WHEN BROADER PRETRAINING HURTS IN-CONTEXT LEARNING: A SPECTRAL DIAGNOSTIC FROM INVERSE LINEAR REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Broader pretraining can improve generalization in in-context learning, but adding source tasks can hurt performance on a fixed target family. We study this effect in ill-posed inverse linear regression (ILR), where the learner estimates a latent task vector from an underdetermined linear system. From a Bayesian viewpoint, pretraining provides a prior for this inference. For a Gaussian target prior with equal variance on a low-dimensional subspace and known rank and ridge ratio, learning the Bayes rule reduces to estimating that subspace. For block-diagonal source covariances, an eigengap determines whether additional tasks preserve the target subspace, create a tie, or displace target directions. We derive a sharp mixing threshold above which target-subspace recovery is lost. We prove this phase transition for a controlled two-stage ridge estimator and use a data-driven ridge estimator and a linear transformer to examine the same diagnostic empirically when estimation, architecture, and optimization also affect performance.

1 INTRODUCTION

Pretraining for in-context learning (ICL) is often broadened by adding related source task families. This is natural when target tasks are scarce: additional tasks may improve the learned representation and expand the range of tasks handled at test time. However, their effect on the original target family is less clear. They change the pretraining task distribution, and hence the prior for in-context inference. We ask when this change helps or hurts in-context learning on the original target family.

We study this question through ICL of ill-posed inverse linear regression (ICL-ILR). Each context is an underdetermined linear system whose latent coefficient vector must be estimated using information shared across tasks. The Bayesian formulation makes this role explicit: the Bayes-optimal in-context estimator is a regularized inverse map determined by the task prior. For a low-rank target prior with equal variance in every target direction and a known ridge ratio, this inverse map depends on the prior only through the target subspace.

The difficulty is to recover that subspace without observing the task vectors. Our key identity shows that the covariance of taskwise minimum-norm least-squares estimates (mLSEs) is a positive multiple of the task-prior covariance plus an isotropic term. Under the block-diagonal source assumption, target-subspace recovery reduces to the sign and size of the informative eigengap: the smallest eigenvalue on the target subspace minus the largest off-target eigenvalue of the pooled mLSE covariance.

Informal main result. Let $\nu \in [0, 1]$ be the fraction of source tasks in the pooled pretraining data. Assume that $\Sigma_0 = \lambda \Pi_\star$, where $\lambda > 0$ and $\Pi_\star$ projects orthogonally onto the target subspace, and that Σ_s is block-diagonal relative to this subspace and its complement. Let β_{on} and β_{off} be its smallest on-target and largest off-target eigenvalues. Additional source tasks can increase or decrease the recovery margin and, when off-target variance is sufficiently strong, prevent target-subspace recovery. These effects are quantified by the gap ratio:

$$R_\nu := \frac{c_{\mathrm{gap},\nu}}{c_{\mathrm{gap},0}} = 1 + \nu \left(\frac{\beta_{\mathrm{on}} - \beta_{\mathrm{off}}}{\lambda} - 1 \right), \qquad c_{\mathrm{gap},0} = a\lambda, \tag{1}$$

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.

Table 1: Effect of broader pretraining on the target-subspace recovery margin, for $0 < \nu \leq 1$. Transfer here is classified by the gap ratio R_ν, rather than by a comparison of finite-sample risks.

Recovery margin	Gap ratio	Condition	Target subspace
Nondecreasing	$R_\nu \geq 1$	$\beta_{\mathrm{on}} - \beta_{\mathrm{off}} \geq \lambda$	Uniquely recovered
Smaller, positive	$0 < R_\nu < 1$	$\lambda(1-\nu^{-1}) < \beta_{\mathrm{on}} - \beta_{\mathrm{off}} < \lambda$	Uniquely recovered
Zero or negative	$R_\nu \leq 0$	$\beta_{\mathrm{on}} - \beta_{\mathrm{off}} \leq \lambda(1-\nu^{-1})$	A valid tied choice at zero; cannot be selected below zero

where $c_{\mathrm{gap},\nu}$ is the informative eigengap of the pooled mLSE covariance and $a > 0$ depends only on the context length and ambient dimension. A larger positive gap tolerates larger covariance fluctuations in empirical approximations. Thus $R_\nu > 1$ enlarges the recovery margin, $R_\nu = 1$ leaves it unchanged, and $0 < R_\nu < 1$ reduces it while preserving the target subspace. At $R_\nu = 0$ there is a tie; $R_\nu < 0$ forces target-subspace loss (Table 1).

Main contributions. We establish two results for the pooled leading-subspace estimator in Gaussian ICL-ILR.

1. *Population recovery.* The covariance identity gives an exact condition for recovering the target subspace and Bayes rule, with a sharp mixing threshold when off-target source variance exceeds the smallest on-target source variance.
2. *Finite-sample behavior.* For a controlled two-stage ridge estimator (cTRE) with known rank and ridge ratio, we prove a target excess-risk bound of order $1/(Mc^2_{\mathrm{gap},\nu})$ for positive gaps, with other parameters fixed. For purely off-support sources under the stated dimension condition, the risk remains positive beyond the threshold as the number M of pretraining tasks grows.

Extensions allow source covariance between the target subspace and its complement, and explain why unequal target eigenvalues require estimating covariance weights as well as the subspace.

Numerical evidence and scope. Synthetic cTRE experiments reproduce the predicted phase transition and finite-sample decay. A data-driven two-stage ridge estimator (dTRE) and a learned linear transformer show the same qualitative distinction between aligned and off-support sources, with smoother changes in risk.

1.1 Related Work

ICL as statistical inference. Transformers can learn simple function classes in context and implement procedures for linear regression, regularization, gradient descent, and algorithm selection (Garg et al., 2022; Bai et al., 2023; Akyürek et al., 2023; von Oswald et al., 2023; Ahn et al., 2023; Mahankali et al., 2024; Zhang et al., 2024). A Bayesian interpretation treats ICL as inference over latent task variables (Xie et al., 2022; Lu & Yu, 2025). For matched linear-regression tasks, Wu et al. (2024) give task-complexity guarantees for learning a nearly Bayes-optimal ridge predictor with linear attention; Lu et al. (2025) derive asymptotic learning-curve formulas.

Task diversity and transfer. The pretraining distribution affects the emergence of ICL (Chan et al., 2022), the transition from Bayesian to ridge-like regression behavior (Raventós et al., 2023), and generalization beyond the training distribution across low-dimensional task subspaces (Kwon et al., 2026; Goddard et al., 2025). Shared representations also improve few-shot transfer in multi-task and meta-learning when downstream tasks align with the learned representation (Tripuraneni et al., 2021; Du et al., 2021). Closest to our question, Letey et al. (2026) analyze how covariance mismatch and task diversity affect asymptotic prediction risk. We derive an exact finite-dimensional recovery threshold and finite-sample coefficient-risk bounds for cTRE, showing how source variance can displace target directions.

Inverse problems and learned priors. Rank-deficient or ill-conditioned inverse problems require regularization (Hoerl & Kennard, 1970; Engl et al., 1996; Hansen, 1998). Lu & Yu (2025) study how transformers learn a task prior and its regularizer in inverse linear regression. Related attention models recover structure in inverse problems and physics discovery (Yu et al., 2024; Cole et al., 2024). Low-dimensional structure also underlies ICL theory for low-dimensional and nonparametric target

classes (Oko et al., 2024; Kim et al., 2024; Ching et al., 2026). We focus on how additional source tasks change the prior subspace used for target inference.

2 Setup and Bayesian Formulation

We study inverse linear regression. In forward linear regression, a learner uses a context of input-output pairs $(x_i, y_i)_{i=1}^n$ from $y_i = x_i^\top w + \varepsilon_i$ to predict a response y_{n+1} at a new query input x_{n+1} (see e.g., Lu et al. (2025); Kwon et al. (2026); Letey et al. (2026)). Here the object of interest is the latent coefficient (task vector) $w \in \mathbb{R}^d$ itself. For each target task, the context $(X, Y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ satisfies $Y = Xw + \varepsilon$, where $X = [x_1, \dots, x_n]^\top$ and $Y = [y_1, \dots, y_n]^\top$. We assume $d > n + 1$ and $\sigma > 0$. The M_0 target pretraining tasks are independent, with coefficients, design rows, and noise drawn independently as follows:

$$y_i^j = x_i^{j,\top} w^j + \varepsilon_i^j, \quad x_i^j \sim \mathcal{N}(0, I_d), \quad w^j \sim \mathcal{N}(0, \Sigma_0), \quad \varepsilon_i^j \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

Writing $Y^j = X^j w^j + \varepsilon^j$, the target pretraining dataset is $\mathcal{D}_0 = \{(X^j, Y^j)\}_{j=1}^{M_0}$. The goal is to learn an inverse map $(X, Y) \mapsto \hat{w}(X, Y)$ that estimates the coefficient of a fresh target task. Since $n < d$, the context leaves some coefficient directions undetermined without prior information. Pretraining supplies information about which directions occur in the task family, and this information provides regularization. The inverse viewpoint is natural for recovering signals or operator kernels (Engl et al., 1996; Yu et al., 2024). Related in-context operator-learning work studies PDE solution maps and optimal-transport maps (Cole et al., 2024; 2026).

To study the effect of additional pretraining tasks, we add an independent source family with the same design and noise distributions:

$$\mathcal{D}_s = \{(X^j, Y^j)\}_{j=M_0+1}^{M_0+M_s}, \quad w^j \sim \mathcal{N}(0, \Sigma_s), j = M_0 + 1, \dots, M_0 + M_s, \tag{3}$$

where Σ_s is the new source task covariance. Write $\mathcal{D}_\nu = \mathcal{D}_0 \cup \mathcal{D}_s$ for the enlarged pretraining collection. Then the total pretraining sample size is $M = M_0 + M_s$ and the mixing fraction is $\nu = M_s/M \in [0, 1]$. At the population level this induces the mixed task-prior covariance

$$\bar{\Sigma}_\nu = (1 - \nu)\Sigma_0 + \nu\Sigma_s. \tag{4}$$

The learner pools these tasks without using their family labels. We keep M and ν separate: varying ν at fixed M changes task composition, whereas adding sources to a fixed target dataset increases both M and ν. The central question is whether the pooled data help or hinder in-context learning on the original target family.

The following assumption fixes the variance on the target subspace and describes the source variance on and off it.

Assumption 1 (Target and block-diagonal source covariances)**.** *The target and additional source task covariances satisfy the following conditions.*

(i) Target covariance. *There exists an orthonormal matrix $U_\star \in \mathbb{R}^{d\times r_w}$ with $1 \le r_w < n$ such that*

$$\Sigma_0 = \lambda\Pi_\star = \lambda U_\star U_\star^\top,$$

for some $\lambda > 0$. Thus $\mathrm{range}(U_\star) = \mathrm{range}(\Sigma_0)$ is the rank-r_w target subspace.

(ii) Additional source covariance. *Let $U_\perp \in \mathbb{R}^{d\times(d-r_w)}$ be any orthonormal basis for $\mathrm{range}(U_\star)^\perp$. The source covariance satisfies $U_\star^\top \Sigma_s U_\perp = 0$, or equivalently,*

$$\Sigma_s = U_\star B U_\star^\top + U_\perp C U_\perp^\top \tag{5}$$

for some nonnegative definite matrices $B \in \mathbb{R}^{r_w\times r_w}$ and $C \in \mathbb{R}^{(d-r_w)\times(d-r_w)}$.

The matrices B and C describe source variance along the target subspace and its complement. Removing covariance between these components allows a direct eigenvalue comparison. Appendix A.3 permits nonzero covariance between them and bounds the change in the selected subspace, using this exactly solvable model as a reference.

2.1 BAYESIAN ANALYSIS: ESTIMATING THE TARGET SUBSPACE

We first identify the optimal inverse map when the target prior is known, and then determine what must be learned from pretraining. Under squared coefficient loss, the Bayes-optimal estimator is the target posterior mean. Since $\Sigma_0 = \lambda\Pi_\star$, it takes the form

$$\begin{aligned} w_\star(X,Y) = \mathbb{E}[w \mid X,Y] &= \Sigma_0 X^\top (X\Sigma_0 X^\top + \sigma^2 I_n)^{-1} Y \\ &= \Pi_\star X^\top (X\Pi_\star X^\top + \rho I_n)^{-1} Y, \qquad \rho = \sigma^2/\lambda. \end{aligned} \tag{6}$$

Appendix A gives the derivation. This is ridge regression restricted to the target subspace. With the rank r_w and ridge ratio ρ known, learning the target Bayes rule therefore reduces to estimating the projector $\Pi_\star$; equivalently, the target subspace $\mathrm{range}(U_\star)$.

For any estimator $\hat{w}(X,Y)$, we measure its performance by the *target excess Bayes risk*

$$\mathcal{E}_0(\hat{w}) := \mathbb{E}\big[\|\hat{w}(X,Y) - w_\star(X,Y)\|_2^2\big], \tag{7}$$

where the expectation is over a fresh target task (X, w, ε), with $Y = Xw + \varepsilon$, conditional on the pretraining data. This risk measures the error of the learned inverse map relative to the target posterior mean. The total coefficient error also includes the irreducible Bayes risk:

$$\mathbb{E}\big[\|\hat{w}(X,Y) - w\|_2^2\big] = \mathcal{E}_0(\hat{w}) + \mathbb{E}\big[\|w_\star(X,Y) - w\|_2^2\big].$$

The identity follows from the orthogonality property of conditional expectation. For a query $x_{\mathrm{q}} \sim \mathcal{N}(0, I_d)$ independent of the context, task coefficient, and pretraining data, the total forward prediction error equals the total coefficient error:

$$\mathbb{E}\big[\big(x_{\mathrm{q}}^\top \hat{w}(X,Y) - x_{\mathrm{q}}^\top w\big)^2\big] = \mathbb{E}\left[\|\hat{w}(X,Y) - w\|_2^2\right].$$

Estimating the target subspace. If the pretraining task vectors w^j were observed, we could estimate their covariance directly. Instead, each context constrains its task vector through an underdetermined system. We use the minimum-norm least-squares estimate (mLSE)

$$\widehat{w}^j(X^j, Y^j) := (X^{j,\top} X^j)^\dagger X^{j,\top} Y^j, \qquad j = 1, \ldots, M, \tag{8}$$

where $\dagger$ denotes the Moore–Penrose pseudoinverse. Although an individual estimate need not recover its task vector, the covariance across these independent, centered estimates retains the task-prior eigenspaces. In fact, their average covariance is

$$\Sigma_{\hat{w},\nu} := \frac{1}{M}\sum_{j=1}^{M} \mathrm{Cov}(\widehat{w}^j) = a\bar{\Sigma}_\nu + \big(b\,\mathrm{tr}(\bar{\Sigma}_\nu) + c\sigma^2\big) I_d. \tag{9}$$

Lemma 7 and Corollary 8 give the constants $a, b, c > 0$ and prove the identity.

Population rule from the pooled subspace. To isolate the effect of the source distribution, suppose first that $\Sigma_{\hat{w},\nu}$ is known. Let Π_ν project onto any leading rank-r_w eigenspace of $\Sigma_{\hat{w},\nu}$. Substituting $\lambda\Pi_\nu$ for the target covariance gives a Gaussian posterior mean.

Definition 2 (Population Gaussian posterior mean)**.** *Let Π_ν be a leading rank-r_w spectral projector of $\Sigma_{\hat{w},\nu}$ in* (9)*, and let $\rho = \sigma^2/\lambda$. The posterior mean under the approximate prior $\mathcal{N}(0, \lambda\Pi_\nu)$ is*

$$\hat{w}_\nu^{\mathrm{PM}}(X,Y) = \Pi_\nu X^\top (X\Pi_\nu X^\top + \rho I_n)^{-1} Y. \tag{10}$$

At $\nu = 0$, the covariance identity gives $\Pi_0 = \Pi_\star$, so $\hat{w}_0^{\mathrm{PM}} = w_\star$.

2.2 POSITIVE AND NEGATIVE TRANSFER

Informative eigengap of $\Sigma_{\hat{w},\nu}$. The population posterior mean in Definition 2 is determined by the rank-r_w projector Π_ν extracted from $\Sigma_{\hat{w},\nu}$. It recovers the target Bayes estimator precisely when its selected projector is $\Pi_\star$. By (9), this occurs uniquely when every eigenvalue on the target subspace exceeds every eigenvalue off it. We now express that comparison in terms of the eigengap.

Write the eigenvalues of the on-target source block B as $\beta_1 \le \cdots \le \beta_{r_w}$ and the eigenvalues of the off-target source block C as $\gamma_1 \ge \cdots \ge \gamma_{d-r_w} \ge 0$. Define

$$\beta_{\mathrm{on}} := \beta_1, \qquad \beta_{\mathrm{off}} := \gamma_1 = \|C\|_{\mathrm{op}}, \tag{11}$$

Here β_{on} is the weakest source component on the target subspace and β_{off} is the strongest source component off the target subspace.

We define the *informative eigengap* by subtracting the largest eigenvalue on the orthogonal complement from the smallest eigenvalue on the target subspace:

$$\begin{aligned} c_{\text{gap},\nu} :=& \lambda_{\min,\text{range}(U_\star)}(\Sigma_{\hat{w},\nu}) - \lambda_{\max,\text{range}(U_\perp)}(\Sigma_{\hat{w},\nu}) \\ =& a\lambda + \nu a(\beta_{\text{on}} - \lambda) - \nu a\,\beta_{\text{off}}, \end{aligned} \tag{12}$$

Here $\lambda_{\min,\text{range}(U_\star)}$ and $\lambda_{\max,\text{range}(U_\perp)}$ denote eigenvalues of the restrictions to the indicated invariant subspaces. Proposition 9 proves the second equality, with $a = \frac{n(dn+d-2)}{d(d-1)(d+2)}$. A positive gap uniquely selects $\Pi_\star$. At zero, the weakest target direction ties the strongest off-target direction; below zero, every leading rank-r_w subspace differs from the target. Unlike a gap between consecutive ordered eigenvalues, this informative gap can be negative.

When $\beta_{\text{off}} > \beta_{\text{on}}$, increasing the source fraction ν eventually makes the gap vanish. We define the corresponding *critical mixing fraction*, with an endpoint convention for the other cases, by

$$\nu_{\text{c}} := \begin{cases} \lambda/(\lambda + \beta_{\text{off}} - \beta_{\text{on}}), & \beta_{\text{off}} > \beta_{\text{on}}, \\ 1, & \beta_{\text{off}} \le \beta_{\text{on}}. \end{cases} \tag{13}$$

If $\beta_{\text{off}} > \beta_{\text{on}}$, the target projector remains a valid choice at $\nu = \nu_{\text{c}}$, but cannot be selected for $\nu > \nu_{\text{c}}$. If $\beta_{\text{on}} = \beta_{\text{off}}$, the gap is positive for $\nu < 1$ and vanishes at $\nu = 1$. If $\beta_{\text{off}} < \beta_{\text{on}}$, the gap remains positive on all of $[0, 1]$, so $\nu_{\text{c}} = 1$ is only a convention.

Transfer in terms of subspace recovery. The size of a positive gap determines how much covariance perturbation can be tolerated before the selected subspace changes substantially. We compare it with the target-only gap through $R_\nu = c_{\text{gap},\nu}/c_{\text{gap},0}$. This comparison defines the following transfer regimes.

Definition 3 (Transfer regimes for subspace recovery)**.** *Fix a source covariance Σ_s and a mixing fraction $\nu \in (0, 1]$. Recall $c_{\text{gap},\nu}$ in* (12) *and $c_{\text{gap},0} = a\lambda > 0$; let R_ν be the gap ratio defined in* (1)*.*

We say that broader pretraining at level ν induces:

- *(i)* Positive transfer *if $R_\nu \ge 1$: the eigengap does not decrease; equality is the neutral boundary;*
- *(ii)* Target-subspace-preserving negative transfer *if $0 < R_\nu < 1$: the eigengap shrinks, but the population top-r_w eigenspace is still $\Pi_\star$;*
- *(iii)* Loss of guaranteed target-subspace recovery *if $R_\nu \le 0$: at equality the leading subspace is nonunique, whereas $R_\nu < 0$ forces every leading rank-r_w subspace to differ from $\Pi_\star$.*

These definitions compare recovery margins. To compare finite-sample risks, one must also account for covariance fluctuations and the total number of tasks, as in Section 4.

3 Population recovery of the target subspace

We first state how the gap determines target excess risk in (7), then derive the covariance identity that gives its explicit form. A positive gap recovers the target Bayes rule. A negative gap forces the estimator to omit part of the target subspace, which creates a positive error even when the pooled covariance is known exactly. Proofs are in Appendix A.

Theorem 4 (Population posterior-mean recovery and target loss)**.** *Consider the Gaussian model* (2) *with $d > n + 1$, $\sigma > 0$, and Assumption 1. For $\nu \in [0, 1]$, let Π_ν be any leading rank-r_w spectral projector of $\Sigma_{\hat{w},\nu}$, and let $\hat{w}_\nu^{\text{PM}}$ be the Gaussian posterior mean in Definition 2. With R_ν as in* (1)*, the following hold.*

- *(a)* Exact target-recovery. *If $R_\nu > 0$ (in particular, for every $\nu < \nu_{\text{c}}$), then $\Pi_\nu = \Pi_\star$, $\hat{w}_\nu^{\text{PM}} = w_\star$, and $\mathcal{E}_0\big(\hat{w}_\nu^{\text{PM}}\big) = 0$.*
- *(b)* Loss of target directions. *If $R_\nu < 0$ (equivalently, $\nu > \nu_{\text{c}}$ when $\beta_{\text{off}} > \beta_{\text{on}}$), then every such projector satisfies*

$$\Pi_\nu \ne \Pi_\star, \qquad \mathcal{E}_0\big(\hat{w}_\nu^{\text{PM}}\big) \ge \mathbb{E}\big[\|(I_d - \Pi_\nu)w_\star(X, Y)\|_2^2\big] > 0.$$

At $R_\nu = 0$, the leading rank-r_w eigenspace is nonunique.

Remark 5 (Loss of all target directions). *If $d - r_w \geq r_w$ and $\nu\gamma_{r_w} > (1-\nu)\lambda + \nu\beta_{\max}$, then at least r_w off-target directions exceed every target direction. Here γ_{r_w} is the r_w-th largest eigenvalue of C and $\beta_{\max} = \lambda_{\max}(B)$. Every leading rank-$r_w$ subspace then lies in* $\mathrm{range}(U_\perp)$*, giving $\Pi_\nu \Pi_\star = 0$ and $\mathcal{E}_0(\hat{w}_\nu^{\mathrm{PM}}) \geq \mathbb{E}\|w_\star\|_2^2$.*

Remark 6 (Target-subspace recovery is weaker than positive transfer). *If $\beta_{\mathrm{off}} \leq \beta_{\mathrm{on}}$, target recovery holds for every $\nu < 1$. A nondecreasing margin requires the stronger condition $\beta_{\mathrm{on}} \geq \lambda + \beta_{\mathrm{off}}$. Thus, when $\beta_{\mathrm{off}} \leq \beta_{\mathrm{on}} < \lambda + \beta_{\mathrm{off}}$, broader pretraining preserves the target subspace but reduces the gap for $0 < \nu < 1$.*

3.1 Covariance of mLSE and spectral gap decomposition

The central calculation is the covariance of the mLSE. Its signal part is the projection of the task vector onto the random row space of the design. Averaging over the isotropic Gaussian design preserves the prior covariance up to a positive factor and adds an isotropic term. The noise contributes another isotropic term.

