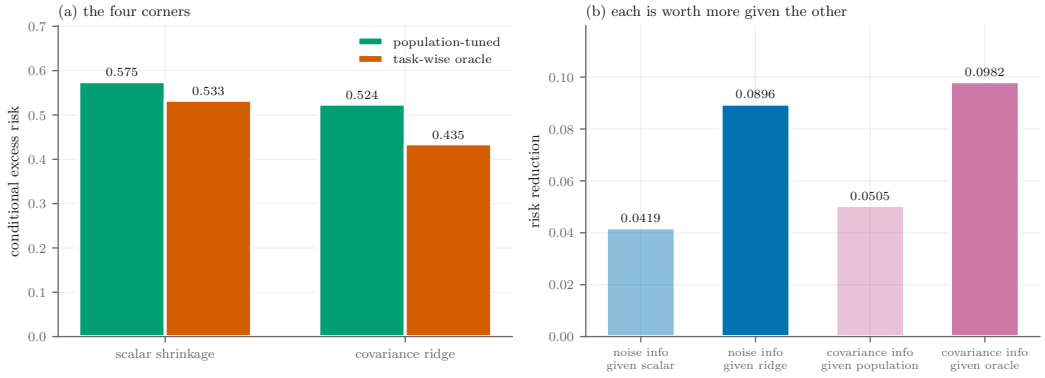

Figure 17: The information 2×2 at $n = 20$, $d = 10$, $\tau^2 = 1$, log-normal task noise with median 1 and unit log standard deviation. Left: conditional excess risk of the four rules. Right: the value of each kind of information given the other. Knowing the task's noise level is worth 0.042 to the scalar rule and 0.090 to ridge; the realized covariance is worth 0.051 at population tuning and 0.098 at task-wise tuning. Each piece of information is worth about twice as much in the presence of the other.

Result. The risks are 0.575, 0.533, 0.524, and 0.435 for population scalar, oracle scalar, population ridge, and oracle ridge, respectively (Figure 17). Noise information gains 0.042 for scalar and 0.090 for ridge; the positive interaction is 0.048 of the total 0.140 gain. This is configuration-specific, not implied by slope ordering: local mean gains depend on $-\frac{1}{2} L''_{\text{oracle}}(\bar{\xi}) \text{Var}(\xi)$. Appendices C.3 and C.6 give the corresponding nonlocal identities, the latter for the limiting ridge curve.

G.9 Additional Parameter Sweeps

Figure 18 checks scalar bias, variance, and risk on Gaussian tasks; Figures 20 and 21 extend its parameter ranges. Figure 19 compares 8-layer/4-head transformers with the benchmarks in small-gap regimes, using a different architecture from Figure 2. The separate seeded study is in Appendix G.5.

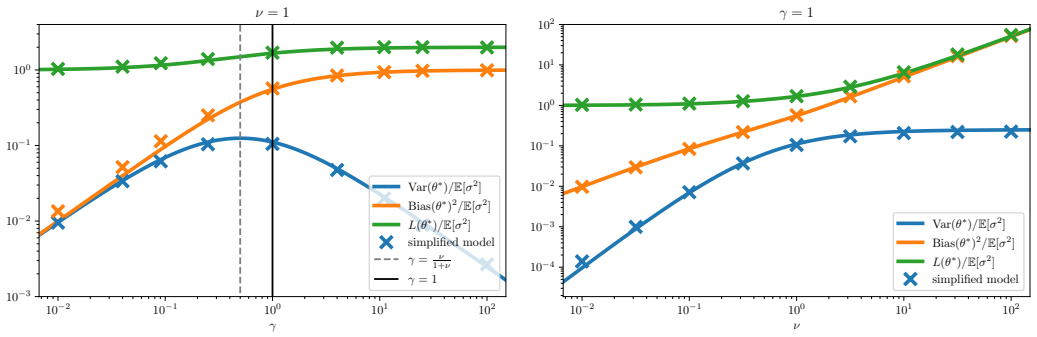

Figure 18: Empirical validation of the scalar-transformer formulas. Solid curves are closed-form predictions under optimal scaling; markers are estimates from trained one-layer linear-attention transformers on Gaussian tasks. Left: $\gamma = d/n$ varies at $\nu = 1$, with dashed and solid vertical lines marking $\gamma = \nu/(1+\nu)$ and $\gamma = 1$. Right: ν varies at $\gamma = 1$. The panels show normalized variance, squared bias, and risk. Markers are trained under squared prediction loss.