Lemma 7 (Covariance of mLSE). *Assume $d > n + 1$ and the independent Gaussian design and noise of* (2)*. Suppose that the min-norm least squares estimator $\widehat{w}^j(X^j, Y^j)$ in* (8) *has w^j drawn from a centered Gaussian distribution with covariance Σ. Then the covariance of $\widehat{w}^j(X^j, Y^j)$ has the closed form*

$$\mathrm{Cov}(\widehat{w}) = a\,\Sigma + \big(b\ \mathrm{tr}(\Sigma) + c\,\sigma^2\big) I_d, \tag{14}$$

where $a = \frac{n(dn+d-2)}{d(d-1)(d+2)}$, $b = \frac{n(d-n)}{d(d-1)(d+2)}$, and $c = \frac{n}{d(d-n-1)}$.

Corollary 8 (Covariance of mLSE under broader pretraining). *For the pooled target and source tasks with fraction ν and $\bar{\Sigma}_\nu = (1-\nu)\Sigma_0 + \nu\Sigma_s$, the average mLSE covariance defined in* (9) *is*

$$\Sigma_{\hat{w},\nu} = a\,\bar{\Sigma}_\nu + \big(b\ \mathrm{tr}(\bar{\Sigma}_\nu) + c\,\sigma^2\big) I_d. \tag{15}$$

In particular, $\Sigma_{\hat{w},\nu}$ shares its eigenvectors with $\bar{\Sigma}_\nu$.

The isotropic term shifts every eigenvalue by the same amount. It therefore cancels when we compute the eigengap. Under Assumption 1, the smallest target eigenvalue of $\bar{\Sigma}_\nu$ is $(1-\nu)\lambda + \nu\beta_{\mathrm{on}}$, and the largest off-target eigenvalue is $\nu\beta_{\mathrm{off}}$. Their difference, multiplied by a, gives the informative gap.

Proposition 9 (Population spectral-gap decomposition). *Under Assumption 1, the eigengap in* (12) *is*

$$c_{\mathrm{gap},\nu} = \underbrace{a\lambda}_{c_{\mathrm{gap},0}} + \underbrace{\nu a(\beta_{\mathrm{on}} - \lambda)}_{\textit{change on target}} - \underbrace{\nu a\,\beta_{\mathrm{off}}}_{\textit{off-target variance}}, \tag{16}$$

where β_{on} and β_{off} are defined in (11)*. In particular, $c_{\mathrm{gap},\nu} > 0$ for every $\nu < \nu_{\mathrm{c}}$. When $\beta_{\mathrm{off}} > \beta_{\mathrm{on}}$, its sign changes at the interior threshold ν_{c} in* (13)*; when $\beta_{\mathrm{off}} \leq \beta_{\mathrm{on}}$, it is non-negative on* $[0, 1]$.

The term $\nu a(\beta_{\mathrm{on}} - \lambda)$ can be positive or negative, depending on whether the weakest on-target source variance exceeds the target variance. The off-target contribution $-\nu a\beta_{\mathrm{off}}$ is always nonpositive. Dividing by $a\lambda$ gives (1) and the transfer regimes in Definition 3.

4 Finite-sample subspace estimation and target risk

The population analysis assumes that the covariance of the taskwise estimates is known. We now replace it with an empirical covariance. The controlled two-stage ridge estimator (cTRE) first estimates the target subspace and then performs ridge regression within that subspace. It uses the true rank and ridge ratio, so its excess risk is determined by the difference between the learned and target subspaces. The data-driven version (dTRE) also estimates the rank and regularization parameters; see Appendix B.2 for its construction and Lu & Yu (2025) for background.

Definition 10 (Controlled two-stage ridge estimator (cTRE)). *Let $\mathcal{D}_\nu = \{(X^j, Y^j)\}_{j=1}^M$, $M \geq 2$, be a pretraining dataset with mixing fraction ν. In Stage 1, the mLSE* (8) *gives the empirical covariance*

$$\widehat{C}_{M,\nu} = \frac{1}{M-1} \sum_{j=1}^{M} (\widehat{w}^j - \bar{w})(\widehat{w}^j - \bar{w})^\top, \qquad \bar{w} = \frac{1}{M} \sum_{j=1}^{M} \widehat{w}^j. \tag{17}$$

Let $\widehat{U}_{r_w} \in \mathbb{R}^{d\times r_w}$ collect the top-r_w eigenvectors of $\widehat{C}_{M,\nu}$. In Stage 2, on a fresh target-task context (X, Y), cTRE performs ridge regression inside $\widehat{U}_{r_w}$ with the oracle ridge ratio $\rho = \sigma^2/\lambda$:

$$\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}(X, Y) = \widehat{U}_{r_w}\big(\widehat{U}^\top_{r_w} X^\top X\, \widehat{U}_{r_w} + \rho I_{r_w}\big)^{-1}\widehat{U}^\top_{r_w} X^\top Y. \tag{18}$$

The next theorem establishes convergence and a phase transition for cTRE. Its proof in Appendix A.4 has three steps: control the empirical covariance error, use the population gap to bound the subspace error, and propagate this error through the ridge map. For a nonzero positive semidefinite matrix A, define its effective rank by

$$r_{\mathrm{eff}}(A) := \frac{\mathrm{tr}(A)}{\|A\|_{\mathrm{op}}}.$$

For almost-sure limits, use initial segments of independent target and source task sequences with $M_s(M)/M \to \nu$; finite-sample bounds use the empirical fraction, as detailed in Appendix A.6.

Theorem 11 (Convergence and phase transition for cTRE)**.** *Under Assumption 1 and the independent Gaussian design and noise in* (2)*, assume $d > n + 3$ and $\sigma > 0$. Let $\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}$ be the estimator in Definition 10, with $M \geq 2$. There are constants C_{cov} and C_{fin}, depending only on $(d, n, r_w, \lambda, \sigma^2)$ and Σ_s, such that the following hold. For $\delta \in (0, 1)$, write*

$$\eta_{M,\nu}(\delta) := C_{\mathrm{cov}}\delta^{-1/2}\|\Sigma_{\hat{w},\nu}\|_{\mathrm{op}}\sqrt{\frac{r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu})}{M}}.$$

(a) Matched aligned sources. *If $\Sigma_s = (\beta_{\mathrm{on}}/\lambda)\Sigma_0$ with $\beta_{\mathrm{on}}/\lambda \geq 1$, $c_{\mathrm{gap},\nu} = a\lambda(1 + \nu(\beta_{\mathrm{on}}/\lambda - 1))$. For every $\nu \in [0, 1]$, if $\eta_{M,\nu}(\delta) \leq c_{\mathrm{gap},\nu}/4$, then with probability at least $1 - \delta$ over pretraining,*

$$\mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big) \leq C_{\mathrm{fin}}\delta^{-1} r_w\, \frac{\|\Sigma_{\hat{w},\nu}\|^2_{\mathrm{op}} r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu})}{M c^2_{\mathrm{gap},\nu}} \tag{19}$$

Consequently, for every fixed $\nu \in [0, 1]$, $\mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big) \xrightarrow[M\to\infty]{\text{a.s.}} 0$.

(b) Pure off-support sources. *If $\Sigma_s = \beta_{\mathrm{off}}\Pi_\perp$ with $\Pi_\perp = I_d - \Pi_\star$, then $\nu_{\mathrm{c}} = \lambda/(\lambda + \beta_{\mathrm{off}})$ and $c_{\mathrm{gap},\nu} = a((1-\nu)\lambda - \nu\beta_{\mathrm{off}})$. For every $\nu < \nu_{\mathrm{c}}$, whenever $\eta_{M,\nu}(\delta) \leq c_{\mathrm{gap},\nu}/4$, with probability at least $1 - \delta$,* (19) *holds, and hence $\mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big) \xrightarrow[M\to\infty]{\text{a.s.}} 0$.*
If, in addition, $d - r_w \geq r_w$, then for every fixed $\nu > \nu_{\mathrm{c}}$, target consistency fails:

$$\liminf_{M\to\infty} \mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big) \geq \mathbb{E}\|w_\star(X, Y)\|_2^2 > 0 \qquad \text{a.s.} \tag{20}$$

For purely off-support sources, subspace error vanishes below the threshold as the number of tasks grows; above it, the estimated subspace omits the target directions. The lower bound in (20) measures the omitted component, while the fitted off-target component can contribute additional error. For positive gaps, (19) also explains why a larger gap alone does not imply lower risk: the covariance norm and effective rank appear in the numerator. Appendix A.6 extends the result to general block-diagonal sources and to sources with covariance between the target subspace and its complement.

5 Experiments

We first use cTRE to test the population threshold and finite-sample convergence, then ask whether the aligned/off-support distinction is also visible for data-driven estimators, including dTRE and a trained linear transformer. Knowing the target rank and ridge ratio lets cTRE isolate the effect of subspace estimation. Appendix B.2 gives the corresponding dTRE and cTRE implementation details. The linear-transformer model is a fully data-driven estimator whose performance depends on optimization and architecture. Here we use the implementation from Yu et al. (2024); see Appendix B.3 for details.

For cTRE, we use $(d, n, r_w, \lambda, \sigma^2) = (100, 20, 5, 1, 0.04)$ and estimate the excess Bayes risk by Monte Carlo on fresh target tasks:

$$\widehat{\mathcal{E}_0} = \frac{1}{N_{\mathrm{test}}}\sum_{i=1}^{N_{\mathrm{test}}} \big\|\hat{w}(X_i, Y_i) - w_\star(X_i, Y_i)\big\|_2^2,$$

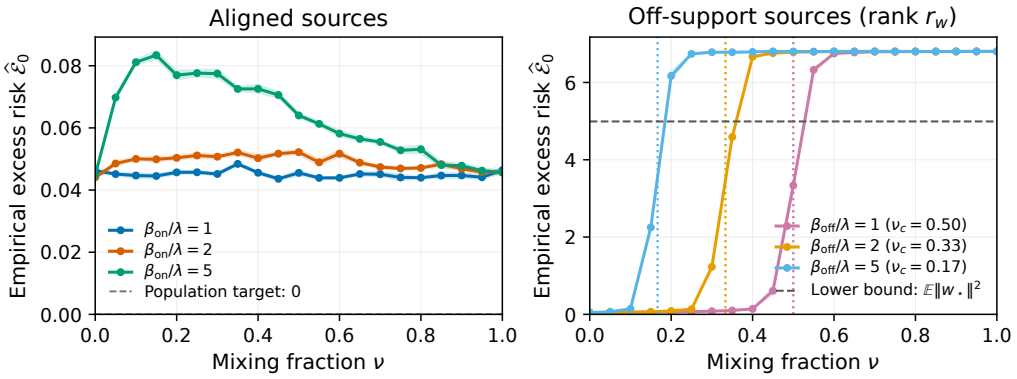

Figure 1: cTRE excess risk versus ν at $M = 5000$. Aligned-source risk remains near zero; off-support risk rises near ν_c and approaches a plateau above the lower bound $\mathbb{E}\|w_\star\|_2^2$.

where $w_\star$ is the target posterior mean in (6). We use 10 independent seeds and $N_{\text{test}} = 2000$ fresh target tasks per seed, reusing each test set across ν and M. Curves in E1–E2 show seed means with one-standard-error bands. Appendix B gives the data generation and estimator details.

E1: Population recovery threshold. Figure 1 uses $M = 5000$ to approximate the population behavior of cTRE. We compare two source families:

(a) *Aligned:* $\Sigma_s = \beta_{\text{on}}\Pi_\star = (\beta_{\text{on}}/\lambda)\Sigma_0$ with $\beta_{\text{on}}/\lambda \in \{1, 2, 5\}$ (so $\beta_{\text{off}} = 0$, $\nu_c = 1$).
(b) *Off-support:* $\Sigma_s = \beta_{\text{off}}\Pi_s$, where Π_s has rank r_w and $\Pi_s\Pi_\star = 0$ with $\beta_{\text{off}}/\lambda \in \{1, 2, 5\}$ (so $\beta_{\text{on}} = 0$ and $\nu_c \in \{0.50, 0.33, 0.17\}$).

The observed transition agrees with Theorem 4. Aligned-source risk remains small, although stronger sources produce a slight increase at intermediate ν: covariance fluctuations also change, so a larger eigengap alone need not reduce finite-sample risk. Off-support risk remains positive above ν_c. Its plateau exceeds $\mathbb{E}\|w_\star\|_2^2$, since ridge regression in the wrong subspace produces a nonzero estimate.

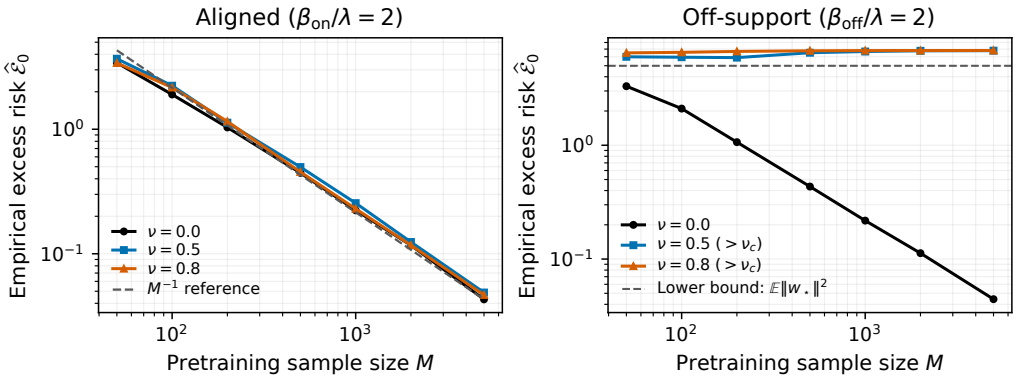

Figure 2: cTRE convergence with increasing M: $1/M$-scale decay for aligned sources; decay below ν_c and a positive plateau above it for off-support sources, consistent with Corollary 25.

E2: Phase transition in cTRE convergence. Figure 2 varies $M \in \{50, 100, 200, 500, 1000, 2000, 5000\}$ for the E1 families with source strength 2λ and $\nu \in \{0, 0.5, 0.8\}$. All aligned curves show $1/M$-scale decay, agreeing with Theorem 11(a). Off support, only $\nu = 0$ lies below $\nu_c = 1/3$ and shows decay; $\nu = 0.5$ and 0.8 give positive plateaus, so the excess risk relative to the target posterior mean does not vanish. Thus more pretraining tasks improve covariance estimation without correcting the population subspace error above the threshold.

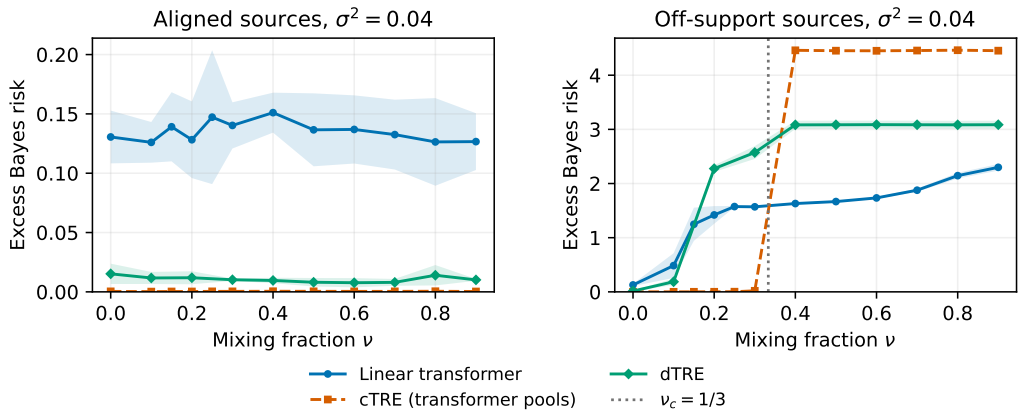

Figure 3: Excess Bayes risk for dTRE and the trained linear transformer at $\sigma^2 = 0.04$, with cTRE as a reference. Curves show seed means and bands show one sample standard deviation.

E3: Data-driven estimators. Figure 3 compares dTRE and a linear transformer with cTRE at $(d, n, r_w) = (20, 10, 3)$. Here the aligned source is $\Sigma_s = 2\Sigma_0$, while the off-support source uses the full complement, $\Sigma_s = 2\lambda\Pi_\perp$. The dTRE curves average 10 runs; the transformer curves average 7–10 aligned and 9–12 off-support runs per cell, with sample-standard-deviation bands. In the aligned setting, dTRE is near the cTRE reference, while the transformer estimator has larger but still small excess risk for all ν. In the off-support setting, dTRE and the transformer estimator degrade before the threshold and do not show a crisp transition. Data-driven rank/ridge selection, architecture, and optimization may contribute, but the experiment does not isolate their effects; see Appendix B.2–B.3 for details and additional tests when $\sigma^2 = 0.0025$. We therefore view this experiment as qualitative evidence that the spectral diagnostic remains relevant for data-driven ICL models, not as a theorem-level validation for the estimators.

6 Conclusion and Limitations

Conclusion. Broader pretraining is not automatically beneficial for a fixed target family. For the pooled subspace estimator in ICL-ILR, added source tasks change the covariance used to infer the task prior, and an eigengap between on-target and off-target directions governs recovery. Aligned sources preserve the target subspace for $\nu < 1$, but widen the eigengap only when their weakest variance exceeds the target variance. Sufficiently strong off-target variance can reverse the gap, forcing positive target excess Bayes risk. Thus, broader pretraining must be evaluated relative to the target directions and the accuracy with which they can be estimated.

Limitations. We note that the current analysis is limited to the Gaussian linear settings. Nonlinear and non-Gaussian extensions require further assumptions and new techniques.

The convergence guarantees only apply to the cTRE, which assumes that the target rank and ridge ratio are known. In practice, these parameters must be estimated from data, which can introduce additional error. The data-driven estimators only achieve qualitative results of the phase transition. It remains to establish rigorous guarantees for rank selection, regularization, transformer training, and target-aware source selection.

The cTRE guarantees assume fixed dimensions, isotropic Gaussian designs, positive noise, and known rank and ridge ratio. Bounds for coupled source directions require a positive block-diagonal reference gap. Unequal target spectra require additional covariance calibration for Bayes consistency and remains future work.

AI USE STATEMENT

LLMs and AI tools were used to polish writing, find related work, keep track of references, draft paper paragraphs, and revise the paper. The authors also used AI tools to assist with proofs for mathematical claims and experimental code, while the authors verified all proofs and results. The authors are responsible for the final content, including all AI-assisted text, claims, and numerical experiments.

REPRODUCIBILITY STATEMENT

All assumptions and proofs are provided in the main text and Appendix A; experimental procedures and model settings are presented in detail in Appendix B.

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. doi: 10.52202/075280-1977. URL https://proceedings.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. doi: 10.52202/075280-2498. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b2e63e36c57e153b9015fece2352a9f9-Abstract-Conference.html.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891, 2022. doi: 10.52202/068431-1371. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html.

Michelle Ching, Ioana Popescu, Nico Smith, Tianyi Ma, William G. Underwood, and Richard J. Samworth. Efficient and minimax optimal in-context nonparametric regression with transformers. *arXiv preprint arXiv:2601.15014*, 2026. URL https://arxiv.org/abs/2601.15014.

Frank Cole, Yulong Lu, Riley O'Neill, and Tianhao Zhang. Provable in-context learning of linear systems and linear elliptic PDEs with transformers. *arXiv preprint arXiv:2409.12293*, 2024. URL https://arxiv.org/abs/2409.12293v1. Version 1, September 2024.

Frank Cole, Dixi Wang, Yineng Chen, Yulong Lu, and Rongjie Lai. In-context operator learning on the space of probability measures. *arXiv preprint arXiv:2601.09979*, 2026. URL https://arxiv.org/abs/2601.09979.

R. Dennis Cook and Liliana Forzani. On the mean and variance of the generalized inverse of a singular Wishart matrix. *Electronic Journal of Statistics*, 5:146–158, 2011. doi: 10.1214/11-EJS602. URL https://doi.org/10.1214/11-EJS602.

Simon S. Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

Heinz W Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*. Kluwer Academic Publishers, Dordrecht, 1996. doi: 10.1007/978-94-009-1740-8. URL https://doi.org/10.1007/978-94-009-1740-8.

Shivam Garg, Dimitris Tsipras, Percy 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. doi: 10.52202/068431-2217. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Chase Goddard, Lindsay M. Smith, Vudtiwat Ngampruetikorn, and David J. Schwab. When can in-context learning generalize out of task distribution? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19585–19599. PMLR, 2025. URL https://proceedings.mlr.press/v267/goddard25a.html.

Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer, 2 edition, 2015. doi: 10.1007/978-3-319-13467-3. URL https://link.springer.com/book/10.1007/978-3-319-13467-3.

Per Christian Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998. doi: 10.1137/1.9780898719697. URL https://epubs.siam.org/doi/book/10.1137/1.9780898719697.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL https://doi.org/10.1080/00401706.1970.10488634.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106667–106713, 2024. doi: 10.52202/079017-3387. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c11daad0a48ea5f3c5c6390c7b060720-Abstract-Conference.html.

Soo Min Kwon, Alec S. Xu, Can Yaras, Laura Balzano, and Qing Qu. Out-of-distribution generalization of in-context learning: A low-dimensional subspace perspective. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4285–4293. PMLR, 2026. URL https://proceedings.mlr.press/v300/kwon26a.html.

Mary I. Letey, Jacob A. Zavatone-Veth, Yue M. Lu, and Cengiz Pehlevan. Pretrain–test task alignment governs generalization in in-context learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KZLeg0MQ2r.

Fei Lu and Yue Yu. Transformer learns the cross-task prior and regularization for in-context learning. *arXiv preprint arXiv:2505.12138*, 2025. URL https://arxiv.org/abs/2505.12138.

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. doi: 10.1073/pnas.2502599122. URL https://doi.org/10.1073/pnas.2502599122.

Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/87f476af4053961667c2c08e9f4b850e-Abstract-Conference.html.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77316–77365, 2024. doi: 10.52202/079017-2459. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8d6c356c82252032a20c749dda36f2d8-Abstract-Conference.html.

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, pp. 14228–14246, 2023. doi: 10.52202/075280-0626. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2e10b2c2e1aa4f8083c37dfe269873f8-Abstract-Conference.html.

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10434–10443, 2021. URL https://proceedings.mlr.press/v139/tripuraneni21a.html.

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

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, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

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://arxiv.org/abs/2310.08391.

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.

Yue Yu, Ning Liu, Fei Lu, Tian Gao, Siavash Jafarzadeh, and Stewart Silling. Nonlocal attention operator: Materializing hidden knowledge towards interpretable physics discovery. In *Advances in Neural Information Processing Systems*, volume 37, pp. 113797–113822, 2024. doi: 10.52202/079017-3613. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ce5b4f79f4752b7f8e983a80ebcd9c7a-Abstract-Conference.html.

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 https://www.jmlr.org/papers/v25/23-1042.html.

A PROOFS OF THE MAIN THEORY

Outline. We first compute the covariance of the per-task mLSE and use its eigenvalues to prove the population recovery threshold. We then study sources that mix target and off-target directions, and pass from population covariances to finite pretraining samples. Finally, we explain how unequal target eigenvalues change support recovery and the choice of ridge regularization.

The target posterior mean. Condition on X. Under the target model, $w \sim \mathcal{N}(0, \Sigma_0)$ and $Y = Xw + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ independent of w. Hence (w, Y) is jointly Gaussian with

$$\mathrm{Cov}(w, Y \mid X) = \Sigma_0 X^\top, \qquad \mathrm{Cov}(Y \mid X) = X\Sigma_0 X^\top + \sigma^2 I_n.$$

The conditional mean formula for jointly Gaussian vectors gives

$$\mathbb{E}[w \mid X, Y] = \Sigma_0 X^\top (X\Sigma_0 X^\top + \sigma^2 I_n)^{-1} Y.$$

With the equal-strength target prior $\Sigma_0 = \lambda \Pi_\star$, this becomes

$$\mathbb{E}[w \mid X, Y] = \lambda \Pi_\star X^\top (\lambda X \Pi_\star X^\top + \sigma^2 I_n)^{-1} Y = \Pi_\star X^\top (X \Pi_\star X^\top + \rho I_n)^{-1} Y, \qquad \rho = \sigma^2/\lambda.$$

A.1 PROOF OF LEMMA 7 AND COROLLARY 8

We write P_X for the orthogonal projector onto the row space of X and $O(d)$ for the orthogonal group:

$$P_X := X^\top (XX^\top)^\dagger X, \quad O(d) := \{Q \in \mathbb{R}^{d\times d} : Q^\top Q = I_d = QQ^\top\}.$$

The covariance calculation uses the rotational invariance of Gaussian design. We first record two elementary consequences, then apply the same idea to the signal covariance.

Lemma 12 (Orthogonal invariance forces a scaled identity)**.** *Let $M \in \mathbb{R}^{d\times d}$ be a symmetric matrix that is orthogonally invariant in the sense that $QM = MQ$ for every $Q \in O(d)$. Then M is a scalar multiple of the identity matrix. That is,*

$$M = Q^\top M Q \quad \textit{for every } Q \in O(d) \quad \Rightarrow \quad M = cI_d \textit{ for some scalar } c.$$

Proof. For each i, take $Q = D_i := I_d - 2e_i e_i^\top$, the diagonal matrix that flips the sign of the ith coordinate and fixes the others. From $M = D_i^\top M D_i = D_i M D_i$ we obtain

$$M_{ij} = -(M_{ij}) \qquad (j \neq i),$$

so $M_{ij} = 0$ whenever $i \neq j$.

Thus M is diagonal.

Next take $Q = P$, any permutation matrix. Since $M = P^\top M P$ and M is diagonal, the diagonal entries of M must all be equal. Therefore $M = cI_d$ for some scalar c. □

Lemma 13 (Rotational invariance of the mean pseudoinverse)**.** *Let $X \in \mathbb{R}^{n\times d}$ be a random matrix such that*

$$XQ \overset{d}{=} X \qquad \textit{for every deterministic } Q \in O(d), \textit{ i.e., } Q^\top Q = I_d = QQ^\top,$$

and suppose $\mathbb{E}\,\mathrm{tr}\left((X^\top X)^\dagger\right) < \infty$. Then we have

$$\mathbb{E}\big[(X^\top X)^\dagger\big] = cI_d \textit{ with} \quad c = \frac{1}{d}\mathbb{E}\,\mathrm{tr}\left((X^\top X)^\dagger\right).$$

Proof. Set $M := \mathbb{E}\big[(X^\top X)^\dagger\big]$. Fix $Q \in O(d)$. Since $XQ \overset{d}{=} X$,

$$\begin{aligned} M = \mathbb{E}\big[((XQ)^\top (XQ))^\dagger\big] &= \mathbb{E}\big[(Q^\top X^\top X Q)^\dagger\big] \\ &= \mathbb{E}\big[Q^\top (X^\top X)^\dagger Q\big] = Q^\top M Q, \end{aligned}$$

where the second-to-last equality uses the following general identity for the Moore–Penrose pseudoinverse: if A is a symmetric positive semidefinite matrix and Q is an orthogonal matrix, then by the spectral decomposition $A = U\Lambda U^\top$, we have

$$(Q^\top AQ)^\dagger = Q^\top U\Lambda^\dagger U^\top Q = Q^\top A^\dagger Q.$$

Applying this identity with $A = X^\top X$ gives

$$((XQ)^\top (XQ))^\dagger = Q^\top (X^\top X)^\dagger Q.$$

Hence $QM = MQ$ for every $Q \in O(d)$. By Lemma 12 above, M is a scalar multiple of the identity matrix. Taking traces yields $cd = \operatorname{tr}(M) = \mathbb{E}\operatorname{tr}\left((X^\top X)^\dagger\right)$, which proves the value for c. □

The next lemma describes linear covariance maps that respect changes of orthogonal coordinates. It is an elementary form of Schur's lemma; see, for instance, Hall (2015, Chapter 4). We give a direct proof to show why the coefficients depend only on the map.

Lemma 14 (Form of an orthogonally equivariant linear map)**.** *Let T be a linear map on the space of symmetric $d \times d$ matrices that is equivariant under orthogonal conjugation, i.e., $T(QAQ^\top) = QT(A)Q^\top$ for every $Q \in O(d)$ and every symmetric matrix $A \in \mathbb{R}^{d\times d}$. Then there exist scalars α, β, determined by T alone and independent of A, such that*

$$T(A) = \alpha A + \beta \operatorname{tr}(A) I_d$$

for every symmetric matrix A.

Proof. Set $E_{11} := e_1 e_1^\top$ and $B := T(E_{11})$. We first show that B acts as a scalar on $e_1^\perp$ and has the form

$$B = a_0 E_{11} + b_0(I_d - E_{11})$$

for some scalars a_0, b_0.

For each $k \in \{2, \dots, d\}$, let $D_k := I_d - 2e_k e_k^\top$, the diagonal matrix that flips the sign of the kth coordinate and fixes the others. Since $D_k^\top E_{11} D_k = E_{11}$, the equivariance property of T gives

$$B = T(E_{11}) = T(D_k^\top E_{11} D_k) = D_k^\top B D_k.$$

Comparing matrix entries,

$$B_{ij} = (D_k^\top B D_k)_{ij} = s_i s_j B_{ij},$$

where $s_k = -1$ and $s_\ell = 1$ for $\ell \neq k$. If exactly one of i, j equals k, then $s_i s_j = -1$, so $B_{ij} = -B_{ij}$ and therefore $B_{ij} = 0$. Taking $k = j$ shows $B_{1j} = 0$ for every $j \geq 2$, and taking $k = i$ shows $B_{ij} = 0$ for every distinct $i, j \geq 2$. Thus B is diagonal.

Now let P be any permutation matrix that fixes e_1 and permutes $e_2, \dots, e_d$. Since $P^\top E_{11} P = E_{11}$, equivariance again yields

$$B = T(E_{11}) = T(P^\top E_{11} P) = P^\top B P.$$

Because B is diagonal, its diagonal entries in positions $2, \dots, d$ must all be equal. Hence $B = a_0 E_{11} + b_0(I_d - E_{11})$.

Let $x \in \mathbb{R}^d$ be any unit vector, and choose $Q \in O(d)$ with $Qe_1 = x$. Then $xx^\top = QE_{11}Q^\top$, so

$$T(xx^\top) = QT(E_{11})Q^\top = (a_0 - b_0)xx^\top + b_0 I_d.$$

Thus the coefficients

$$\alpha := a_0 - b_0, \qquad \beta := b_0$$

are fixed once T is fixed; in particular, they do not depend on x.

Now note that every symmetric matrix A has a spectral decomposition $A = \sum_{m=1}^d \lambda_m u_m u_m^\top$ with eigenvalues $\lambda_1, \dots, \lambda_d$ and orthonormal eigenvectors $u_1, \dots, u_d$. By linearity of T,

$$\begin{aligned} T(A) &= \sum_{m=1}^d \lambda_m T(u_m u_m^\top) = \sum_{m=1}^d \lambda_m \left(\alpha u_m u_m^\top + \beta I_d\right) \\ &= \alpha \sum_{m=1}^d \lambda_m u_m u_m^\top + \beta \Big(\sum_{m=1}^d \lambda_m\Big) I_d = \alpha A + \beta \operatorname{tr}(A) I_d. \end{aligned}$$

The constants α, β depend only on T and not on A. □

Lemma 15 (Form of the signal covariance map)**.** *Assume that $X \in \mathbb{R}^{n\times d}$ has full row rank almost surely and $XQ \overset{d}{=} X$ for every deterministic $Q \in O(d)$, and let*

$$T(A) := \mathbb{E}[P_X A P_X], \quad P_X := X^\top (XX^\top)^{-1} X,$$

for any symmetric $d \times d$ matrix A. Then there exist scalars α, β, depending only on the law of X, such that

$$T(A) = \alpha A + \beta \operatorname{tr}(A) I_d$$

for every symmetric matrix A.

Proof. The map T is linear in A. The claim follows from Lemma 14 once we verify the equivariance property $T(QAQ^\top) = QT(A)Q^\top$ for every $Q \in O(d)$. The latter follows from the $O(d)$-invariance of the distribution of X and the definition of P_X as a function of X. Fix $Q \in O(d)$. Since $XQ \overset{d}{=} X$, we may compute with XQ instead of X. Also,

$$P_{XQ} = (XQ)^\top \big((XQ)(XQ)^\top\big)^{-1}(XQ) = Q^\top X^\top (XX^\top)^{-1} XQ = Q^\top P_X Q.$$

Therefore, for every symmetric A,

$$\begin{aligned} T(Q^\top AQ) &= \mathbb{E}[P_X Q^\top AQ P_X] = \mathbb{E}[P_{XQ} Q^\top AQ P_{XQ}] \\ &= \mathbb{E}[Q^\top P_X Q Q^\top AQ Q^\top P_X Q] = Q^\top \mathbb{E}[P_X A P_X] Q = Q^\top T(A) Q. \end{aligned}$$

Replacing Q by $Q^\top$ gives

$$T(QAQ^\top) = QT(A)Q^\top \qquad \text{for every } Q \in O(d).$$

Since T is determined by the law of X, the coefficients depend only on that law. □

Proof of Lemma 7. Decompose

$$\widehat{w} = (X^\top X)^\dagger X^\top Y = P_X w + X^\top (XX^\top)^{-1} \varepsilon,$$

which is well defined for $n < d$ on the event $\{XX^\top \text{ invertible}\}$, of probability one under Gaussian design. Since w and ε are centered, mutually independent, and independent of X,

$$\operatorname{Cov}(\widehat{w}) = \mathbb{E}\big[P_X \Sigma P_X\big] + \sigma^2\, \mathbb{E}\big[X^\top (XX^\top)^{-2} X\big] = \mathbb{E}\big[P_X \Sigma P_X\big] + \sigma^2\, \mathbb{E}\big[(X^\top X)^\dagger\big]. \tag{21}$$

Noise term. Since $XQ \overset{d}{=} X$ for every $Q \in O(d)$, Lemma 13 gives $\mathbb{E}[(X^\top X)^\dagger] = cI_d$ for some scalar c. Taking traces and using

$$\operatorname{tr}\big((X^\top X)^\dagger\big) = \operatorname{tr}\big((XX^\top)^{-1}\big),$$

together with the inverse-Wishart identity $\mathbb{E}[(XX^\top)^{-1}] = \frac{1}{d-n-1} I_n$ for $XX^\top \sim \operatorname{Wishart}_n(I_n, d)$ with $d > n+1$ gives

$$cd = \mathbb{E}\operatorname{tr}((XX^\top)^{-1}) = \frac{n}{d-n-1},$$

hence $c = \frac{n}{d(d-n-1)}$.

Signal term. The same right-orthogonal invariance of X implies, via Lemma 15, that

$$T(A) = \alpha A + \beta\ \operatorname{tr}(A)\, I_d \tag{22}$$

for some scalars α, β depending only on the law of X, which depends on (n, d).

To determine α, β, we use two test matrices. First, since $T(I_d) = \mathbb{E}[P_X]$, Lemma 12 gives $\mathbb{E}[P_X] = \gamma I_d$ for some scalar γ. Taking traces and using $\operatorname{tr}(P_X) = n$ almost surely,

$$\gamma d = \operatorname{tr}(\mathbb{E}[P_X]) = \mathbb{E}[\operatorname{tr}(P_X)] = n,$$

so $\gamma = n/d$. On the other hand,

$$T(I_d) = \alpha I_d + \beta \operatorname{tr}(I_d) I_d = (\alpha + d\beta) I_d.$$

Comparing the two expressions for $T(I_d)$ gives

$$\alpha + d\beta = \frac{n}{d}.$$

Second, set $A = E_{11} = e_1 e_1^\top$ and denote $P_{11} = e_1^\top P_X e_1$. Then

$$e_1^\top T(E_{11}) e_1 = e_1^\top \mathbb{E}[P_X e_1 e_1^\top P_X] e_1 = \mathbb{E}[P_{11}^2].$$

The row space of X is uniformly distributed over the n-dimensional subspaces. Thus P_{11} has distribution $\mathrm{Beta}(n/2, (d-n)/2)$, and its first two moments are $\mathbb{E}[P_{11}] = n/d$ and $\mathbb{E}[P_{11}^2] = n(n+2)/(d(d+2))$. Since $\mathrm{tr}(E_{11}) = 1$ and $e_1^\top E_{11} e_1 = 1$, Eq. (22) yields

$$\alpha + \beta = \frac{n(n+2)}{d(d+2)}.$$

Solving the two linear equations gives

$$\alpha = \frac{n(dn + d - 2)}{d(d-1)(d+2)} =: a, \qquad \beta = \frac{n(d-n)}{d(d-1)(d+2)} =: b.$$

Substituting $A = \Sigma$ into (22) and combining with the noise term in (21) gives the claimed identity (14). The eigenvector statement follows because $a\Sigma + \kappa I_d$ shares its eigenvectors with Σ for any scalar κ. □

Proof of Corollary 8. The mLSE has mean zero under each task distribution. Therefore the covariance of a uniformly selected task from the pooled collection is $(1-\nu)\,\mathrm{Cov}(\widehat{w} \mid \Sigma_0) + \nu\,\mathrm{Cov}(\widehat{w} \mid \Sigma_s)$. Applying Lemma 7 to both terms gives (15), because the covariance map is affine in Σ and $\bar{\Sigma}_\nu = (1-\nu)\Sigma_0 + \nu\Sigma_s$. □

A.2 Proof of the spectral gap decomposition and Theorem 4

Proof of Proposition 9. Substituting the source decomposition (5) (i.e., $\Sigma_s = U_\star B U_\star^\top + U_\perp C U_\perp^\top$) into Corollary 8, we obtain

$$\begin{aligned} \Sigma_{\hat{w},\nu} &= a\,\bar{\Sigma}_\nu + \big(b\ \mathrm{tr}(\bar{\Sigma}_\nu) + c\,\sigma^2\big) I_d \\ &= \tau_\nu I_d + a\big((1-\nu)\lambda \Pi_\star + \nu\, U_\star B U_\star^\top\big) + a\nu\, U_\perp C U_\perp^\top \end{aligned} \tag{23}$$

with isotropic constant $\tau_\nu := b\ \mathrm{tr}(\bar{\Sigma}_\nu) + c\sigma^2$. The target subspace and its orthogonal complement are invariant under $\Sigma_{\hat{w},\nu}$. Its eigenvalues on the two subspaces are

$$\begin{aligned} \text{on the target subspace } \mathrm{range}(U_\star): &\quad \tau_\nu + a\big((1-\nu)\lambda + \nu\beta_k\big), \quad k = 1, \dots, r_w, \\ \text{on its orthogonal complement } \mathrm{range}(U_\perp): &\quad \tau_\nu + a\nu\gamma_\ell, \quad \ell = 1, \dots, d - r_w, \end{aligned}$$

where $\{\beta_k\} = \sigma(B)$ and $\{\gamma_\ell\} = \sigma(C)$ (zeros included). The smallest on-support eigenvalue minus the largest off-support eigenvalue gives

$$c_{\mathrm{gap},\nu} = a\big((1-\nu)\lambda + \nu\beta_{\mathrm{on}}\big) - a\nu\beta_{\mathrm{off}} = a\lambda + \nu a(\beta_{\mathrm{on}} - \lambda) - \nu a\beta_{\mathrm{off}},$$

which is exactly (16). The sign follows by solving $c_{\mathrm{gap},\nu} \geq 0$: when $\beta_{\mathrm{off}} \leq \beta_{\mathrm{on}}$, the affine gap is nonnegative at both endpoints and hence throughout $[0,1]$, so $\nu_{\mathrm{c}} = 1$; when $\beta_{\mathrm{off}} > \beta_{\mathrm{on}}$, $c_{\mathrm{gap},\nu} \geq 0$ iff $\nu \leq \lambda/(\lambda + \beta_{\mathrm{off}} - \beta_{\mathrm{on}}) = \nu_{\mathrm{c}}$. □

Lemma 16 (Projector-mismatch lower bound)**.** *Let* Π *be any rank-*r_w *orthogonal projector, and let*

$$\hat{w}_\Pi(X, Y) := \Pi X^\top \big(X \Pi X^\top + \rho I_n\big)^{-1} Y, \qquad \rho = \sigma^2/\lambda.$$

Then

$$\mathcal{E}_0(\hat{w}_\Pi) \ \geq\ \mathbb{E}\big[\|(I_d - \Pi) w_\star(X, Y)\|_2^2\big].$$

If $\mathrm{range}(\Pi)$ *does not contain* $\mathrm{range}(\Pi_\star)$*, the right-hand side is strictly positive. If* $\Pi\Pi_\star = 0$*, then* $\mathcal{E}_0(\hat{w}_\Pi) \geq \mathbb{E}\|w_\star(X, Y)\|_2^2 > 0$*.*

Proof of Lemma 16. Fix (X, Y). By construction $\hat{w}_\Pi(X, Y) \in \mathrm{range}(\Pi)$, while $w_\star(X, Y) = \Sigma_0 X^\top (X\Sigma_0 X^\top + \sigma^2 I_n)^{-1} Y \in \mathrm{range}(\Sigma_0) = \mathrm{range}(\Pi_\star)$. Projecting the error onto $\mathrm{range}(\Pi)^\perp$ gives

$$(I_d - \Pi)(\hat{w}_\Pi - w_\star) = -(I_d - \Pi)w_\star,$$

so

$$\|\hat{w}_\Pi - w_\star\|_2^2 \geq \|(I_d - \Pi)w_\star\|_2^2.$$

Taking expectation gives the first claim. For strict positivity, set $Z = XU_\star$. Conditional on X, the target coordinates of the oracle have covariance

$$\mathbb{E}\big[U_\star^\top w_\star w_\star^\top U_\star \mid X\big] = \lambda^2 Z^\top (\lambda Z Z^\top + \sigma^2 I_n)^{-1} Z \succ 0,$$

because Z has full column rank almost surely. Thus $(I_d - \Pi)\Pi_\star \neq 0$ implies $\mathbb{E}\|(I_d - \Pi)w_\star\|_2^2 > 0$. If $\Pi\Pi_\star = 0$, then $(I_d - \Pi)w_\star = w_\star$ pointwise, giving the stated orthogonal lower bound. □

Proof of Theorem 4. Set $\rho = \sigma^2/\lambda$. By Corollary 8 and Proposition 9, the representation (23) gives the informative gap $c_{\mathrm{gap},\nu}$, with $c_{\mathrm{gap},0} = \lambda a$.

(a) Target recovery. If $R_\nu > 0$, then $c_{\mathrm{gap},\nu} > 0$, so the leading r_w-dimensional eigenspace of $\Sigma_{\hat{w},\nu}$ is $\mathrm{range}(\Pi_\star)$. Hence $\Pi_\nu = \Pi_\star$, and (10) gives

$$\hat{w}_\nu^{\mathrm{PM}}(X, Y) = \Pi_\star X^\top \big(X\Pi_\star X^\top + \rho I_n\big)^{-1} Y,$$

so $\hat{w}_\nu^{\mathrm{PM}} = w_\star$ pointwise and $\mathcal{E}_0(\hat{w}_\nu^{\mathrm{PM}}) = 0$.

(b) Loss of target directions. If $R_\nu < 0$, then the largest off-target eigenvalue is strictly larger than the smallest target eigenvalue. Therefore every leading rank-r_w projector Π_ν differs from $\Pi_\star$, including when a tie allows more than one choice of Π_ν. Since both projectors have rank r_w, $\mathrm{range}(\Pi_\nu)$ cannot contain $\mathrm{range}(\Pi_\star)$. Lemma 16 gives

$$\mathcal{E}_0(\hat{w}_\nu^{\mathrm{PM}}) \;\geq\; \mathbb{E}\big[\|(I_d - \Pi_\nu)w_\star\|_2^2\big] > 0.$$

The critical case. When $c_{\mathrm{gap},\nu} = 0$, the weakest target eigenvalue equals the strongest off-target eigenvalue. The leading r_w-dimensional eigenspace is then nonunique, and $\mathrm{range}(\Pi_\star)$ remains one valid choice. Thus a zero margin removes unique identification without forcing target loss. □

Justification of Remark 5. If $R_\nu < 0$, no leading rank-r_w subspace equals the target subspace, as shown above. Now suppose $d - r_w \geq r_w$ and $\nu\gamma_{r_w} > (1 - \nu)\lambda + \nu\beta_{\max}$. Then at least r_w off-target eigenvalues exceed every target eigenvalue. Every leading rank-r_w subspace therefore lies in $\mathrm{range}(U_\perp)$, so $\Pi_\nu\Pi_\star = 0$. Lemma 16 yields $\mathcal{E}_0(\hat{w}_\nu^{\mathrm{PM}}) \geq \mathbb{E}\|w_\star\|_2^2 > 0$. □

Remark 17 (The off-support plateau exceeds the lower bound)**.** *For a fixed rank-r_w projector $P = VV^\top$ orthogonal to $\Pi_\star$, let $Z = XV$ and $W = Z^\top Z$. Isotropic Gaussian design makes Z independent of Y, with $\mathbb{E}[YY^\top] = (\lambda r_w + \sigma^2)I_n$. The ridge output lies in $\mathrm{range}(P)$ and is orthogonal to $w_\star$ pointwise. Therefore*

$$\mathcal{E}_0(\hat{w}_P) = \mathbb{E}\|w_\star\|_2^2 + (\lambda r_w + \sigma^2)\mathbb{E}\,\mathrm{tr}\big[W(W + \rho I_{r_w})^{-2}\big] > \mathbb{E}\|w_\star\|_2^2.$$

The added term follows by conditioning on Z and taking the trace of the squared ridge map. It is finite for $\rho > 0$ and strictly positive because Z has full column rank almost surely. Thus the horizontal oracle-energy line in E1–E2 is a lower bound, not the exact population plateau.