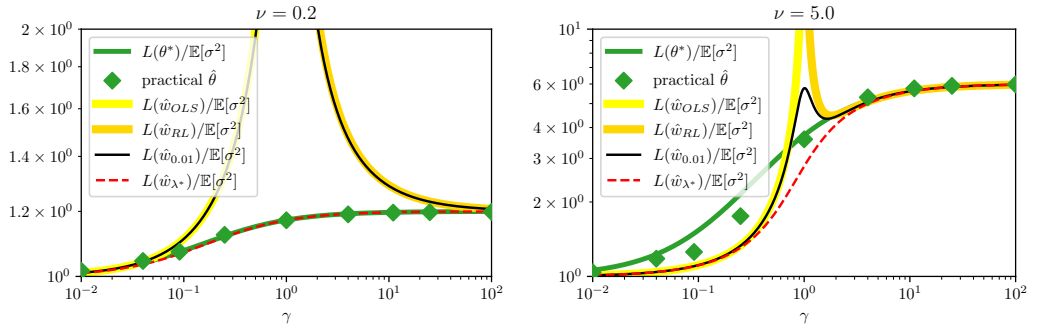

Figure 19: Additional practical-transformer comparisons related to Figure 2, with 8-layer, 4-head softmax-attention transformers. The scalar-to-ridge gap becomes small when $\gamma \gg 1$, $\gamma \ll 1$, and ν is small. Curves are asymptotic theoretical benchmarks, while markers are finite experiments with practical transformers.

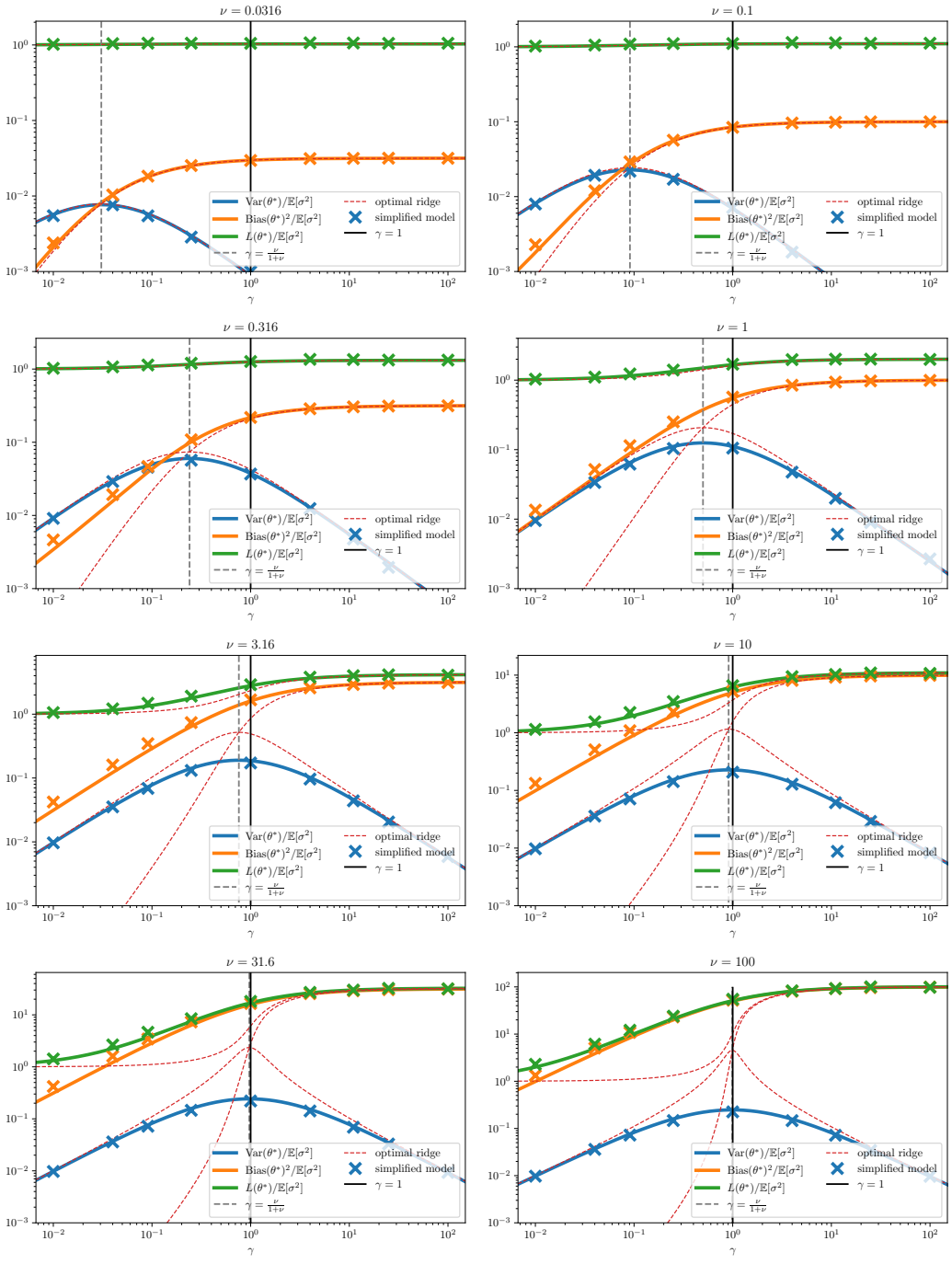

Figure 20: Additional panels for Figure 18 (left), with varying γ across a wider range of fixed ν values. The scalar-to-ridge gap becomes large in high-signal regimes and when γ is near 1.