A.3 Beyond block-diagonal source covariances

The block-diagonal source model preserves the target subspace and its orthogonal complement. We now allow nonzero covariance between the target subspace and its orthogonal complement, represented by a block D. Comparing with the covariance obtained by setting $D = 0$ shows how this coupling rotates the learned subspace.

Proposition 18 (Stability under non-block-diagonal source covariance)**.** *Consider the Gaussian ILR setup of Section 2, assume $d > n + 1$, and retain the target covariance*

$$\Sigma_0 = \lambda\Pi_\star = \lambda U_\star U_\star^\top, \qquad \lambda > 0,$$

from Assumption 1(i). Let $U_\perp \in \mathbb{R}^{d\times(d-r_w)}$ be an orthonormal basis for $\mathrm{range}(U_\star)^\perp$ and set $Q = [U_\star\ U_\perp] \in O(d)$. Suppose the source covariance is positive semidefinite and has the representation

$$\Sigma_s = Q\begin{bmatrix} B & D \\ D^\top & C \end{bmatrix} Q^\top \succeq 0, \qquad \|D\|_{\mathrm{op}} \le \epsilon, \tag{24}$$

where $B \in \mathbb{R}^{r_w\times r_w}$ and $C \in \mathbb{R}^{(d-r_w)\times(d-r_w)}$ are symmetric nonnegative definite matrices, and $D \in \mathbb{R}^{r_w\times(d-r_w)}$. Define the block-diagonal reference source and its mixed covariance by

$$\Sigma_s^{\mathrm{bd}} := Q\begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix} Q^\top, \qquad \bar\Sigma_\nu^{\mathrm{bd}} := (1-\nu)\Sigma_0 + \nu\Sigma_s^{\mathrm{bd}},$$

and let

$$\Sigma_{\hat w,\nu}^{\mathrm{bd}} := a\bar\Sigma_\nu^{\mathrm{bd}} + \big(b\ \mathrm{tr}(\bar\Sigma_\nu^{\mathrm{bd}}) + c\sigma^2\big) I_d$$

be the corresponding population mLSE covariance. Write the eigenvalues of a symmetric matrix in nonincreasing order and define the block-reference margins

$$h_\nu := (1-\nu)\lambda + \nu\lambda_{\min}(B) - \nu\lambda_{\max}(C), \qquad g_\nu := ah_\nu. \tag{25}$$

Fix $\nu \in [0,1]$ and suppose $h_\nu > 0$. Then $g_\nu = \lambda_{r_w}(\Sigma_{\hat w,\nu}^{\mathrm{bd}}) - \lambda_{r_w+1}(\Sigma_{\hat w,\nu}^{\mathrm{bd}})$. Let $\Sigma_{\hat w,\nu}$ be the population mLSE covariance induced by the actual source covariance (24)*, and let Π_ν be its leading rank-r_w spectral projector. Then:*

(a) The cross-block induces the exact covariance perturbation

$$\|\Sigma_{\hat w,\nu} - \Sigma_{\hat w,\nu}^{\mathrm{bd}}\|_{\mathrm{op}} = a\nu\|D\|_{\mathrm{op}} \le a\nu\epsilon. \tag{26}$$

(b) The leading spectral cluster remains separated without an additional smallness condition:

$$\lambda_{r_w}(\Sigma_{\hat w,\nu}) - \lambda_{r_w+1}(\Sigma_{\hat w,\nu}) \ge g_\nu > 0. \tag{27}$$

In particular, Π_ν is unique.

(c) The deviation from the target subspace satisfies

$$\begin{aligned}\|\Pi_\nu - \Pi_\star\|_{\mathrm{op}} &= \big\|\sin\Theta\big(\mathrm{range}(\Pi_\nu), \mathrm{range}(\Pi_\star)\big)\big\|_{\mathrm{op}} \\ &\le \frac{a\nu\|D\|_{\mathrm{op}}}{g_\nu} = \frac{\nu\|D\|_{\mathrm{op}}}{h_\nu} \le \frac{a\nu\epsilon}{g_\nu}.\end{aligned} \tag{28}$$

Moreover, $\Pi_\nu = \Pi_\star$ if and only if $\nu D = 0$.

Proof idea. Setting $D = 0$ leaves the diagonal blocks and trace unchanged, so the covariance perturbation has norm $a\nu\|D\|_{\mathrm{op}}$. The Courant–Fischer principle shows that this off-diagonal perturbation cannot close the gap between the leading r_w eigenvalues and the rest when $h_\nu > 0$. A Sylvester equation then bounds the rotation of the leading subspace.

Proof of Proposition 18. The source covariance and its block-diagonal reference have the same trace:

$$\mathrm{tr}(\Sigma_s) = \mathrm{tr}(B) + \mathrm{tr}(C) = \mathrm{tr}(\Sigma_s^{\mathrm{bd}}).$$

Consequently, $\mathrm{tr}(\bar\Sigma_\nu) = \mathrm{tr}(\bar\Sigma_\nu^{\mathrm{bd}})$. Corollary 8 therefore gives, in the orthogonal basis Q,

$$Q^\top\Sigma_{\hat w,\nu}Q = \begin{bmatrix} A_\nu & K_\nu \\ K_\nu^\top & F_\nu \end{bmatrix}, \qquad Q^\top\Sigma_{\hat w,\nu}^{\mathrm{bd}}Q = \begin{bmatrix} A_\nu & 0 \\ 0 & F_\nu \end{bmatrix}, \tag{29}$$

where

$$\begin{aligned} A_\nu &= \tau_\nu I_{r_w} + a\big((1-\nu)\lambda I_{r_w} + \nu B\big), \\ F_\nu &= \tau_\nu I_{d-r_w} + a\nu C, \\ K_\nu &= a\nu D, \qquad \tau_\nu := b\ \mathrm{tr}(\bar\Sigma_\nu) + c\sigma^2. \end{aligned}$$

The square of the symmetric off-diagonal perturbation is

$$\begin{bmatrix} 0 & K_\nu \\ K_\nu^\top & 0 \end{bmatrix}^2 = \begin{bmatrix} K_\nu K_\nu^\top & 0 \\ 0 & K_\nu^\top K_\nu \end{bmatrix}.$$

Hence its operator norm is $\|K_\nu\|_{\mathrm{op}} = a\nu\|D\|_{\mathrm{op}}$. Orthogonal conjugation preserves the operator norm, proving (26).

Set

$$\alpha_\nu := \lambda_{\min}(A_\nu), \qquad \beta_\nu := \lambda_{\max}(F_\nu).$$

The definitions in (29) give

$$\alpha_\nu - \beta_\nu = a\big((1-\nu)\lambda + \nu\lambda_{\min}(B) - \nu\lambda_{\max}(C)\big) = g_\nu > 0.$$

Let $H_\nu := Q^\top \Sigma_{\hat{w},\nu} Q$. By the Courant–Fischer variational principle, using the target coordinate subspace $\mathbb{R}^{r_w} \times \{0\}$ as a trial subspace,

$$\lambda_{r_w}(H_\nu) = \max_{\substack{\mathcal{L}\subset\mathbb{R}^d \\ \dim(\mathcal{L})=r_w}} \min_{\substack{x\in\mathcal{L} \\ \|x\|_2=1}} x^\top H_\nu x \geq \alpha_\nu.$$

Using the complementary coordinate subspace $\{0\} \times \mathbb{R}^{d-r_w}$ in the dual variational formula gives

$$\lambda_{r_w+1}(H_\nu) = \min_{\substack{\mathcal{L}\subset\mathbb{R}^d \\ \dim(\mathcal{L})=d-r_w}} \max_{\substack{x\in\mathcal{L} \\ \|x\|_2=1}} x^\top H_\nu x \leq \beta_\nu.$$

It follows that

$$\lambda_{r_w}(\Sigma_{\hat{w},\nu}) - \lambda_{r_w+1}(\Sigma_{\hat{w},\nu}) = \lambda_{r_w}(H_\nu) - \lambda_{r_w+1}(H_\nu) \geq \alpha_\nu - \beta_\nu = g_\nu,$$

which proves (27).

It remains to control the rotation of the leading eigenspace. Let $\widehat{V}_- \in \mathbb{R}^{d\times(d-r_w)}$ have orthonormal columns spanning the bottom $(d-r_w)$-dimensional eigenspace of H_ν, and write

$$H_\nu \widehat{V}_- = \widehat{V}_- \Lambda_-.$$

Let

$$U_0 := \begin{bmatrix} I_{r_w} \\ 0 \end{bmatrix}, \qquad V_0 := \begin{bmatrix} 0 \\ I_{d-r_w} \end{bmatrix}, \qquad S := \widehat{V}_-^\top U_0.$$

From (29),

$$H_\nu U_0 = U_0 A_\nu + V_0 K_\nu^\top.$$

Premultiplying this identity by $\widehat{V}_-^\top$ and using $\widehat{V}_-^\top H_\nu = \Lambda_- \widehat{V}_-^\top$ yields the Sylvester equation

$$\Lambda_- S - S A_\nu = R, \qquad R := \widehat{V}_-^\top V_0 K_\nu^\top. \tag{30}$$

By (27),

$$\lambda_{\max}(\Lambda_-) = \lambda_{r_w+1}(H_\nu) \leq \beta_\nu, \qquad \lambda_{\min}(A_\nu) = \alpha_\nu,$$

so the spectra of Λ_- and A_ν are separated by at least $\alpha_\nu - \beta_\nu = g_\nu$. In particular, the unique solution of (30) has the convergent integral representation

$$S = -\int_0^\infty e^{t\Lambda_-} R e^{-tA_\nu}\, dt.$$

Indeed, differentiating the integrand and integrating over $[0,\infty)$ recovers (30). Moreover,

$$\begin{aligned} \|S\|_{\mathrm{op}} &\leq \int_0^\infty \|e^{t\Lambda_-}\|_{\mathrm{op}} \|R\|_{\mathrm{op}} \|e^{-tA_\nu}\|_{\mathrm{op}}\, dt \\ &\leq \|K_\nu\|_{\mathrm{op}} \int_0^\infty e^{-t(\alpha_\nu-\beta_\nu)}\, dt = \frac{a\nu\|D\|_{\mathrm{op}}}{g_\nu}. \end{aligned}$$

The singular values of $\widehat{V}_-^\top U_0$ are the sines of the principal angles between the target coordinate subspace and the leading eigenspace of H_ν. Since orthogonal conjugation by Q preserves principal angles, $\|S\|_{\mathrm{op}} = \|\Pi_\nu - \Pi_\star\|_{\mathrm{op}}$. This proves (28).

Finally, if $\nu D = 0$, then $K_\nu = 0$ and (29), together with $g_\nu > 0$, shows that $\Pi_\nu = \Pi_\star$. Conversely, if $\Pi_\nu = \Pi_\star$, then the target coordinate subspace is invariant under H_ν. Its off-target component must therefore vanish:

$$V_0^\top H_\nu U_0 = K_\nu^\top = a\nu D^\top = 0.$$

Thus $\nu D = 0$, completing the proof. □

A.4 FINITE-SAMPLE CONCENTRATION AND CONVERGENCE

To prove Theorem 11, we first bound the covariance estimation error

$$\varepsilon_{M,\nu} := \|\widehat{C}_{M,\nu} - \Sigma_{\hat{w},\nu}\|_{\mathrm{op}} \tag{31}$$

and use an eigenspace perturbation bound to control Stage 1. When the population leading subspace is uniquely determined, the empirical projector converges to it, and a stability estimate for Stage 2 gives risk convergence. For a source whose covariance is supported on the full target complement, the leading rank-r_w population projector may be nonunique; in that case we control its overlap with the target subspace to obtain the risk lower bound.

We first record the auxiliary concentration and perturbation statements used by Theorem 11.

Lemma 19 (Stage-1 covariance concentration)**.** *Assume $d > n + 3$ and fix the target and source covariance families. For the empirical covariance $\widehat{C}_{M,\nu}$ in Definition 10, there exists a constant C_{cov}, depending only on the fixed problem parameters $(d, n, \sigma^2, \Sigma_0, \Sigma_s)$, such that for every $M \geq 2$ and every $\delta \in (0, 1)$,*

$$\Pr\left(\varepsilon_{M,\nu} \leq C_{\mathrm{cov}}\, \delta^{-1/2}\, \|\Sigma_{\hat{w},\nu}\|_{\mathrm{op}} \sqrt{\frac{r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu})}{M}}\right) \geq 1 - \delta, \tag{32}$$

where the probability is over the M pretraining tasks with empirical mixing fraction ν. Moreover, for each fixed ν, $\varepsilon_{M,\nu} \to 0$ almost surely as $M \to \infty$.

Proposition 20 (Convergence of cTRE: aligned sources)**.** *Assume $\Sigma_s = (\beta_{\mathrm{on}}/\lambda)\Sigma_0$ with $\beta_{\mathrm{on}}/\lambda \geq 1$, and let $\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}$ be the cTRE estimator from Definition 10 built from M pretraining tasks at mixing fraction ν. There exists a constant C_{ridge}, depending only on the fixed target-model parameters, such that whenever $\varepsilon_{M,\nu} \leq c_{\mathrm{gap},\nu}/4$,*

$$\mathcal{E}_0\big(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}\big) \leq C_{\mathrm{ridge}}\, r_w \Big(\frac{\varepsilon_{M,\nu}}{c_{\mathrm{gap},\nu}}\Big)^2. \tag{33}$$

In the matched aligned model $\Sigma_s = (\beta_{\mathrm{on}}/\lambda)\Sigma_0$, set $\xi_\nu := 1 + \nu(\beta_{\mathrm{on}}/\lambda - 1)$. Then

$$\Sigma_{\hat{w},\nu} = a\xi_\nu \lambda \Pi_\star + \tau_\nu I_d, \qquad \tau_\nu := b\xi_\nu \lambda r_w + c\sigma^2, \tag{34}$$

so $c_{\mathrm{gap},\nu} = a\xi_\nu\lambda$ and

$$r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu}) = \frac{\mathrm{trace}(\Sigma_{\hat{w},\nu})}{\|\Sigma_{\hat{w},\nu}\|_{\mathrm{op}}} = \frac{d\tau_\nu + r_w a\xi_\nu\lambda}{\tau_\nu + a\xi_\nu\lambda}. \tag{35}$$

Moreover, $c_{\mathrm{gap},\nu} = a\lambda(1 + \nu(\beta_{\mathrm{on}}/\lambda - 1))$ is nondecreasing in ν, and $r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu})$ in (35) is nonincreasing in ν.

Proposition 21 (Convergence of cTRE: pure off-support phase transition)**.** *Assume the pure full-complement off-support source model $\Sigma_s = \beta_{\mathrm{off}}\Pi_\perp$, where $\beta_{\mathrm{off}} > 0$ and $\Pi_\perp = I_d - \Pi_\star$. Let $\nu_{\mathrm{c}} = \lambda/(\lambda + \beta_{\mathrm{off}})$ and let $\hat{\Pi}_{M,\nu} := \widehat{U}_{r_w}\widehat{U}_{r_w}^\top$ be the Stage 1 projector in Definition 10. There exists a constant C_{ridge}, depending only on the fixed target-model parameters, such that the following two regimes hold.*

(a) Subcritical convergence. *If $0 \leq \nu < \nu_{\mathrm{c}}$, set*

$$\gamma_\nu^- := a\big((1-\nu)\lambda - \nu\beta_{\mathrm{off}}\big) > 0.$$

Whenever $\varepsilon_{M,\nu} \leq \gamma_\nu^-/4$,

$$\mathcal{E}_0\big(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}\big) \leq C_{\mathrm{ridge}}\, r_w \Big(\frac{\varepsilon_{M,\nu}}{\gamma_\nu^-}\Big)^2. \tag{36}$$

Consequently, if $\varepsilon_{M,\nu} \to 0$ as $M \to \infty$, then $\hat{\Pi}_{M,\nu} \to \Pi_\star$ and $\mathcal{E}_0(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}) \to 0$.

(b) Supercritical target loss. *If $\nu > \nu_{\mathrm{c}}$ and $d - r_w \geq r_w$, then almost surely over the pretraining data,*

$$\liminf_{M\to\infty} \mathcal{E}_0\big(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}\big) \geq \mathbb{E}\big[\|w_\star(X,Y)\|_2^2\big] > 0. \tag{37}$$

Proof of Lemma 19. Write $Z_j := \widehat{w}^j$. Let $S_j \in \{\Sigma_0, \Sigma_s\}$ denote the task-prior covariance used by task j. For a fixed empirical mixing fraction ν, the average task covariance is $M^{-1}\sum_{j=1}^M S_j = \bar{\Sigma}_\nu$; equivalently, a task drawn uniformly from the pooled pretraining collection has task covariance $\bar{\Sigma}_\nu$. Lemma 7 gives

$$\mathbb{E}[Z_j] = 0, \qquad \mathrm{Cov}(Z_j) = \Sigma_{\hat{w},j} := aS_j + \big(b\,\mathrm{tr}(S_j) + c\sigma^2\big)I_d, \qquad \frac{1}{M}\sum_{j=1}^M \Sigma_{\hat{w},j} = \Sigma_{\hat{w},\nu}.$$

Step 1: fourth moment bounds. We first show a fourth-moment bound for Z_j and $\bar{Z} = M^{-1}\sum_{j=1}^M Z_j$. Decompose the mLSE vector as

$$Z_j = P_{X^j} w^j + X^{j,\top}(X^j X^{j,\top})^{-1}\varepsilon^j, \qquad P_{X^j} := X^{j,\top}(X^j X^{j,\top})^{-1}X^j.$$

For the signal term $P_{X^j} w^j$, P_{X^j} is an orthogonal projector, so $\|P_{X^j}\|_{\mathrm{op}} \le 1$. Hence, for $w^j \sim \mathcal{N}(0, S_j)$,

$$\mathbb{E}\|P_{X^j} w^j\|_2^4 \le \mathbb{E}\|w^j\|_2^4 = 2\,\mathrm{tr}(S_j^2) + \{\mathrm{tr}(S_j)\}^2 \le 3\{\mathrm{tr}(S_j)\}^2. \tag{38}$$

For the noise term $N_j := X^{j,\top} W_j^{-1}\varepsilon^j$ with $W_j := X^j X^{j,\top}$, note that $\varepsilon^j \sim \mathcal{N}(0, \sigma^2 I_n)$ is independent of X^j, so conditional on X^j,

$$N_j \mid X^j \quad \sim \mathcal{N}\big(0, A_j(X^j)\big), \qquad A_j(X^j) := \sigma^2 X^{j,\top} W_j^{-2} X^j.$$

For a centered Gaussian vector $g \sim \mathcal{N}(0, A)$, $\mathbb{E}\|g\|_2^4 = 2\,\mathrm{tr}(A^2) + \{\mathrm{tr}(A)\}^2$. The nonzero eigenvalues of $X^{j,\top} W_j^{-2} X^j$ are the eigenvalues of W_j^{-1}. Therefore

$$\mathbb{E}\big[\|N_j\|_2^4 \mid X^j\big] = \sigma^4\left\{2\,\mathrm{tr}(W_j^{-2}) + \big(\mathrm{tr}(W_j^{-1})\big)^2\right\} \le 3\sigma^4\big(\mathrm{tr}(W_j^{-1})\big)^2.$$

Since $W_j \sim \mathrm{Wishart}_n(I_n, d)$, the inverse-Wishart second moment $\mathbb{E}\{\mathrm{tr}(W_j^{-1})\}^2$ is finite under condition $d > n + 3$. Thus there is a finite constant $K_{d,n}$ such that

$$\mathbb{E}\|X^{j,\top}(X^j X^{j,\top})^{-1}\varepsilon^j\|_2^4 \le K_{d,n}\sigma^4. \tag{39}$$

Combining (38) and (39) with $(u+v)^4 \le 8(u^4 + v^4)$ gives

$$\mathbb{E}\|Z_j\|_2^4 \le 24\{\mathrm{tr}(S_j)\}^2 + 8K_{d,n}\sigma^4.$$

Moreover, $\Sigma_{\hat{w},\nu} = a\bar{\Sigma}_\nu + (b\,\mathrm{tr}(\bar{\Sigma}_\nu) + c\sigma^2)I_d \succeq c\sigma^2 I_d$, where $c > 0$ is the coefficient in Lemma 7; hence $\mathrm{tr}(\Sigma_{\hat{w},\nu}) \ge dc\sigma^2 > 0$ uniformly in ν. Since $\mathrm{tr}(\bar{\Sigma}_\nu) = (1-\nu)\,\mathrm{tr}(\Sigma_0) + \nu\,\mathrm{tr}(\Sigma_s)$,

$$B_4^{\max} := 24\max\big\{\{\mathrm{tr}(\Sigma_0)\}^2, \{\mathrm{tr}(\Sigma_s)\}^2\big\} + 8K_{d,n}\sigma^4$$

bounds $\mathbb{E}\|Z_j\|_2^4$ for every target or source task. Therefore, with

$$K_4 := \frac{B_4^{\max}}{(dc\sigma^2)^2},$$

which depends only on $(d, n, \sigma^2, \Sigma_0, \Sigma_s)$ and not on M or ν, we have

$$\frac{1}{M}\sum_{j=1}^M \mathbb{E}\|Z_j\|_2^4 \le K_4\{\mathrm{tr}(\Sigma_{\hat{w},\nu})\}^2 \tag{40}$$

for every M and ν.

To get a fourth-moment bound for $\bar{Z}$, let $S_M := \sum_{j=1}^M Z_j$, so $\bar{Z} = M^{-1}S_M$. Since

$$\|S_M\|_2^2 = \sum_{j=1}^M \|Z_j\|_2^2 + 2\sum_{1\le i<j\le M}\langle Z_i, Z_j\rangle,$$

independence and centering imply that all mixed products containing an unpaired centered vector have expectation zero. Therefore

$$\mathbb{E}\|S_M\|_2^4 = \sum_{j=1}^M \mathbb{E}\|Z_j\|_2^4 + 2\sum_{1\le i<j\le M}\mathbb{E}\big[\|Z_i\|_2^2\|Z_j\|_2^2\big] + 4\sum_{1\le i<j\le M}\mathbb{E}\langle Z_i, Z_j\rangle^2.$$

For $i \neq j$, Cauchy–Schwarz gives

$$\mathbb{E}\big[\|Z_i\|_2^2\|Z_j\|_2^2\big] \le \big(\mathbb{E}\|Z_i\|_2^4\big)^{1/2}\big(\mathbb{E}\|Z_j\|_2^4\big)^{1/2},$$

and the same bound holds for $\mathbb{E}\langle Z_i, Z_j\rangle^2$ because $\langle Z_i, Z_j\rangle^2 \le \|Z_i\|_2^2\|Z_j\|_2^2$. Writing $a_j := \mathbb{E}\|Z_j\|_2^4$, we obtain

$$\mathbb{E}\|S_M\|_2^4 \le \sum_{j=1}^{M} a_j + 6 \sum_{1\le i<j\le M} a_i^{1/2} a_j^{1/2} \le 4M \sum_{j=1}^{M} a_j,$$

where the last step uses $\sum_{i<j} a_i^{1/2} a_j^{1/2} \le \frac{1}{2}(\sum_j a_j^{1/2})^2 \le \frac{M}{2}\sum_j a_j$. Thus, by (40),

$$\mathbb{E}\|\bar{Z}\|_2^4 = \frac{1}{M^4}\mathbb{E}\|S_M\|_2^4 \le \frac{4}{M^3}\sum_{j=1}^{M}\mathbb{E}\|Z_j\|_2^4 \le \frac{4K_4\{\mathrm{tr}(\Sigma_{\hat{w},\nu})\}^2}{M^2}.$$

Step 2: covariance estimation error. Let

$$\widetilde{C}_M := \frac{1}{M}\sum_{j=1}^{M} Z_j Z_j^\top, \qquad \bar{Z} := \frac{1}{M}\sum_{j=1}^{M} Z_j.$$

Independence, centering of $Z_j Z_j^\top - \Sigma_{\hat{w},j}$, and $\mathbb{E}[\|Z_j Z_j^\top\|_F^2] = \mathbb{E}[\|Z_j\|_2^4]$ give

$$\mathbb{E}\|\widetilde{C}_M - \Sigma_{\hat{w},\nu}\|_F^2 = \frac{1}{M^2}\sum_{j=1}^{M}\mathbb{E}\|Z_j Z_j^\top - \Sigma_{\hat{w},j}\|_F^2 \le \frac{1}{M^2}\sum_{j=1}^{M}\mathbb{E}\|Z_j\|_2^4 \le \frac{K_4\{\mathrm{tr}(\Sigma_{\hat{w},\nu})\}^2}{M}. \quad (41)$$

The empirical covariance in (17) satisfies $\widehat{C}_{M,\nu} = \frac{M}{M-1}\big(\widetilde{C}_M - \bar{Z}\bar{Z}^\top\big)$, so

$$\widehat{C}_{M,\nu} - \Sigma_{\hat{w},\nu} = \frac{M}{M-1}(\widetilde{C}_M - \Sigma_{\hat{w},\nu}) + \frac{1}{M-1}\Sigma_{\hat{w},\nu} - \frac{M}{M-1}\bar{Z}\bar{Z}^\top.$$

Let $T_\nu := \mathrm{tr}(\Sigma_{\hat{w},\nu})$. For $M \ge 2$, we have $(M/(M-1))^2 \le 4$, $(M-1)^{-2} \le 4/M^2 \le 4/M$, and $\|\Sigma_{\hat{w},\nu}\|_F^2 \le T_\nu^2$. Using $(x+y+z)^2 \le 3(x^2+y^2+z^2)$, (41), and the bound on $\mathbb{E}\|\bar{Z}\|_2^4$ from Step 1,

$$\begin{aligned}\mathbb{E}\|\widehat{C}_{M,\nu} - \Sigma_{\hat{w},\nu}\|_F^2 &\le 12\mathbb{E}\|\widetilde{C}_M - \Sigma_{\hat{w},\nu}\|_F^2 + \frac{3}{(M-1)^2}\|\Sigma_{\hat{w},\nu}\|_F^2 + 12\mathbb{E}\|\bar{Z}\|_2^4 \\ &\le \frac{12K_4T_\nu^2}{M} + \frac{12T_\nu^2}{M} + \frac{48K_4T_\nu^2}{M^2} \le \frac{(60K_4+12)T_\nu^2}{M}.\end{aligned}$$

Since $\varepsilon_{M,\nu} = \|\widehat{C}_{M,\nu} - \Sigma_{\hat{w},\nu}\|_{\mathrm{op}} \le \|\widehat{C}_{M,\nu} - \Sigma_{\hat{w},\nu}\|_F$, this gives

$$\mathbb{E}\varepsilon_{M,\nu}^2 \le \frac{(60K_4+12)\{\mathrm{tr}(\Sigma_{\hat{w},\nu})\}^2}{M}.$$

Finally,

$$\mathrm{tr}(\Sigma_{\hat{w},\nu}) = \|\Sigma_{\hat{w},\nu}\|_{\mathrm{op}} r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu}) \le \sqrt{d}\,\|\Sigma_{\hat{w},\nu}\|_{\mathrm{op}}\sqrt{r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu})},$$

where $1 \le r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu}) \le d$. Therefore

$$\mathbb{E}\varepsilon_{M,\nu}^2 \le C_{\mathrm{cov}}^2\|\Sigma_{\hat{w},\nu}\|_{\mathrm{op}}^2\frac{r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu})}{M}, \quad C_{\mathrm{cov}} := \sqrt{d(60K_4+12)}.$$

Then, Markov's inequality applied to $\varepsilon_{M,\nu}^2$ proves (32).

For almost-sure statements, couple the samples by taking the first $M_0(M)$ and $M_s(M)$ observations from two independent infinite iid target and source streams, with $\nu_M = M_s(M)/M \to \nu$. The entries of each stream's $ZZ^\top$ and Z have finite first moments. Apply the strong law separately to each stream and weight by its sample fraction; a stream with bounded size has vanishing weight. This gives $\widetilde{C}_M \to \Sigma_{\hat{w},\nu}$ and $\bar{Z} \to 0$ entrywise almost surely. Since the covariance map is continuous in ν, the error relative to $\Sigma_{\hat{w},\nu_M}$ also tends to zero. Fixed-ν statements use an admissible sequence of integer task counts. □

We use the following Davis–Kahan bound, whose short residual proof we include: if $\widetilde{\Sigma}$ and $\Sigma_{\hat{w},\nu}$ are symmetric with $\|\widetilde{\Sigma} - \Sigma_{\hat{w},\nu}\|_{\mathrm{op}} \le \Delta$ and the eigengap of $\Sigma_{\hat{w},\nu}$ between its top-r_w eigenspace $\mathrm{range}(\Pi)$ and the rest is $\gamma > 2\Delta$, then the top-r_w projector $\hat{\Pi}$ of $\widetilde{\Sigma}$ satisfies

$$\big\|\hat{\Pi} - \Pi\big\|_F \le \frac{\sqrt{2r_w}\,\Delta}{\gamma - \Delta}. \tag{42}$$

To verify (42), write $A = \Sigma_{\hat{w},\nu}$, $E = \widetilde{\Sigma} - A$, and choose orthonormal eigenbases $U_\perp$ for the bottom population subspace and $\widehat{U}$ for the top empirical subspace. If their diagonal eigenvalue matrices are $\Lambda_\perp$ and $\widehat{\Lambda}$, then $V = U_\perp^\top \widehat{U}$ satisfies

$$\Lambda_\perp V - V\widehat{\Lambda} = -U_\perp^\top E \widehat{U}.$$

Weyl's inequality separates these two spectra by at least $\gamma - \Delta$. Entrywise division therefore gives $\|V\|_F \le \sqrt{r_w}\Delta/(\gamma - \Delta)$, and the identity $\|\hat{\Pi} - \Pi\|_F^2 = 2\|V\|_F^2$ proves the claim.

Proof of Proposition 20. Set $\Delta = \varepsilon_{M,\nu}$ and $\xi_\nu = 1 + \nu(\beta_{\mathrm{on}}/\lambda - 1)$. Under $\Sigma_s = (\beta_{\mathrm{on}}/\lambda)\Sigma_0$, the covariance $\Sigma_{\hat{w},\nu}$ has the aligned form (34) by Corollary 8. Its top-r_w eigenspace is $\mathrm{range}(\Pi_\star)$, and the eigengap between the top-r_w eigenvalues and the remaining eigenvalues is

$$c_{\mathrm{gap},\nu} = a\xi_\nu\lambda = a\lambda(1 + \nu(\beta_{\mathrm{on}}/\lambda - 1)).$$

The hypothesis $\Delta \le c_{\mathrm{gap},\nu}/4$ plus (42) applied with $\Pi = \Pi_\star$ and $\gamma = c_{\mathrm{gap},\nu}$ yields

$$\|\hat{\Pi} - \Pi_\star\|_F^2 \le \frac{2r_w\Delta^2}{(c_{\mathrm{gap},\nu} - \Delta)^2} \le \frac{2r_w\Delta^2}{(3c_{\mathrm{gap},\nu}/4)^2} = \frac{32}{9}\, r_w \Big(\frac{\Delta}{c_{\mathrm{gap},\nu}}\Big)^2. \tag{43}$$

For any rank-r_w projector Π, define the ridge map

$$\hat{w}_\Pi(X, Y) := \Pi X^\top (X\Pi X^\top + \rho I_n)^{-1} Y.$$

For the Stage 2 ridge map at $\rho = \sigma^2/\lambda$, the error bound

$$\mathcal{E}_0\big(\hat{w}_\Pi\big) = \mathbb{E}\big[\|\hat{w}_\Pi(X, Y) - w_\star(X, Y)\|_2^2\big] \le C_0 \big\|\Pi - \Pi_\star\big\|_F^2 \tag{44}$$

holds with a constant C_0 depending only on the fixed target-model parameters. To see this, set

$$\Delta_\Pi := \Pi - \Pi_\star, \qquad A := X\Pi X^\top + \rho I_n, \qquad B := X\Pi_\star X^\top + \rho I_n.$$

Since $A \succeq \rho I_n$ and $B \succeq \rho I_n$,

$$\|A^{-1}\|_{\mathrm{op}} \le \rho^{-1}, \qquad \|B^{-1}\|_{\mathrm{op}} \le \rho^{-1}.$$

Decompose the estimator difference as

$$\begin{aligned}\hat{w}_\Pi(X, Y) - \hat{w}_{\Pi_\star}(X, Y) &= \Pi X^\top A^{-1} Y - \Pi_\star X^\top B^{-1} Y \\ &= \underbrace{\Delta_\Pi X^\top A^{-1} Y}_{T_1} + \underbrace{\Pi_\star X^\top (A^{-1} - B^{-1}) Y}_{T_2}.\end{aligned}$$

The direct projector term satisfies

$$\|T_1\|_2 \le \|\Delta_\Pi\|_{\mathrm{op}} \|X\|_{\mathrm{op}} \|A^{-1}\|_{\mathrm{op}} \|Y\|_2 \le \rho^{-1} \|\Delta_\Pi\|_F \|X\|_{\mathrm{op}} \|Y\|_2. \tag{45}$$

For the resolvent term, use

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}, \qquad B - A = X(\Pi_\star - \Pi)X^\top = -X\Delta_\Pi X^\top.$$

Thus

$$\begin{aligned}\|T_2\|_2 &\le \|X\|_{\mathrm{op}} \|A^{-1}\|_{\mathrm{op}} \|X\|_{\mathrm{op}} \|\Delta_\Pi\|_{\mathrm{op}} \|X\|_{\mathrm{op}} \|B^{-1}\|_{\mathrm{op}} \|Y\|_2 \\ &\le \rho^{-2} \|\Delta_\Pi\|_F \|X\|_{\mathrm{op}}^3 \|Y\|_2.\end{aligned} \tag{46}$$

Combining (45) and (46) with $(u + v)^2 \le 2u^2 + 2v^2$ gives

$$\|\hat{w}_\Pi(X, Y) - w_\star(X, Y)\|_2^2 \le \big(2\rho^{-2}\|X\|_{\mathrm{op}}^2 \|Y\|_2^2 + 2\rho^{-4}\|X\|_{\mathrm{op}}^6 \|Y\|_2^2\big)\, \|\Delta_\Pi\|_F^2.$$

Taking expectation over a fresh target task gives (44) with, for example,

$$C_0 := 2\rho^{-2}\mathbb{E}\big[\|X\|_{\mathrm{op}}^2 \|Y\|_2^2\big] + 2\rho^{-4}\mathbb{E}\big[\|X\|_{\mathrm{op}}^6 \|Y\|_2^2\big].$$

This constant is finite because X, w, and ε are Gaussian in fixed dimensions and $Y = Xw + \varepsilon$ is a polynomial function of them; it depends only on the fixed target-model parameters. Combining (43) with (44) gives (33) with $C_{\mathrm{ridge}} = (32/9)C_0$. □

A.5 Closed-form expressions for C_{cov} and C_{fin}

We give explicit choices of the constants in Lemma 19, Proposition 20, and Theorem 11. These constants depend on the fixed model parameters $(d, n, \sigma^2, \Sigma_0, \Sigma_s)$ and are independent of (M, ν, δ). Their bounds use inverse-Wishart moments, the coefficients of the covariance map, and Gaussian moments of the fresh-task design.

Inverse-Wishart and trace bounds. For $W \sim \mathrm{Wishart}_n(I_n, d)$ with $d > n + 3$, the inverse-Wishart moments give the closed forms (e.g., (Cook & Forzani, 2011))

$$\mathbb{E}\ \mathrm{tr}(W^{-1}) = \frac{n}{d-n-1}, \qquad \mathrm{Var}\ \mathrm{tr}(W^{-1}) = \frac{2n(d-1)}{(d-n)(d-n-1)^2(d-n-3)},$$

so $\mathbb{E}\{\mathrm{tr}(W^{-1})\}^2 = \mathrm{Var}(\mathrm{tr}(W^{-1})) + (\mathbb{E}\ \mathrm{tr}(W^{-1}))^2$ is bounded by

$$K_{d,n} := 3\left(\tfrac{n^2}{(d-n-1)^2} + \tfrac{2n(d-1)}{(d-n)(d-n-1)^2(d-n-3)}\right) = O\big(n^2/(d-n)^2\big).$$

Setting $T_{\max} := \max\{\mathrm{tr}(\Sigma_0), \mathrm{tr}(\Sigma_s)\}$ and using $\Sigma_{\hat{w},\nu} \succeq c\sigma^2 I_d$, so that $\mathrm{tr}(\Sigma_{\hat{w},\nu}) \geq dc\sigma^2$, the unitless ratio

$$K_4 = \frac{24\, T_{\max}^2 + 8K_{d,n}\sigma^4}{(d\, c\, \sigma^2)^2}$$

collects the fourth-moment dependence of the covariance concentration bound. The dimension d also appears explicitly in C_{cov} below.

Covariance concentration constant C_{cov}. By the proof of Lemma 19 (Step 2),

$$C_{\mathrm{cov}} = \sqrt{d\,(60\, K_4 + 12)}. \tag{47}$$

Thus all dependence of C_{cov} on $(d, n, \sigma^2, \Sigma_0, \Sigma_s)$ is explicit through d, K_4, and the covariance-map coefficient $c = n/[d(d-n-1)]$. There is no hidden dependence on M, ν, or δ.

Stage-2 sensitivity constant C_0 and C_{ridge}. From the proof of Proposition 20,

$$C_0 = 2\rho^{-2}\,\mathbb{E}\big[\|X\|_{\mathrm{op}}^2\,\|Y\|_2^2\big] + 2\rho^{-4}\,\mathbb{E}\big[\|X\|_{\mathrm{op}}^6\,\|Y\|_2^2\big], \qquad \rho = \sigma^2/\lambda. \tag{48}$$

Conditioning on X,

$$\mathbb{E}\big[\|Y\|_2^2 \mid X\big] = \mathrm{tr}(X\Sigma_0 X^\top) + n\sigma^2 \leq \|X\|_{\mathrm{op}}^2\,\mathrm{tr}(\Sigma_0) + n\sigma^2.$$

Therefore

$$C_0 \leq 2\rho^{-2}\big\{\mathrm{tr}(\Sigma_0)\mathbb{E}\|X\|_{\mathrm{op}}^4 + n\sigma^2\mathbb{E}\|X\|_{\mathrm{op}}^2\big\} + 2\rho^{-4}\big\{\mathrm{tr}(\Sigma_0)\mathbb{E}\|X\|_{\mathrm{op}}^8 + n\sigma^2\mathbb{E}\|X\|_{\mathrm{op}}^6\big\}.$$

Using $\|X\|_{\mathrm{op}}^2 \leq \mathrm{tr}(X^\top X)$ for $X \in \mathbb{R}^{n\times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries gives $\mathbb{E}\|X\|_{\mathrm{op}}^{2k} = O((nd)^k)$, and hence the convenient polynomial bound

$$C_0 = O\Big(\rho^{-2}n^2d^2\,(\mathrm{tr}(\Sigma_0) + \sigma^2) + \rho^{-4}n^4d^4\,(\mathrm{tr}(\Sigma_0) + \sigma^2)\Big).$$

The Davis–Kahan factor adds a multiplicative $32/9$, so

$$C_{\mathrm{ridge}} = \tfrac{32}{9}\,C_0 = O\big((\rho^{-2}n^2d^2 + \rho^{-4}n^4d^4)(\mathrm{tr}(\Sigma_0) + \sigma^2)\big). \tag{49}$$

Finite-sample constant C_{fin}. Combining (47) and (49), the constant C_{fin} in Theorem 11 is

$$C_{\mathrm{fin}} = C_{\mathrm{ridge}}\,C_{\mathrm{cov}}^2 = \tfrac{32}{9}\,C_0\,d\,(60\,K_4 + 12) = O\big((\rho^{-2}n^2d^3 + \rho^{-4}n^4d^5)(\mathrm{tr}(\Sigma_0) + \sigma^2)(1 + K_4)\big). \tag{50}$$

For the aligned model $\Sigma_s = (\beta_{\mathrm{on}}/\lambda)\Sigma_0$, we have $\mathrm{tr}(\Sigma_0) = r_w\lambda$, giving

$$C_{\mathrm{fin}} = O\big((\rho^{-2}n^2d^3 + \rho^{-4}n^4d^5)(r_w\lambda + \sigma^2)(1 + K_4)\big).$$

The dependence on M and δ appears explicitly in the risk bound as $1/(M\delta)$; the dependence on ν remains in the covariance norm, effective rank, and recovery margin.

Dependence on dimension. The factor $\sqrt{d}$ in C_{cov} comes from $r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu}) \le d$ when rewriting the trace-moment bound in effective-rank form. Further dimension dependence enters through K_4, including the inverse-Wishart moments and the lower bound $\mathrm{tr}(\Sigma_{\hat{w},\nu}) \ge dc\sigma^2$. Replacing the second-moment/Frobenius step by a matrix Bernstein argument (e.g., Theorem 6.1 of (Tropp, 2012)) can reduce the dimension order under stronger tail assumptions on the per-task mLSE, but does not change the qualitative phase-transition message of Theorem 11. Similarly, the d^4 trace-moment factor in C_{ridge} is a loose consequence of bounding powers of $\|X\|_{\mathrm{op}}$ by raw trace moments; sharper Gaussian spectral-norm bounds improve the polynomial but not the M^{-1} rate.

Proof of Proposition 21. Under the pure full-complement off-support model,

$$\bar{\Sigma}_\nu = (1-\nu)\lambda\Pi_\star + \nu\beta_{\mathrm{off}}\Pi_\perp, \qquad \Pi_\perp = I_d - \Pi_\star,$$

and Corollary 8 gives

$$\Sigma_{\hat{w},\nu} = a(1-\nu)\lambda\Pi_\star + a\nu\beta_{\mathrm{off}}\Pi_\perp + \tau_\nu^{\mathrm{iso}} I_d, \quad \tau_\nu^{\mathrm{iso}} := b\big(r_w(1-\nu)\lambda + \nu\beta_{\mathrm{off}}(d-r_w)\big) + c\sigma^2. \tag{51}$$

Thus, after removing the common isotropic term, the covariance has one eigenvalue on the target subspace and one on its orthogonal complement.

By Lemma 19, the pooled Stage 1 empirical covariance is consistent:

$$\varepsilon_{M,\nu} = \big\|\widehat{C}_{M,\nu} - \Sigma_{\hat{w},\nu}\big\|_{\mathrm{op}} \xrightarrow[M\to\infty]{\text{a.s.}} 0. \tag{52}$$

First suppose $0 \le \nu < \nu_{\mathrm{c}}$. Then the target eigenvalue exceeds the off-target eigenvalue by

$$c_{\mathrm{gap},\nu} = a\big((1-\nu)\lambda - \nu\beta_{\mathrm{off}}\big) > 0,$$

and (51) shows that the population top-r_w eigenspace is $\mathrm{range}(\Pi_\star)$, separated from the next eigenspace by the gap $c_{\mathrm{gap},\nu}$. If $\varepsilon_{M,\nu} \le c_{\mathrm{gap},\nu}/4$, Davis–Kahan (42) gives

$$\|\hat{\Pi}_{M,\nu} - \Pi_\star\|_F^2 \le \frac{32}{9}\, r_w \Big(\frac{\varepsilon_{M,\nu}}{c_{\mathrm{gap},\nu}}\Big)^2.$$

The bound (44), applied with $\Pi = \hat{\Pi}_{M,\nu}$, then yields (36) after increasing C_{ridge} if necessary. In particular, (52) implies $\hat{\Pi}_{M,\nu} \to \Pi_\star$ a.s. and $\mathcal{E}_0(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}) \to 0$ a.s.