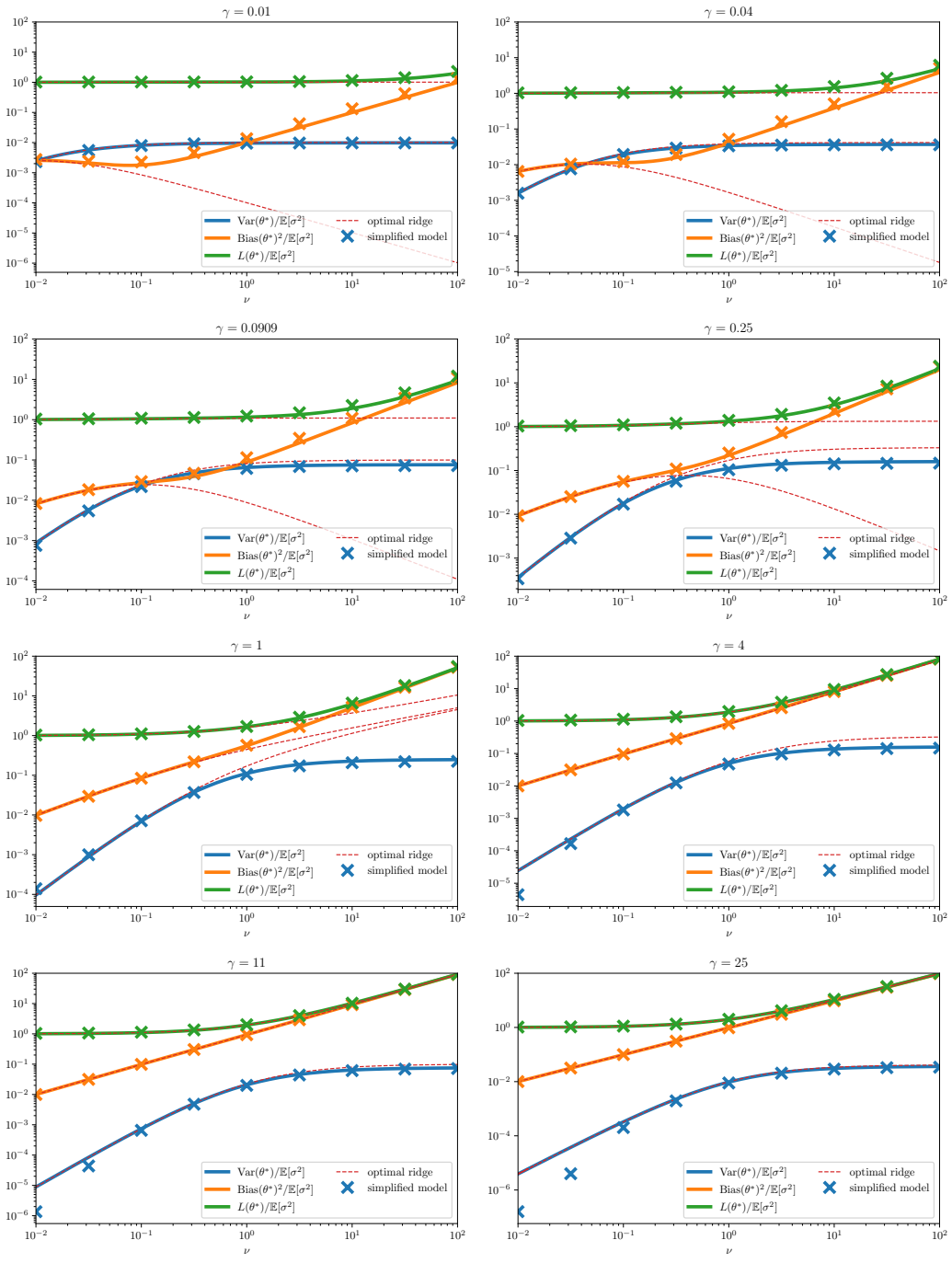

Figure 21: Additional panels for Figure 18 (right), with varying ν across a wider range of fixed γ values. The scalar-to-ridge gap becomes large when γ is near 1, especially in high-signal regimes. When γ is small, prediction risk remains close to the irreducible noise floor over a broad range of ν. As γ rises, both squared bias and total risk become more sensitive to changes in ν, consistent with the theoretical formulas.