Now suppose $\nu > \nu_{\mathrm{c}}$. Then

$$c_{\mathrm{gap},\nu} = -a\big(\nu\beta_{\mathrm{off}} - (1-\nu)\lambda\big) < 0,$$

and (51) shows that the full off-support cluster $\mathrm{range}(\Pi_\perp)$ is separated from the target cluster by the gap $|c_{\mathrm{gap},\nu}|$. Let $\widehat{P}_{M,\nu}^{\perp}$ be the projector onto the top $(d-r_w)$-dimensional eigenspace of $\widehat{C}_{M,\nu}$. Applying Davis–Kahan (42) to this off-support cluster together with (52) gives

$$\|\widehat{P}_{M,\nu}^{\perp} - \Pi_\perp\|_F \to 0 \qquad \text{a.s.}$$

Since $d - r_w \ge r_w$, the cTRE top-r_w projector $\hat{\Pi}_{M,\nu}$ is a subprojector of $\widehat{P}_{M,\nu}^{\perp}$. Therefore

$$\|\hat{\Pi}_{M,\nu}\Pi_\star\|_{\mathrm{op}} \le \|\widehat{P}_{M,\nu}^{\perp}\Pi_\star\|_{\mathrm{op}} = \|(\widehat{P}_{M,\nu}^{\perp} - \Pi_\perp)\Pi_\star\|_{\mathrm{op}} \to 0 \qquad \text{a.s.}$$

Then, on this event,

$$\begin{aligned}\mathcal{E}_0(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}) &\ge \mathbb{E}\big[\|(I_d - \hat{\Pi}_{M,\nu})w_\star\|_2^2\big] \qquad \text{(Lemma 16, applied with } \Pi = \hat{\Pi}_{M,\nu}) \\ &= \mathbb{E}\big[\|w_\star\|_2^2\big] - \mathbb{E}\big[\|\hat{\Pi}_{M,\nu}w_\star\|_2^2\big].\end{aligned}$$

Because $w_\star \in \mathrm{range}(\Pi_\star)$ almost surely,

$$\mathbb{E}\big[\|\hat{\Pi}_{M,\nu}w_\star\|_2^2\big] \le \|\hat{\Pi}_{M,\nu}\Pi_\star\|_{\mathrm{op}}^2\, \mathbb{E}\|w_\star\|_2^2 \to 0.$$

Taking $\liminf$ gives (37). □

Proof of Theorem 11. For the aligned sources, Proposition 20 applies and $c_{\mathrm{gap},\nu} > 0$. On the event in Lemma 19, the assumed sample-size condition implies $\varepsilon_{M,\nu} \le c_{\mathrm{gap},\nu}/4$, and hence

$$\mathcal{E}_0\big(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}\big) \le C_{\mathrm{ridge}} r_w \left(\frac{C_{\mathrm{cov}} \delta^{-1/2} \|\Sigma_{\hat{w},\nu}\|_{\mathrm{op}}}{c_{\mathrm{gap},\nu}} \sqrt{\frac{r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu})}{M}} \right)^2 .$$

This is (19) after setting $C_{\mathrm{fin}} = C_{\mathrm{ridge}} C_{\mathrm{cov}}^2$. The almost-sure convergence follows from Lemma 19, because $c_{\mathrm{gap},\nu} > 0$ is fixed for each fixed $\nu \in [0,1]$ and Proposition 20 then gives an upper bound tending to zero.

For pure off-support sources with $\nu < \nu_{\mathrm{c}}$, the same argument uses Proposition 21(a). This gives the stated high-probability rate and almost-sure convergence to zero. For $\nu > \nu_{\mathrm{c}}$, Proposition 21(b), together with the almost-sure covariance consistency from Lemma 19, gives (20). □

A.6 Extensions of the finite-sample convergence theorem

Theorem 11 treats two source families: sources proportional to the target covariance, and sources isotropic on its full orthogonal complement. We now extend the argument to general block-diagonal sources and to sources with a nonzero cross-block.

The first question is whether cTRE converges to its population rule. If the leading r_w population eigenvalues are separated from the rest, covariance concentration, eigenspace perturbation, and ridge stability give this convergence. The population rule need not be the target oracle.

For a block-diagonal source, the sign of $c_{\mathrm{gap},\nu}$ determines target recovery. A positive margin gives vanishing excess risk. A negative margin forces at least one target direction to be omitted and yields a positive risk lower bound. The target directions can be lost at different mixing fractions. Even when an eigenvalue tie prevents a unique limiting projector, compactness and the Ky Fan variational principle control every possible limit, so the lower bound still holds.

For sources that couple target and off-target directions, the cross-block rotates the population subspace. A positive margin for the block-diagonal reference bounds this rotation. The resulting risk estimate separates the population error from the sampling error, which vanishes as M grows.

Proof idea. We first compare the Stage 2 ridge rules associated with arbitrary projectors. The reverse triangle inequality then bounds the difference between their square-root excess risks, including when the limiting risk is positive. Next, the isotropy of the target oracle converts missing target directions into a risk lower bound. Combining these two facts with the covariance and eigenspace bounds proves the three results below.

Dependence on the recovery margin. For a positive margin, the risk bound is proportional to $(M c_{\mathrm{gap},\nu}^2)^{-1}$, with the other problem parameters fixed. Thus the sufficient sample size grows as the margin approaches zero. When the margin vanishes at ν_{c}, its linear dependence on ν implies that this bound ceases to be small when $|\nu - \nu_{\mathrm{c}}|$ is of order $M^{-1/2}$. This is a consequence of the bound, not a claim that the stochastic transition has been characterized sharply. When the population subspace differs from the target, increasing M removes the sampling error but leaves the population error. Let

$$\hat{\Pi}_{M,\nu} := \widehat{U}_{r_w} \widehat{U}_{r_w}^{\top}$$

denote the Stage 1 projector in Definition 10. For a rank-r_w orthogonal projector P, write

$$\hat{w}_P(X,Y) := PX^{\top}(XPX^{\top} + \rho I_n)^{-1} Y, \qquad \mathcal{R}(P) := \mathcal{E}_0(\hat{w}_P).$$

Thus $\mathcal{R}(\Pi_\star) = 0$. We also use the concentration radius

$$\eta_{M,\nu}(\delta) := C_{\mathrm{cov}}\, \delta^{-1/2} \|\Sigma_{\hat{w},\nu}\|_{\mathrm{op}} \sqrt{\frac{r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu})}{M}}, \qquad \delta \in (0,1), \tag{53}$$

from Lemma 19, where

$$r_{\mathrm{eff}}(A) := \frac{\mathrm{tr}(A)}{\|A\|_{\mathrm{op}}} \qquad \text{for } A \succeq 0,\ A \neq 0.$$

All finite-sample statements below assume $M \geq 2$, and their risks are conditional test risks given the Stage 1 data. For the almost-sure statements, take two infinite independent sequences of target and source pretraining tasks and let the empirical fractions $\nu_M = M_s(M)/M$ converge to ν. Finite-M quantities are evaluated at ν_M; we suppress this subscript to simplify notation. The affine covariance map is continuous in ν, so the empirical covariance converges to its counterpart at the limiting fraction; every strict limiting gap persists for all sufficiently large M, while the tie-robust argument below requires no cutoff gap. If $M\nu$ is integral along an admissible sequence, one may simply take $\nu_M = \nu$.

Lemma 22 (Pairwise stability of the Stage 2 ridge map)**.** *Under the target model of Section 2, let P and Q be deterministic rank-r_w orthogonal projectors. Then the constant C_0 in* (48) *satisfies*

$$\mathbb{E}\big[\|\hat{w}_P(X,Y) - \hat{w}_Q(X,Y)\|_2^2\big] \leq C_0 \|P - Q\|_F^2. \tag{54}$$

Consequently,

$$\Big|\sqrt{\mathcal{R}(P)} - \sqrt{\mathcal{R}(Q)}\Big| \leq \sqrt{C_0}\,\|P - Q\|_F. \tag{55}$$

The same statements hold conditionally when P and Q are random but independent of the fresh target task.

Lemma 23 (Isotropy of the target oracle)**.** *Under the target model of Section 2,*

$$\mathbb{E}[w_\star w_\star^\top] = \kappa_\star \Pi_\star, \qquad \kappa_\star := \frac{1}{r_w}\mathbb{E}\|w_\star(X,Y)\|_2^2 > 0. \tag{56}$$

Consequently, for every rank-r_w projector P that is deterministic or independent of the fresh target task,

$$\mathbb{E}\big[\|(I_d - P)w_\star\|_2^2 \mid P\big] = \kappa_\star\big(r_w - \mathrm{tr}(P\Pi_\star)\big) \quad \textit{almost surely}. \tag{57}$$

Proposition 24 (Finite-sample convergence to an isolated population subspace)**.** *Assume $d > n + 3$, fix $\nu \in [0,1]$, and let*

$$\Gamma_\nu := \lambda_{r_w}(\Sigma_{\hat{w},\nu}) - \lambda_{r_w+1}(\Sigma_{\hat{w},\nu}) > 0. \tag{58}$$

Let Π_ν be the unique leading rank-r_w spectral projector of $\Sigma_{\hat{w},\nu}$ and set

$$\mathcal{R}_\nu^{\mathrm{pop}} := \mathcal{R}(\Pi_\nu).$$

If $\varepsilon_{M,\nu} \leq \Gamma_\nu/4$, then

$$\|\hat{\Pi}_{M,\nu} - \Pi_\nu\|_F^2 \leq \frac{32}{9}\, r_w \left(\frac{\varepsilon_{M,\nu}}{\Gamma_\nu}\right)^2, \tag{59}$$

$$\left|\sqrt{\mathcal{E}_0\big(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}\big)} - \sqrt{\mathcal{R}_\nu^{\mathrm{pop}}}\right| \leq \sqrt{\frac{32 C_0 r_w}{9}}\,\frac{\varepsilon_{M,\nu}}{\Gamma_\nu}. \tag{60}$$

In particular, if $\eta_{M,\nu}(\delta) \leq \Gamma_\nu/4$, then both bounds hold with $\varepsilon_{M,\nu}$ replaced by $\eta_{M,\nu}(\delta)$, with probability at least $1 - \delta$. Moreover, for each fixed ν satisfying (58)*,*

$$\hat{\Pi}_{M,\nu} \longrightarrow \Pi_\nu, \qquad \mathcal{E}_0\big(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}\big) \longrightarrow \mathcal{R}_\nu^{\mathrm{pop}} \qquad \textit{almost surely}. \tag{61}$$

Corollary 25 (General block-diagonal sources)**.** *Assume $d > n + 3$ and Assumption 1. Fix $\nu \in [0,1]$ and define the on-target and off-target scores*

$$u_i := (1-\nu)\lambda + \nu\beta_i, \quad 1 \leq i \leq r_w, \qquad v_j := \nu\gamma_j, \quad 1 \leq j \leq d - r_w. \tag{62}$$

Let $\zeta_{\nu,1} \geq \cdots \geq \zeta_{\nu,d}$ be the decreasing rearrangement of $\{u_i\}_{i=1}^{r_w} \cup \{v_j\}_{j=1}^{d-r_w}$, and let

$$\mathfrak{P}_\nu := \operatorname*{arg\,max}_{\substack{P = P^\top = P^2 \\ \mathrm{tr}(P) = r_w}} \mathrm{tr}(P\Sigma_{\hat{w},\nu}), \qquad \ell_\nu := r_w - \max_{P \in \mathfrak{P}_\nu} \mathrm{tr}(P\Pi_\star). \tag{63}$$

Then $\ell_\nu \in \{0, \ldots, r_w\}$ is the smallest number of target directions omitted by a population leading subspace. The maximization permits the best choice for the target when an eigenvalue tie occurs. The following conclusions hold.

(a) Exact recovery. *If* $c_{\mathrm{gap},\nu} > 0$, *then* $\mathfrak{P}_\nu = \{\Pi_\star\}$,

$$\Gamma_\nu = c_{\mathrm{gap},\nu} = a\big(\zeta_{\nu,r_w} - \zeta_{\nu,r_w+1}\big) > 0, \qquad \ell_\nu = 0.$$

Whenever $\eta_{M,\nu}(\delta) \le c_{\mathrm{gap},\nu}/4$, *with probability at least* $1-\delta$,

$$\mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big) \le C_{\mathrm{fin}}\,\delta^{-1} r_w \frac{\|\Sigma_{\hat{w},\nu}\|_{\mathrm{op}}^2 r_{\mathrm{eff}}(\Sigma_{\hat{w},\nu})}{M c_{\mathrm{gap},\nu}^2}. \tag{64}$$

In particular, the excess risk converges to zero almost surely.

(b) Target loss without a cutoff-gap assumption. *With* $\kappa_\star$ *defined in Lemma 23, almost surely,*

$$\liminf_{M\to\infty} \mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big) \ge \kappa_\star \ell_\nu = \frac{\ell_\nu}{r_w}\mathbb{E}\|w_\star(X,Y)\|_2^2. \tag{65}$$

If $c_{\mathrm{gap},\nu} < 0$, *then* $\ell_\nu \ge 1$. *If, in addition,* $d - r_w \ge r_w$ *and*

$$\nu\gamma_{r_w} > (1-\nu)\lambda + \nu\beta_{r_w}, \tag{66}$$

then $\ell_\nu = r_w$, *so the right-hand side of* (65) *is* $\mathbb{E}\|w_\star(X,Y)\|_2^2$.

(c) Convergence under an isolated cutoff. *If*

$$\Gamma_\nu = a\big(\zeta_{\nu,r_w} - \zeta_{\nu,r_w+1}\big) > 0, \tag{67}$$

then $\mathfrak{P}_\nu = \{\Pi_\nu\}$ *and the conclusions of Proposition 24 hold. In particular, if* $\eta_{M,\nu}(\delta) \le \Gamma_\nu/4$, *then, with probability at least* $1-\delta$,

$$\left|\sqrt{\mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big)} - \sqrt{\mathcal{R}(\Pi_\nu)}\right| \le \sqrt{\frac{32 C_0 r_w}{9}}\,\frac{\eta_{M,\nu}(\delta)}{\Gamma_\nu}. \tag{68}$$

Moreover,

$$\mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big) \longrightarrow \mathcal{R}(\Pi_\nu) \quad \textit{a.s.}, \qquad \mathcal{R}(\Pi_\nu) \ge \kappa_\star \ell_\nu.$$

Corollary 26 (Finite-sample convergence beyond block-diagonal sources)**.** *Assume* $d > n+3$ *and the setting of Proposition 18. Fix* $\nu \in [0,1]$ *with* $h_\nu > 0$, *and define*

$$q_\nu := \frac{a\nu\|D\|_{\mathrm{op}}}{g_\nu} = \frac{\nu\|D\|_{\mathrm{op}}}{h_\nu}, \qquad \bar{q}_\nu := \min\{1, q_\nu\}, \qquad \mathcal{R}^{\mathrm{obl}}_\nu := \mathcal{R}(\Pi_\nu). \tag{69}$$

For a scalar t, *write* $(t)_+ := \max\{t, 0\}$. *Then*

$$\mathcal{R}^{\mathrm{obl}}_\nu \le 2 C_0 r_w \bar{q}_\nu^2. \tag{70}$$

If $\nu D \neq 0$, *also define the spectral width*

$$\omega_\nu := \lambda_{\max}(\bar{\Sigma}_\nu) - \lambda_{\min}(\bar{\Sigma}_\nu) > 0, \qquad b_\nu := \sqrt{\kappa_\star}\,\frac{\nu\|D\|_{\mathrm{op}}}{\omega_\nu};$$

if $\nu D = 0$, *set* $b_\nu := 0$. *The population excess risk satisfies*

$$b_\nu^2 \le \mathcal{R}^{\mathrm{obl}}_\nu \le 2 C_0 r_w \bar{q}_\nu^2. \tag{71}$$

Suppose $\eta_{M,\nu}(\delta) \le g_\nu/4$ *and set*

$$s_{M,\nu}(\delta) := \sqrt{\frac{32 C_0 r_w}{9}}\,\frac{\eta_{M,\nu}(\delta)}{g_\nu}. \tag{72}$$

Then, with probability at least $1-\delta$,

$$\big(b_\nu - s_{M,\nu}(\delta)\big)_+^2 \le \mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big) \le \Big(\sqrt{2 C_0 r_w}\,\bar{q}_\nu + s_{M,\nu}(\delta)\Big)^2. \tag{73}$$

Moreover,

$$\mathcal{E}_0\big(\hat{w}^{\mathrm{cTRE},\rho}_{M,\nu}\big) \longrightarrow \mathcal{R}^{\mathrm{obl}}_\nu \qquad \textit{almost surely}. \tag{74}$$

Thus, for fixed (ν, δ), *the square root of the excess risk differs from its population value by* $O(M^{-1/2})$. *When* $\nu D \neq 0$, *that population value is strictly positive.*

Proof of Lemma 22. Set

$$A_P := XPX^\top + \rho I_n, \qquad A_Q := XQX^\top + \rho I_n.$$

Both inverse norms are at most ρ^{-1}. Adding and subtracting $QX^\top A_P^{-1}Y$ and using the resolvent identity gives

$$\hat{w}_P - \hat{w}_Q = (P-Q)X^\top A_P^{-1}Y + QX^\top A_P^{-1}X(Q-P)X^\top A_Q^{-1}Y.$$

Therefore

$$\|\hat{w}_P - \hat{w}_Q\|_2 \le \left(\rho^{-1}\|X\|_{\mathrm{op}} + \rho^{-2}\|X\|_{\mathrm{op}}^3\right)\|P-Q\|_F\|Y\|_2.$$

After squaring, applying $(x+y)^2 \le 2x^2 + 2y^2$, and taking expectation, we obtain (54) with the value of C_0 in (48). Since

$$\sqrt{\mathcal{R}(P)} = \|\hat{w}_P - w_\star\|_{L^2},$$

the reverse triangle inequality in the Hilbert space $L^2(\Omega;\mathbb{R}^d)$, followed by (54), proves (55). Conditioning on the Stage 1 data proves the random-projector version. □

Proof of Lemma 23. The oracle satisfies $w_\star \in \mathrm{range}(\Pi_\star)$ almost surely. Moreover, the law of (X, w, ε) is invariant under every orthogonal change of coordinates within $\mathrm{range}(\Pi_\star)$, and the oracle map is equivariant under the same changes of coordinates. Therefore $\mathbb{E}[w_\star w_\star^\top]$ is supported on $\mathrm{range}(\Pi_\star)$ and commutes with every orthogonal transformation on that subspace. It must consequently be a scalar multiple of $\Pi_\star$. Taking the trace identifies the scalar as $\kappa_\star = r_w^{-1}\mathbb{E}\|w_\star\|_2^2$.

To see that $\kappa_\star > 0$, condition on X. The matrix $XU_\star$ has rank r_w almost surely because $r_w < n$, and the conditional covariance of the target coordinates of $w_\star$ is then positive definite. Finally,

$$\mathbb{E}\|(I_d - P)w_\star\|_2^2 = \mathrm{tr}\big((I_d - P)\mathbb{E}[w_\star w_\star^\top]\big) = \kappa_\star\big(r_w - \mathrm{tr}(P\Pi_\star)\big),$$

which proves (57). Conditioning first on P proves the random-projector statement. □

Proof of Proposition 24. Apply (42) with $\Delta = \varepsilon_{M,\nu}$, $\gamma = \Gamma_\nu$, and $\Pi = \Pi_\nu$. If $\Delta \le \Gamma_\nu/4$, then

$$\|\hat{\Pi}_{M,\nu} - \Pi_\nu\|_F^2 \le \frac{2r_w\Delta^2}{(\Gamma_\nu - \Delta)^2} \le \frac{32}{9}r_w\left(\frac{\Delta}{\Gamma_\nu}\right)^2,$$

which is (59). Conditional on the Stage 1 data, Lemma 22 with $P = \hat{\Pi}_{M,\nu}$ and $Q = \Pi_\nu$ gives

$$\left|\sqrt{\mathcal{R}(\hat{\Pi}_{M,\nu})} - \sqrt{\mathcal{R}(\Pi_\nu)}\right| \le \sqrt{C_0}\|\hat{\Pi}_{M,\nu} - \Pi_\nu\|_F.$$

This proves (60). The high-probability statement follows from Lemma 19. The same lemma gives $\varepsilon_{M,\nu_M} \to 0$ almost surely. For empirical fractions $\nu_M \to \nu$, continuity of the affine covariance map gives

$$\|\widehat{C}_{M,\nu_M} - \Sigma_{\hat{w},\nu}\|_{\mathrm{op}} \le \varepsilon_{M,\nu_M} + \|\Sigma_{\hat{w},\nu_M} - \Sigma_{\hat{w},\nu}\|_{\mathrm{op}} \to 0.$$

Apply Davis–Kahan to this fixed limiting covariance. Rates to the limiting rule additionally include an $O(|\nu_M - \nu|)$ term; the displayed rates compare rules at the same empirical fraction. Since $\Gamma_\nu > 0$ is fixed, (59) and (55) then give (61). □

Proof of Corollary 25. By Corollary 8, the isotropic term in $\Sigma_{\hat{w},\nu}$ does not affect its eigenspaces. In the basis $[U_\star\ U_\perp]$, the nonisotropic eigenvalues, before multiplication by a, are exactly the scores in (62). Hence the cutoff gap is $a(\zeta_{\nu,r_w} - \zeta_{\nu,r_w+1})$ whenever it is positive, and the Ky Fan maximum principle identifies $\mathfrak{P}_\nu$ with the set of leading rank-r_w spectral projectors.

If $c_{\mathrm{gap},\nu} > 0$, every on-target score is strictly larger than every off-target score. Thus $\Pi_\star$ is the unique population projector, its cutoff gap is $c_{\mathrm{gap},\nu}$, and $\ell_\nu = 0$. Proposition 24, (53), and $C_{\mathrm{fin}} = (32/9)C_0C_{\mathrm{cov}}^2$ yield (64) and the asserted almost-sure convergence.

We next allow a tie at the cutoff. The set of rank-r_w projectors is compact. On the almost-sure event $\|\widehat{C}_{M,\nu} - \Sigma_{\hat{w},\nu}\|_{\mathrm{op}} \to 0$, every subsequence of $\{\hat{\Pi}_{M,\nu}\}$ has a further convergent subsequence. If its limit is P, then for any rank-r_w projector Q, the variational inequality

$$\mathrm{tr}(\hat{\Pi}_{M,\nu}\widehat{C}_{M,\nu}) \ge \mathrm{tr}(Q\widehat{C}_{M,\nu})$$

passes to the limit and gives $\mathrm{tr}(P\Sigma_{\hat{w},\nu}) \geq \mathrm{tr}(Q\Sigma_{\hat{w},\nu})$. Hence $P \in \mathfrak{P}_\nu$. It follows that

$$\liminf_{M\to\infty}\{r_w - \mathrm{tr}(\hat{\Pi}_{M,\nu}\Pi_\star)\} \geq \ell_\nu \qquad \text{a.s.} \tag{75}$$

Lemma 16, followed by (57) and (75), proves (65).

For completeness, the integer-valued assertion about ℓ_ν follows by separating the score eigenspaces strictly above, equal to, and strictly below ζ_{ν,r_w}. In the tied eigenspace, which is the orthogonal sum of an on-target and an off-target eigenspace, the target overlap is maximized by choosing as many on-target basis directions as the remaining rank permits. This maximum is an integer. If $c_{\mathrm{gap},\nu} < 0$, at least one off-target score strictly exceeds an on-target score, so every maximizer loses at least one target direction and $\ell_\nu \geq 1$. Under (66), at least r_w off-target scores strictly exceed every on-target score; hence every leading rank-r_w subspace is off target and $\ell_\nu = r_w$.

Finally, if (67) holds, the population projector is unique. Proposition 24 gives (68) and the risk limit. Since $\mathfrak{P}_\nu = \{\Pi_\nu\}$, (57) and Lemma 16 give $\mathcal{R}(\Pi_\nu) \geq \kappa_\star \ell_\nu$. □

Proof of Corollary 26. Proposition 18 gives an actual cutoff gap at least g_ν and

$$\|\Pi_\nu - \Pi_\star\|_{\mathrm{op}} \leq \bar{q}_\nu.$$

Equal-rank projector geometry implies $\|\Pi_\nu - \Pi_\star\|_F^2 \leq 2r_w\|\Pi_\nu - \Pi_\star\|_{\mathrm{op}}^2$. The sensitivity bound (44) therefore proves (70).

We next prove the lower bound. Since Π_ν is a spectral projector of $\bar{\Sigma}_\nu$, it commutes with $\bar{\Sigma}_\nu$. In the basis $Q = [U_\star\ U_\perp]$,

$$\|[\bar{\Sigma}_\nu, \Pi_\star]\|_{\mathrm{op}} = \left\|\begin{bmatrix} 0 & -\nu D \\ \nu D^\top & 0 \end{bmatrix}\right\|_{\mathrm{op}} = \nu\|D\|_{\mathrm{op}}.$$

If $\nu D \neq 0$, then $\omega_\nu > 0$. With $s = (\lambda_{\max}(\bar{\Sigma}_\nu) + \lambda_{\min}(\bar{\Sigma}_\nu))/2$, the commutator identity and submultiplicativity give

$$\begin{aligned}\nu\|D\|_{\mathrm{op}} &= \|[\bar{\Sigma}_\nu - sI_d, \Pi_\star - \Pi_\nu]\|_{\mathrm{op}} \\ &\leq 2\|\bar{\Sigma}_\nu - sI_d\|_{\mathrm{op}}\|\Pi_\star - \Pi_\nu\|_{\mathrm{op}} = \omega_\nu\|\Pi_\star - \Pi_\nu\|_{\mathrm{op}}.\end{aligned}$$

Furthermore,

$$r_w - \mathrm{tr}(\Pi_\nu\Pi_\star) = \sum_{i=1}^{r_w}\sin^2\theta_i \geq \|\Pi_\nu - \Pi_\star\|_{\mathrm{op}}^2,$$

where θ_i are the principal angles. Lemmas 16 and 23 now yield

$$\mathcal{R}_\nu^{\mathrm{obl}} \geq \kappa_\star\left(\frac{\nu\|D\|_{\mathrm{op}}}{\omega_\nu}\right)^2 = b_\nu^2.$$

The same inequality is trivial when $\nu D = 0$, proving (71).

Because the actual cutoff gap is at least g_ν, Proposition 24 and Lemma 19 imply, on an event of probability at least $1-\delta$,

$$\left|\sqrt{\mathcal{E}_0\big(\hat{w}_{M,\nu}^{\mathrm{cTRE},\rho}\big)} - \sqrt{\mathcal{R}_\nu^{\mathrm{obl}}}\right| \leq s_{M,\nu}(\delta).$$

Combining this inequality with (71) and squaring proves (73). Finally, $\varepsilon_{M,\nu} \to 0$ almost surely and the actual population gap is positive, so Proposition 24 gives (74). □

A.7 Unequal target eigenvalues: support recovery and ridge regularization

We now explain which parts of the analysis depend essentially on the equal-strength target prior. Retain an orthonormal $U_\star \in \mathbb{R}^{d\times r_w}$ with $r_w < n$, and replace only the equal-eigenvalue identity in Assumption 1(i) by

$$\Sigma_0 = U_\star\Lambda U_\star^\top, \qquad \Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_{r_w}), \qquad \lambda_1 \geq \cdots \geq \lambda_{r_w} > 0. \tag{76}$$

The target support is still $\mathrm{range}(\Pi_\star)$, where $\Pi_\star = U_\star U_\star^\top$, but the variances of the target coefficients can now differ across directions.

Where equal strength is used. The equal-spectrum condition enters the paper at two distinct levels. First, it simplifies the Stage 1 spectral geometry. In Proposition 9, the on-target block is

$$(1-\nu)\lambda I_{r_w} + \nu B,$$

whose smallest eigenvalue is exactly $(1-\nu)\lambda + \nu\lambda_{\min}(B)$, regardless of the eigenvectors of B. This identity gives the gap decomposition in (16). For unequal Λ, the corresponding block is the matrix sum $(1-\nu)\Lambda + \nu B$, and its smallest eigenvalue depends on the relative eigendirections of Λ and B.

Second, and more fundamentally, equal strength makes the Bayes rule depend on the target prior only through the projector $\Pi_\star$ and one scalar ridge ratio $\rho = \sigma^2/\lambda$. Under (76), the Bayes rule instead uses a matrix-valued regularizer. Consequently, recovering $\Pi_\star$ still solves the support-recovery problem, but it no longer recovers the full target prior or the target posterior mean.

Proof idea. The covariance map remains valid, so support recovery reduces to comparing the smallest eigenvalue of the target block with the largest eigenvalue of the off-target block. A change of coordinates gives an exact criterion for the sign of this margin. After support recovery, a separate issue remains: the target posterior uses the full matrix Λ, whereas cTRE uses one scalar ridge parameter. We therefore first prove support recovery results, then state the additional covariance estimation needed for Bayes consistency.

Proposition 27 (Unequal-spectrum Stage 1 geometry)**.** *Assume* (76)*, let* $Q = [U_\star\ U_\perp]$*, and suppose first that the source is block diagonal:*

$$\Sigma_s = U_\star B U_\star^\top + U_\perp C U_\perp^\top, \qquad B \succeq 0, \quad C \succeq 0.$$

Write

$$\gamma_1 := \lambda_{\max}(C), \qquad A_\nu^{\mathrm{on}} := (1-\nu)\Lambda + \nu B,$$

and define the unequal-spectrum informative margin

$$c_{\mathrm{gap},\nu}^{\Lambda} := a\left\{\lambda_{\min}(A_\nu^{\mathrm{on}}) - \nu\gamma_1\right\} = a\,\lambda_{\min}((1-\nu)\Lambda + \nu(B - \gamma_1 I_{r_w}))\,. \tag{77}$$

Then:

(a) The population mLSE covariance is

$$\Sigma_{\hat{w},\nu} = \vartheta_\nu I_d + a U_\star A_\nu^{\mathrm{on}} U_\star^\top + a\nu U_\perp C U_\perp^\top, \tag{78}$$

where

$$\vartheta_\nu := b\left\{(1-\nu)\,\mathrm{tr}(\Lambda) + \nu\big(\mathrm{tr}(B) + \mathrm{tr}(C)\big)\right\} + c\sigma^2.$$

Thus $c_{\mathrm{gap},\nu}^{\Lambda}$ *is the exact margin between the weakest on-target direction and the strongest off-target direction. If it is positive, the unique leading rank-*r_w *projector is* $\Pi_\star$ *and its cutoff eigengap is* $c_{\mathrm{gap},\nu}^{\Lambda}$*. If it is negative, no leading rank-*r_w *subspace equals* $\mathrm{range}(\Pi_\star)$*; equality marks a nonunique cutoff. At* $\nu = 0$*,* $c_{\mathrm{gap},0}^{\Lambda} = a\lambda_{r_w}$*, so one may still define the gap ratio* $R_\nu^{\Lambda} := c_{\mathrm{gap},\nu}^{\Lambda}/(a\lambda_{r_w})$*. Unlike* (1)*, however, this ratio depends on the relative orientation of* Λ *and* B*.*

(b) Let

$$\mu_- := \lambda_{\min}\Big(\Lambda^{-1/2}(B - \gamma_1 I_{r_w})\Lambda^{-1/2}\Big), \qquad m_\nu^{\Lambda} := (1-\nu) + \nu\mu_-. \tag{79}$$

Then

$$\mathrm{sign}(c_{\mathrm{gap},\nu}^{\Lambda}) = \mathrm{sign}(m_\nu^{\Lambda}). \tag{80}$$

Accordingly, the first nonpositive-gap threshold, with the convention that it equals one when the gap stays positive on $[0,1)$*, is*

$$\nu_{\mathrm{c}}^{\Lambda} := \begin{cases} (1-\mu_-)^{-1}, & \mu_- < 0, \\ 1, & \mu_- \geq 0. \end{cases} \tag{81}$$

Whenever $m_\nu^{\Lambda} > 0$*, the useful explicit lower bound*

$$c_{\mathrm{gap},\nu}^{\Lambda} \geq a\lambda_{r_w} m_\nu^{\Lambda} > 0 \tag{82}$$

holds.

(c) Without specifying the relative eigenvectors of B and Λ, Weyl's inequality gives

$$c^{\Lambda}_{\mathrm{gap},\nu} \geq a\left\{(1-\nu)\lambda_{r_w} + \nu\lambda_{\min}(B) - \nu\gamma_1\right\}. \tag{83}$$

If B commutes with Λ, choose a common eigenbasis and denote the paired eigenvalues of B by β_i. Then the exact formula is

$$c^{\Lambda}_{\mathrm{gap},\nu} = a \min_{1\leq i\leq r_w}\left\{(1-\nu)\lambda_i + \nu(\beta_i - \gamma_1)\right\}. \tag{84}$$

The scalar decomposition in Proposition 9 is the special case $\Lambda = \lambda I_{r_w}$.

(d) More generally, suppose

$$\Sigma_s = Q\begin{bmatrix} B & D \\ D^\top & C\end{bmatrix}Q^\top \succeq 0$$

and let $\Sigma^{\mathrm{bd}}_{\hat{w},\nu}$ denote the reference covariance obtained by deleting D. If $c^{\Lambda}_{\mathrm{gap},\nu} > 0$, then the unequal-spectrum version of Proposition 18 holds:

$$\|\Sigma_{\hat{w},\nu} - \Sigma^{\mathrm{bd}}_{\hat{w},\nu}\|_{\mathrm{op}} = a\nu\|D\|_{\mathrm{op}}, \tag{85}$$

$$\lambda_{r_w}(\Sigma_{\hat{w},\nu}) - \lambda_{r_w+1}(\Sigma_{\hat{w},\nu}) \geq c^{\Lambda}_{\mathrm{gap},\nu}, \tag{86}$$

$$\|\Pi_\nu - \Pi_\star\|_{\mathrm{op}} \leq \frac{a\nu\|D\|_{\mathrm{op}}}{c^{\Lambda}_{\mathrm{gap},\nu}}. \tag{87}$$

Moreover, $\Pi_\nu = \Pi_\star$ if and only if $\nu D = 0$. No commutation condition between B and Λ is needed for these support-level statements.

Corollary 28 (Finite recovery of an unequal-spectrum target support)**.** *Assume $d > n+3$, $M \geq 2$, the block-diagonal setting of Proposition 27, and $c^{\Lambda}_{\mathrm{gap},\nu} > 0$. For an arbitrary working ridge parameter $\rho > 0$, define*

$$\mathcal{R}_\rho(P) := \mathbb{E}\Big[\big\|PX^\top(XPX^\top + \rho I_n)^{-1}Y - w_\star\big\|_2^2\Big]$$

and

$$C_0(\rho) := 2\rho^{-2}\mathbb{E}\big[\|X\|_{\mathrm{op}}^2\|Y\|_2^2\big] + 2\rho^{-4}\mathbb{E}\big[\|X\|_{\mathrm{op}}^6\|Y\|_2^2\big].$$

If $\varepsilon_{M,\nu} \leq c^{\Lambda}_{\mathrm{gap},\nu}/4$, then

$$\|\hat{\Pi}_{M,\nu} - \Pi_\star\|_F^2 \leq \frac{32}{9}r_w\left(\frac{\varepsilon_{M,\nu}}{c^{\Lambda}_{\mathrm{gap},\nu}}\right)^2, \tag{88}$$

$$\left|\sqrt{\mathcal{R}_\rho(\hat{\Pi}_{M,\nu})} - \sqrt{\mathcal{R}_\rho(\Pi_\star)}\right| \leq \sqrt{\frac{32C_0(\rho)r_w}{9}}\,\frac{\varepsilon_{M,\nu}}{c^{\Lambda}_{\mathrm{gap},\nu}}. \tag{89}$$

If $\eta_{M,\nu}(\delta) \leq c^{\Lambda}_{\mathrm{gap},\nu}/4$, the same bounds hold with $\varepsilon_{M,\nu}$ replaced by $\eta_{M,\nu}(\delta)$, with probability at least $1-\delta$. Consequently, covariance consistency gives

$$\hat{\Pi}_{M,\nu} \longrightarrow \Pi_\star, \qquad \mathcal{R}_\rho(\hat{\Pi}_{M,\nu}) \longrightarrow \mathcal{R}_\rho(\Pi_\star) \quad \textit{almost surely}.$$

In general $\mathcal{R}_\rho(\Pi_\star) > 0$; thus this is a support recovery theorem, not a Bayes-consistency theorem for the unchanged scalar-ridge cTRE.

Dependence on the source eigenvectors. Equation (83) is generally a strict inequality: the spectra of B and Λ do not determine the gap without their relative orientation. For example, let

$$\Lambda = \mathrm{diag}(10, 1), \qquad \gamma_1 = 2.$$

The two source blocks $B_1 = \mathrm{diag}(3, 0)$ and $B_2 = \mathrm{diag}(0, 3)$ have the same eigenvalues. Nevertheless, (79) gives $\mu_- = -2$ for B_1 and $\mu_- = -1/5$ for B_2, so their exact recovery thresholds are respectively $1/3$ and $5/6$. The equal target spectrum removes precisely this orientation dependence.

Corollary 29 (Sequential loss under a pure off-support source)**.** *Assume* (76)*, $\Sigma_s = \beta_{\mathrm{off}}\Pi_\perp$ with $\beta_{\mathrm{off}} > 0$, $\Pi_\perp := I_d - \Pi_\star$, and $d - r_w \geq r_w$. Define*

$$\nu_i := \frac{\lambda_i}{\lambda_i + \beta_{\mathrm{off}}}, \qquad 1 \leq i \leq r_w. \tag{90}$$

Away from the tie points $\{\nu_i\}$, the target direction associated with λ_i is retained by the population leading rank-r_w subspace exactly when $\nu < \nu_i$. Hence

$$\nu < \frac{\lambda_{r_w}}{\lambda_{r_w} + \beta_{\text{off}}} \quad \Longrightarrow \quad \Pi_\nu = \Pi_\star, \tag{91}$$

$$\nu > \frac{\lambda_1}{\lambda_1 + \beta_{\text{off}}} \quad \Longrightarrow \quad \Pi_\nu \Pi_\star = 0. \tag{92}$$

Between these two thresholds, the weaker target directions are displaced in increasing order of their eigenvalues, with simultaneous losses when target eigenvalues are repeated. More precisely, the number of strictly displaced target directions is

$$k_\nu = \#\{i : (1-\nu)\lambda_i < \nu\beta_{\text{off}}\} = \#\{i : \nu > \nu_i\}.$$

When all λ_i are equal, the thresholds in (91) *and* (92) *coincide, recovering the single all-at-once transition in Theorems 4 and 11.*

Proposition 30 (Risk from omitted target directions and scalar ridge regularization)**.** *Under* (76)*, let $Z := XU_\star$. Then:*

(a) The target posterior mean is the matrix-ridge rule

$$\begin{aligned} w_\star &= U_\star \Lambda Z^\top (Z\Lambda Z^\top + \sigma^2 I_n)^{-1} Y \\ &= U_\star (Z^\top Z + \sigma^2 \Lambda^{-1})^{-1} Z^\top Y. \end{aligned} \tag{93}$$

(b) Its second moment has the form

$$\mathbb{E}[w_\star w_\star^\top] = U_\star K_\Lambda U_\star^\top, \tag{94}$$

where

$$K_\Lambda := \mathbb{E}_Z\left[\Lambda Z^\top (Z\Lambda Z^\top + \sigma^2 I_n)^{-1} Z\Lambda\right] \succ 0. \tag{95}$$

The matrix K_Λ commutes with Λ but is generally not a scalar matrix. Set $\underline{\kappa}_\star := \lambda_{\min}(K_\Lambda) > 0$. Let P be any rank-r_w projector that is deterministic or Stage 1-measurable and independent of the fresh target task. For any estimator $\hat{w}$ whose output lies in $\mathrm{range}(P)$*,*

$$\begin{aligned} \mathcal{E}_0(\hat{w}) &\geq \mathrm{tr}\left[(I_d - P) U_\star K_\Lambda U_\star^\top\right] \\ &\geq \underline{\kappa}_\star \left\{r_w - \mathrm{tr}(P\Pi_\star)\right\}. \end{aligned} \tag{96}$$

(c) For a scalar $\rho > 0$, the projector-only ridge rule at the exact target support is

$$\hat{w}_{\Pi_\star,\rho} = U_\star (Z^\top Z + \rho I_{r_w})^{-1} Z^\top Y.$$

It equals $w_\star$ almost surely if and only if

$$\sigma^2 \Lambda^{-1} = \rho I_{r_w}, \tag{97}$$

equivalently, if and only if Λ is a scalar matrix with the matching ridge ratio.

Consequences for the existing results. The preceding propositions separate the results that extend from those that do not.

- *Results that extend at the support level.* Lemma 7 and Corollary 8 already hold for every Gaussian task covariance. The covariance concentration in Lemma 19, Davis–Kahan projector bounds, isolated-subspace convergence in Proposition 24, and the Ky Fan tie argument in Corollary 25 continue to hold after replacing the scalar block scores by the eigenvalues of A_ν^{on}. In particular, a positive $c_{\text{gap},\nu}^\Lambda$ gives the same finite-sample projector rate with $c_{\text{gap},\nu}^\Lambda$ in the denominator. For the matched proportional source $\Sigma_s = t\Sigma_0$, this gap reduces to $a\{1 + \nu(t-1)\}\lambda_{r_w}$, so the aligned-source Stage 1 conclusion remains explicit. The block target-loss lower bound also survives, with $\kappa_\star \ell_\nu$ replaced by $\underline{\kappa}_\star \ell_\nu$; the exact weighted missed-oracle-energy lower bound is obtained by minimizing the first trace in (96) over the population spectral maximizers. The covariance perturbation, spectral separation, and subspace-angle conclusions for sources that couple target and

off-target directions extend through (85)–(87). The commutator argument from Corollary 26 also gives the unequal-spectrum lower bound when $\nu D \neq 0$ (which implies $\omega_\nu > 0$):

$$\mathcal{E}_0(\hat{w}_{\Pi_\nu}) \geq \underline{\kappa}_\star \left(\frac{\nu \|D\|_{\mathrm{op}}}{\omega_\nu} \right)^2, \qquad \omega_\nu := \lambda_{\max}(\bar{\Sigma}_\nu) - \lambda_{\min}(\bar{\Sigma}_\nu).$$

This holds for any population rule supported on Π_ν; when $\nu D = 0$, take this particular lower bound to be zero.

- *Results that do not extend for the unchanged cTRE.* The implication "recover $\Pi_\star$, hence recover the Bayes rule" in Theorem 4(a) fails by (97). For the same reason, the zero-risk parts of Theorem 11 and Corollary 25(a), the sensitivity estimate (44) anchored at $w_\star$, and the population upper bound in Corollary 26 do not extend from projector control alone. Pairwise stability of two scalar-ridge maps remains valid for a chosen working ρ, so the isolated-subspace result still describes convergence to that scalar-ridge population rule; in general, however, its limiting excess risk is positive even when $\Pi_\nu = \Pi_\star$. Likewise, the oracle-isotropy identity (56) must be replaced by the weighted identity (94). Finally, under a source supported on the target complement, unequal target eigenvalues give the sequential losses in Corollary 29.

Estimating the target covariance for Stage 2. For unequal target eigenvalues, Bayes consistency requires estimating the target covariance, including its eigenvalues. For a positive semidefinite matrix S, define

$$m_S(X, Y) := SX^\top (XSX^\top + \sigma^2 I_n)^{-1} Y.$$

The following stability bound is the covariance-valued analogue of Lemma 22.

Lemma 31 (Stability of covariance-aware Stage 2)**.** *Let $S, T \succeq 0$ satisfy $\|S\|_{\mathrm{op}}, \|T\|_{\mathrm{op}} \leq L$. Then*

$$\mathbb{E}\|m_S(X, Y) - m_T(X, Y)\|_2^2 \leq C_L \|S - T\|_F^2, \tag{98}$$

where one may take

$$C_L := 2\sigma^{-4}\mathbb{E}\big[\|X\|_{\mathrm{op}}^2 \|Y\|_2^2\big] + 2L^2\sigma^{-8}\mathbb{E}\big[\|X\|_{\mathrm{op}}^6 \|Y\|_2^2\big] < \infty. \tag{99}$$

Consequently, if $L \geq \|\Sigma_0\|_{\mathrm{op}} = \lambda_1$ and a Stage 1 procedure supplies $\widehat{\Sigma}_{0,M} \succeq 0$, independently of the fresh target task, with $\|\widehat{\Sigma}_{0,M}\|_{\mathrm{op}} \leq L$ and $\widehat{\Sigma}_{0,M} \to \Sigma_0$ almost surely in Frobenius norm, then the conditional excess risk of the covariance-aware Stage 2 rule satisfies

$$\mathcal{E}_0(m_{\widehat{\Sigma}_{0,M}}) \longrightarrow 0 \qquad \textit{almost surely}.$$

The analogous conditional-risk implication holds in probability.

This lemma identifies the additional statistical requirement. For $\nu > 0$, when the affine-map coefficients and σ^2 are known or separately calibrated, the pooled Stage 1 covariance alone determines the mixed on-target block $(1-\nu)\Lambda + \nu B$, but for an arbitrary unknown B it does not identify Λ. Full Bayes recovery therefore needs further information, such as target-only calibration tasks, a known source covariance that can be subtracted from the mixture when $\nu < 1$, or a structured source model such as a known proportional source. Commutation of B and Λ can preserve the target eigenvectors, but by itself it does not identify the target eigenvalues. These additional assumptions concern identification of the target eigenvalues; the support recovery results do not require them.

Proof of Proposition 27. Corollary 8 and the block decomposition give (78), so the on-target and off-target nonisotropic eigenvalues are respectively the eigenvalues of aA_ν^{on} and $a\nu C$. This proves part (a) and (77).

For part (b), set

$$H := \Lambda^{-1/2}(B - \gamma_1 I_{r_w})\Lambda^{-1/2}.$$

The matrix inside the final expression of (77) factorizes as

$$(1-\nu)\Lambda + \nu(B - \gamma_1 I_{r_w}) = \Lambda^{1/2}\big((1-\nu)I_{r_w} + \nu H\big)\Lambda^{1/2}.$$

Sylvester's law of inertia shows that its smallest eigenvalue has the same sign as $(1-\nu)+\nu\lambda_{\min}(H) = m_\nu^\Lambda$, proving (80) and (81). If $m_\nu^\Lambda > 0$, the standard product bound for positive definite matrices yields

$$\lambda_{\min}\Big(\Lambda^{1/2}\big((1-\nu)I_{r_w} + \nu H\big)\Lambda^{1/2}\Big) \geq \lambda_{r_w} m_\nu^\Lambda,$$

which proves (82). Weyl's inequality gives (83), while simultaneous diagonalization gives (84).

For part (d), deleting D does not change the trace of the mixed covariance. The affine covariance map therefore gives the exact off-diagonal perturbation

$$Q^\top(\Sigma_{\hat{w},\nu} - \Sigma^{\mathrm{bd}}_{\hat{w},\nu})Q = a\nu \begin{bmatrix} 0 & D \\ D^\top & 0 \end{bmatrix},$$

whose operator norm is $a\nu\|D\|_{\mathrm{op}}$. The Courant–Fischer argument in the proof of Proposition 18, with A^{on}_ν in place of $(1-\nu)\lambda I_{r_w} + \nu B$, gives (86). The same Sylvester-equation argument then gives (87). The invariance argument at the end of that proof shows that equality of the projectors is equivalent to $\nu D = 0$. □

Proof of Corollary 28. By Proposition 27(a), the population leading projector is $\Pi_\star$ and its cutoff gap is $c^\Lambda_{\mathrm{gap},\nu}$. Applying (42) with this gap gives (88). The resolvent calculation in Lemma 22 is algebraic and does not use equal target eigenvalues; with the chosen working ρ, it gives

$$\left|\sqrt{\mathcal{R}_\rho(P)} - \sqrt{\mathcal{R}_\rho(Q)}\right| \le \sqrt{C_0(\rho)}\|P - Q\|_F.$$

Taking $P = \hat{\Pi}_{M,\nu}$ and $Q = \Pi_\star$ proves (89). The high-probability and almost-sure conclusions follow from Lemma 19. □

Proof of Corollary 29. The nonisotropic on-target eigenvalues are $a(1-\nu)\lambda_i$, whereas every off-target eigenvalue is $a\nu\beta_{\mathrm{off}}$. Since $d - r_w \ge r_w$, every target direction whose score is strictly below $a\nu\beta_{\mathrm{off}}$ is displaced, while every target direction whose score is strictly above it is retained. Solving $(1-\nu)\lambda_i = \nu\beta_{\mathrm{off}}$ gives (90) and all stated conclusions. □

Proof of Proposition 30. The first equality in (93) is the general Gaussian posterior-mean formula restricted to $\mathrm{range}(U_\star)$. The second follows from the standard push-through identity.

Conditional on Z, the covariance of Y is $Z\Lambda Z^\top + \sigma^2 I_n$. Since the conditional mean of $w_\star$ is zero, taking its conditional second moment and then averaging over Z gives (94)–(95). The matrix Z has full column rank almost surely because $r_w < n$, so $K_\Lambda \succ 0$. Orthogonal invariance under transformations that commute with Λ shows that K_Λ commutes with Λ.

Condition on P when it is random. If $\hat{w} \in \mathrm{range}(P)$, projecting its error onto $\mathrm{range}(P)^\perp$ yields

$$\mathcal{E}_0(\hat{w}) \ge \mathbb{E}\|(I_d - P)w_\star\|_2^2 = \mathrm{tr}\left[(I_d - P)U_\star K_\Lambda U_\star^\top\right].$$

Because $K_\Lambda \succeq \underline{\kappa}_\star I_{r_w}$, this proves (96).

Finally, conditional on a full-rank Z, the law of Y has full support in $\mathbb{R}^n$. Equality of the two linear maps in (93) and part (c) therefore implies

$$(Z^\top Z + \rho I_{r_w})^{-1} Z^\top = (Z^\top Z + \sigma^2 \Lambda^{-1})^{-1} Z^\top.$$

Since $Z^\top$ has a right inverse, the two inverse matrices, and hence their regularizers, are equal. This proves the necessity of (97); sufficiency is immediate. □

Proof of Lemma 31. Set

$$A_S := XSX^\top + \sigma^2 I_n, \qquad A_T := XTX^\top + \sigma^2 I_n.$$

Then $\|A_S^{-1}\|_{\mathrm{op}}, \|A_T^{-1}\|_{\mathrm{op}} \le \sigma^{-2}$, and the resolvent identity gives

$$m_S - m_T = (S - T)X^\top A_S^{-1} Y + TX^\top A_S^{-1} X(T - S)X^\top A_T^{-1} Y.$$

Consequently,

$$\|m_S - m_T\|_2 \le \left(\sigma^{-2}\|X\|_{\mathrm{op}} + L\sigma^{-4}\|X\|_{\mathrm{op}}^3\right)\|S - T\|_F\|Y\|_2.$$

Squaring, applying $(x+y)^2 \le 2x^2 + 2y^2$, and taking expectation proves (98)–(99). The consistency statement follows by conditioning on the Stage 1 estimate and taking $T = \Sigma_0$. □

B ADDITIONAL NUMERICAL DETAILS

B.1 COMMON EXPERIMENTAL DETAILS

Data generation. All experiments use the regression model in Section 2. For each random seed, we draw an orthonormal target basis $U_\star$ uniformly, so that $\Pi_\star = U_\star U_\star^\top$ is uniform on $\mathrm{Gr}(r_w, d)$. In E1–E2, this subspace and the target test set are fixed across ν and M within that seed. Of the M pretraining tasks, $M_0 = \lceil (1-\nu)M \rceil$ are target tasks and the rest are source tasks. Independently for each task, we draw

$$w^j \sim \mathcal{N}(0, \Sigma_\nu^j), \qquad X_{ik}^j \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad \varepsilon^j \sim \mathcal{N}(0, \sigma^2 I_n), \qquad Y^j = X^j w^j + \varepsilon^j,$$

where $\Sigma_\nu^j = \Sigma_0$ for $j \le M_0$ and $\Sigma_\nu^j = \Sigma_s$ otherwise. Test tasks are independent draws from the target distribution. For the off-support family in E1–E2, we draw r_w orthonormal columns of V_s uniformly in $\mathrm{range}(U_\star)^\perp$ and set $\Pi_s = V_s V_s^\top$. The dTRE and transformer experiments instead use the full complement $\Pi_\perp = I_d - \Pi_\star$.

Compute. The cTRE experiments in Figures 1–2 are NumPy CPU runs. Rerunning the public scripts with Python 3.10.4 and NumPy 2.2.6 took 336.6 seconds (5.6 minutes) for E1 and 34.5 seconds (0.6 minutes) for E2 on an iMac18,3 workstation with a 3.4 GHz quad-core Intel Core i5-7500 CPU, 32 GB RAM, and macOS 13.7.8. The learned linear-transformer runs used an NVIDIA 2080 GPU. The updated logs report per-run times of 16.1–40.8 hours for $\sigma^2 = 0.04$ and 15.6–31.4 hours for $\sigma^2 = 0.0025$.

B.2 TWO-STAGE RIDGE ESTIMATOR DETAILS

Both cTRE and dTRE use the same two-stage structure. First, we compute the minimum-norm least-squares estimate $\widehat{w}^j$ in (8), and form the empirical covariance across tasks,

$$\widehat{C}_{M,\nu} = (M-1)^{-1} \sum_{j=1}^{M} (\widehat{w}^j - \bar{w})(\widehat{w}^j - \bar{w})^\top.$$

Second, the estimator takes a subspace $\widehat{U}$ from the leading eigenvectors of $\widehat{C}_{M,\nu}$ and performs ridge regression only inside this subspace:

$$\hat{w}_{\widehat{U},\rho}(X,Y) = \widehat{U}\big(\widehat{U}^\top X^\top X \widehat{U} + \rho I\big)^{-1} \widehat{U}^\top X^\top Y. \tag{100}$$

The estimators differ only in how $\widehat{U}$ and ρ are chosen.

Controlled TRE (cTRE). The cTRE uses the true target rank r_w and ridge ratio $\rho = \sigma^2/\lambda$. It takes $\widehat{U} = \widehat{U}_{r_w}$ from the leading r_w eigenvectors of $\widehat{C}_{M,\nu}$, then applies (18). Knowing the rank and ridge ratio lets us isolate the effect of subspace estimation. E1–E2 use this estimator to compare the observed risks with the population and finite-sample results.

Data-driven TRE (dTRE). The dTRE uses the same Stage 1 covariance but does not use the true r_w or ρ. Let $\hat{\eta}_1 \ge \cdots \ge \hat{\eta}_d$ be the eigenvalues of $\widehat{C}_{M,\nu}$. We estimate the isotropic part of the spectrum by the median of its smallest eigenvalues,

$$\widehat{\ell} = \mathrm{median}\{\hat{\eta}_{d-k+1}, \dots, \hat{\eta}_d\}, \qquad k = \max\{1, \lfloor d/2 \rfloor\},$$

and use the affine covariance constants (a, b, c) from Lemma 7 to form the following estimate of the noise variance,

$$\widehat{\sigma}^2 = \max\left\{10^{-6}, \frac{\widehat{\ell} - b\sum_{k=1}^{d} \hat{\eta}_k}{c}\right\}.$$

Set $\widehat{\ell}_{\mathrm{rank}} = \max\{\widehat{\ell}, c\widehat{\sigma}^2, 10^{-12}\}$. Using a threshold multiplier $m = 2$, we choose the rank

$$\widehat{r} = \min\left\{n-1, \max\left(1, \#\{k : \hat{\eta}_k > m\widehat{\ell}_{\mathrm{rank}}\}\right)\right\},$$

and $\widehat{U} = \widehat{U}_{\widehat{r}}$ is the leading $\widehat{r}$-dimensional eigenspace. To select the ridge parameter, we first construct a covariance estimate of rank $\widehat{r}$ by subtracting $\widehat{\ell}_{\mathrm{proxy}} = b\,\mathrm{tr}(\widehat{C}_{M,\nu}) + c\widehat{\sigma}^2$ on this subspace,

$$\widehat{\Sigma}_{\mathrm{proxy}} = \widehat{U}\left[\frac{1}{a}\left(\widehat{U}^\top \widehat{C}_{M,\nu} \widehat{U} - \widehat{\ell}_{\mathrm{proxy}} I_{\widehat{r}}\right)_+\right] \widehat{U}^\top,$$

Here $(\cdot)_+$ replaces every eigenvalue below 10^{-8} by 10^{-8}, as in the implementation. Finally, ρ is selected on the logarithmic grid $\{10^{-3}, \ldots, 10^{1}\}$ with 17 equally spaced points in log scale by minimizing squared coefficient error on 200 freshly simulated validation tasks sampled from $\widehat{\Sigma}_{\text{proxy}}$ with noise level $\widehat{\sigma}^2$. The dTRE estimate is then (100) with $(\widehat{U}, \widehat{\rho})$. In the reduced dTRE diagnostic, we use $(d, n, r_w, \lambda) = (20, 10, 3, 1)$, the same aligned and off-support source families as Appendix B.3, $M = 20{,}000$ pretraining tasks, 1,000 held-out target tasks, and $N_{\text{seed}} = 10$ independent seeds; reported dTRE summaries give seed means and standard deviations.

Figure 4 shows the estimated rank, ridge parameter, and noise variance for the dTRE runs in Figure 3. The estimated noise variance exceeds $\sigma^2 = 0.04$ in both source families, and the selected ridge parameter exceeds the target value $\sigma^2/\lambda = 0.04$. For aligned sources, all runs retain the correct rank $r_w = 3$, and the excess risk remains small. For off-support sources, the retained rank drops to 1 at $\nu = 0.2$, before the population threshold $\nu_c = 1/3$; the selected ridge parameter reaches the upper end of its search grid at $\nu = 0.3$. Thus the rank and ridge estimates can differ substantially from the target values even when the population eigenspace still equals the target subspace. The experiment does not isolate the contribution of each estimated quantity to the risk. Establishing conditions under which a data-driven choice retains the cTRE convergence behavior remains an open problem.

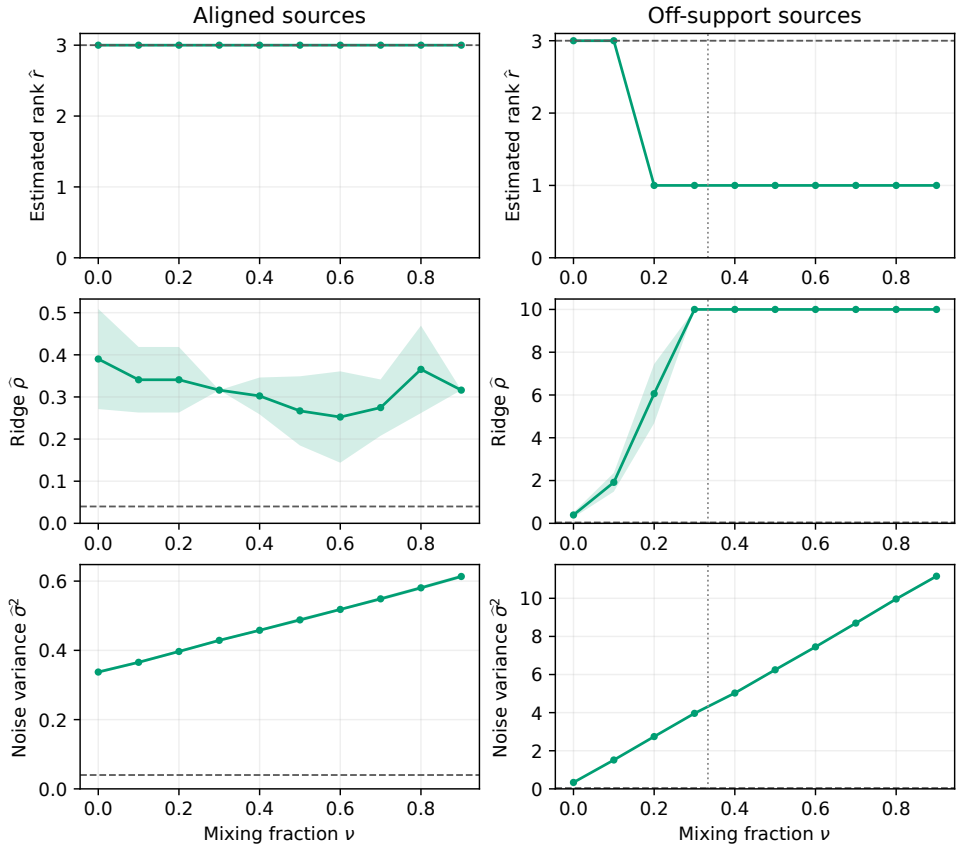

Figure 4: dTRE diagnostics for the setting at $\sigma^2 = 0.04$. The plot shows seed means and one-standard-deviation bands over $N_{\text{seed}} = 10$ runs for the estimated rank $\widehat{r}$, ridge $\widehat{\rho}$, and noise variance $\widehat{\sigma}^2$. Dashed lines mark the target values.

B.3 Linear-transformer model details

We consider a reduced dimension $(d, n, r_w) = (20, 10, 3)$ and $\lambda = 1$. The main transformer experiment uses $\sigma^2 = 0.04$ and tests the following source families: aligned sources $\Sigma_s = 2\Sigma_0$ ($\nu_c = 1$) and pure off-support sources $\Sigma_s = 2\lambda\Pi_\perp$ ($\nu_c = 1/3$). The tested mixing fractions are $\nu \in \{0, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Independent seed runs regenerate the

data; within each run the task pools remain fixed during optimization. For each (family, ν) cell, the learned model is trained on a $M = 50{,}000$-task pretraining set, with $\lceil (1-\nu)M \rceil$ target tasks and the remaining tasks from the source family. Validation is target-only, and the reported model is the checkpoint with the lowest target-validation squared loss. We evaluate on $1{,}000$ held-out target tasks and report both the identity-based inverse excess risk $\|\hat{w} - w_\star\|_2^2$ and an independent forward Monte Carlo check with 500 isotropic queries per target task.

We use a linear-attention model based on the nonlocal attention operator of Yu et al. (2024). Given a regression context (X, Y), it returns a coefficient estimate $\hat{w}_\theta(X, Y) \in \mathbb{R}^d$. For one task, its input is

$$Z^{(0)} = \mathrm{LN}_0\big([X^\top; Y^\top]\big) \in \mathbb{R}^{(d+1)\times n}.$$

Here $[X^\top; Y^\top]$ stacks the two arrays by rows, and LN denotes layer normalization along each row. The rows serve as tokens: row $i \le d$ is the vector of values taken by covariate i across the n context examples, and row $d+1$ is the response vector. The model therefore represents each variable by its n observed values in the context. It has no learned embedding dimension or feed-forward subnetwork.

Each hidden layer lets the variables interact through attention. For head h in layer ℓ, affine maps form queries and keys:

$$Q_h^{(\ell)} = Z^{(\ell)} W_{Q,h}^{(\ell)} + \mathbf{1}(b_{Q,h}^{(\ell)})^\top, \qquad K_h^{(\ell)} = Z^{(\ell)} W_{K,h}^{(\ell)} + \mathbf{1}(b_{K,h}^{(\ell)})^\top \in \mathbb{R}^{(d+1)\times d_k}.$$

Here $W_{Q,h}^{(\ell)}, W_{K,h}^{(\ell)} \in \mathbb{R}^{n\times d_k}$ are trainable matrices, $b_{Q,h}^{(\ell)}, b_{K,h}^{(\ell)} \in \mathbb{R}^{d_k}$ are trainable biases, and $\mathbf{1} \in \mathbb{R}^{d+1}$ is the vector of ones. The attention scores are used directly, without a softmax. Taking $Z^{(\ell)}$ itself as the value matrix, one head produces

$$A_h^{(\ell)}(Z^{(\ell)}) = Q_h^{(\ell)} \big(K_h^{(\ell)}\big)^\top Z^{(\ell)} / \sqrt{d_k} \in \mathbb{R}^{(d+1)\times n}.$$

The heads are summed, normalized, and added back through a residual connection,

$$Z^{(\ell+1)} = \mathrm{LN}_{\ell+1}\Big(\sum_h A_h^{(\ell)}(Z^{(\ell)})\Big) + Z^{(\ell)}, \qquad \ell = 0, \dots, L-2.$$

The last layer returns one coefficient for each covariate. Writing $Z_i^{(L-1)}$ for row i, it forms a shared query vector by a trainable linear combination of all $d+1$ query rows,

$$Q_{\mathrm{out}} = \left[\sum_{i=1}^{d+1} \alpha_i \left(Z_i^{(L-1)} W_Q^{(L-1)} + (b_Q^{(L-1)})^\top\right)\right]^\top \in \mathbb{R}^{d_k},$$

where α_i are trainable scalar weights. It also forms a key vector for each of the d covariate rows,

$$K_j = \big(Z_j^{(L-1)} W_K^{(L-1)} + (b_K^{(L-1)})^\top\big)^\top \in \mathbb{R}^{d_k}, \qquad j = 1, \dots, d.$$

The coefficient of covariate j is then

$$\big[\hat{w}_\theta(X, Y)\big]_j = Q_{\mathrm{out}}^\top K_j / \sqrt{d_k}, \qquad j = 1, \dots, d,$$

Contributions are summed over heads when multiple heads are used. A new query point is then predicted by the scalar rule $\hat{y}_{\mathrm{q}} = x_{\mathrm{q}}^\top \hat{w}_\theta(X, Y)$.

We train all parameters by Adam. In each simulation, optimization uses a fixed pretraining set and minimizes the empirical reconstruction loss

$$\mathcal{L}_M(\theta) := \frac{1}{M} \sum_{j=1}^{M} \big\| X^j \hat{w}_\theta(X^j, Y^j) - Y^j \big\|_2^2.$$

We select the checkpoint with the smallest loss on a separate target-only validation set.

For both reported noise grids, we use $L = 10$ layers including the readout layer, one head, and attention rank $d_k = 10$. Each run uses $E = 10000$ epochs, where one epoch is one full pass through the fixed $M = 50{,}000$-task pool, minibatch size $B \in \{64, 1024\}$, Adam learning rate 10^{-5}, and the default constant learning-rate schedule.

Repeated runs and retained records. For each source family and mixing fraction, we average 10 dTRE runs and 7–12 transformer runs at $\sigma^2 = 0.04$, or 10–12 transformer runs at $\sigma^2 = 0.0025$. The bands show sample standard deviations with denominator $N_{\text{seed}} - 1$. The cTRE reference is evaluated on the corresponding transformer task pools, while dTRE uses its separate 10-seed experiment. The inverse and independent forward Monte Carlo excess-risk estimates differ by at most 0.014 across the saved runs.

Figure 5 shows the excess Bayes risk at $\sigma^2 = 0.0025$, with the same estimator settings as Figure 3. The transformer again deteriorates before the cTRE threshold for off-support sources. The same qualitative patterns remain: aligned-source risk stays low, while both data-driven estimators deteriorate before the off-support threshold. The dTRE remains close to cTRE for aligned sources, but its off-support excess risk is already about 2.29 at $\nu = 0.2$.

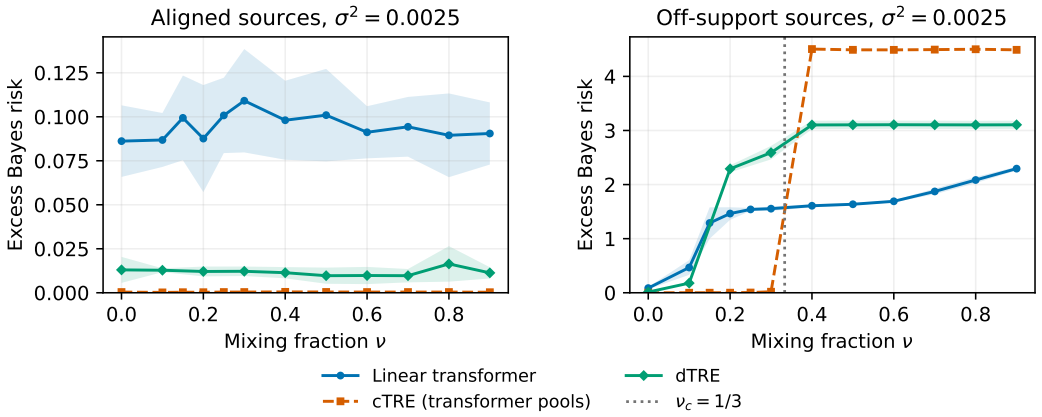

Figure 5: Excess Bayes risk at $\sigma^2 = 0.0025$, with the same estimator settings as Figure 3. Curves show means and bands show one sample standard deviation. The transformer again deteriorates before the cTRE threshold for off-support sources.