MINIMAX PRICE OF APPROXIMATE EQUIVARIANCE IN LINEAR REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariance is an important source of data efficiency in physical prediction and geometric perception. However, practical symmetries are often only approximate. Exact enforcement can remove predictive departures, while an unconstrained fit sacrifices parameter sharing and increases variance with limited labels. The challenge is to determine how strongly to enforce a known symmetry and what prediction error is unavoidable as the departure grows. In this paper, we formulate this choice as a minimax prediction problem in multivariate linear regression and propose projected shrinkage between exact equivariance and an unconstrained fit. We derive an exact finite-sample risk identity for this family and prove matching upper and lower bounds that characterize the global minimax prediction rate up to universal constants. The result shows that retention is governed by the balance between the possible departure and the variance required to estimate it. We extend this principle to anisotropic covariance, restricted high-dimensional designs, and sampled group averaging. Simulations recover the predicted retention transition, demonstrate the value of residual information in underdetermined designs, and connect projector accuracy to downstream prediction. These results provide a finite-sample answer to how much structure to share through an approximate symmetry when data are limited.

1 INTRODUCTION

Machine learning for physical prediction and geometric perception often has to learn from limited labeled data. One way to make these data more informative is to share information across inputs related by a known transformation. If a model predicts a direction from a scene, rotating the scene should rotate the predicted direction in the same way. This principle, called equivariance, lets related inputs provide shared evidence and helps a model learn from fewer labels (Cohen & Welling, 2016; Ravanbakhsh et al., 2017). In real applications, however, the relationship is often only approximate. Treating the symmetry as exact can discard a real directional effect and bias predictions that depend on it. Leaving the model unconstrained preserves that effect, but forces the learner to estimate relationships separately from the same limited labels, increasing variance (Figure 1). With limited labels, how much structure should the learner share through the symmetry, and how much deviation should it learn from data?

Existing work turns this sharing principle into model design. Exact equivariant models share parameters across transformed inputs (Cohen & Welling, 2016; Ravanbakhsh et al., 2017; Kondor & Trivedi, 2018). Theory shows that projection onto this shared structure can reduce estimation variance when the symmetry is correct (Elesedy & Zaidi, 2021). Recent methods soften the constraint, for example by learning partial equivariance or adding residual pathways (Romero & Lohit, 2022; Finzi et al., 2021; Wang et al., 2022; Huang et al., 2023; Petrache & Trivedi, 2023). These approaches let models preserve effects that a hard constraint would remove, yet they do not determine how much freedom to allow when labels are scarce. How much freedom should a learner allow for a specified symmetry, and what prediction error must even the best possible estimator incur as the permitted departure grows?

In this paper, we answer this question by viewing approximate equivariance in multivariate linear regression as a controlled estimation problem. We separate the coefficient into the part that follows the specified symmetry and a departure that may carry additional signal. Projection removes the

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

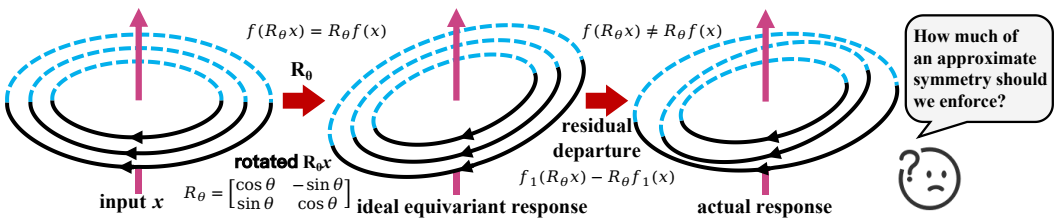

Figure 1: A rotation R_θ maps the input x to $R_\theta x$. Under exact equivariance, the response follows the same transformation, $f_0(R_\theta x) = R_\theta f_0(x)$. However, a fixed context that remains unchanged under the rotation can produce an actual response f_1 with $f_1(R_\theta x) \neq R_\theta f_1(x)$. The resulting residual $f_1(R_\theta x) - R_\theta f_1(x)$ represents the signal outside the equivariant component that the learner must decide whether to retain.

departure and enforces full sharing, and shrinkage retains it to an intermediate degree. These choices form a continuum between exact equivariance and an unconstrained fit, along which prediction trades bias against variance. This balance determines a minimax prediction rate, and a matching lower bound shows that its order cannot be improved. Together, these results give a finite-sample answer to how much approximate symmetry a learner should enforce when data are limited.

Our contributions are as follows:

(1) We formulate approximate equivariance as a minimax prediction problem for multivariate linear regression with a known symmetry representation. For the isotropic Gaussian model, we derive an exact finite-sample risk identity for the projected shrinkage family, construct a suitably tuned estimator that achieves the global minimax prediction rate up to universal constants, and prove a matching lower bound for every measurable estimator.

(2) We extend the analysis beyond isotropic covariance and a single symmetry component. The resulting rates capture how covariance geometry, the relative orientation of the symmetry subspace, and multiple representation blocks affect the cost of retaining directions that break the symmetry.

(3) We characterize practical boundary regimes in which full ordinary least squares is unavailable or the exact theoretical rule must be approximated. In particular, we analyze restricted Gram identifiability and leakage in high dimensions, together with finite group averaging and calibration of the noise level needed for implementation.

2 Preliminaries

To determine how much of a known symmetry a predictor should enforce, we first identify the coefficient matrices that respect the input and output transformations. Let G be a compact group with normalized Haar probability measure μ. The input and output spaces carry known continuous orthogonal representations $\phi : G \to O(p)$ and $\psi : G \to O(q)$, respectively. We represent a linear predictor by a matrix in $\mathcal{H} = \mathbb{R}^{q\times p}$ and equip this space with the Frobenius inner product $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$. For $f_A(x) = Ax$, the desired equivariance relation is $f_A(\phi(g)x) = \psi(g)f_A(x)$ for every $g \in G$. This relation defines the intertwiner subspace

$$\mathcal{E} = \{A \in \mathcal{H} : \psi(g)A = A\phi(g) \text{ for every } g \in G\}. \tag{1}$$

The symmetry constraint separates the coefficient space into directions that obey the relation and directions that can violate it. Let $P_\mathcal{E}$ denote the Frobenius orthogonal projector onto $\mathcal{E}$, and let $P_{\mathcal{E}^\perp} = I - P_\mathcal{E}$ denote its complement. We write $r = \dim(\mathcal{E})$ and $s = pq - r = \dim(\mathcal{E}^\perp)$, and assume $r, s \geq 1$ so that both types of directions are present. To obtain the projector from the group action, define $U_g(A) = \psi(g)A\phi(g)^{-1}$. Its fixed point space is $\mathcal{E}$, and Haar averaging gives

$$P_\mathcal{E} A = \int_G \psi(g)A\phi(g)^{-1}\, d\mu(g). \tag{2}$$

Thus $P_\mathcal{E}$ retains the coefficient component that obeys the specified symmetry, while $P_{\mathcal{E}^\perp}$ isolates directions that may carry symmetry departure. The question of how strongly to enforce the symmetry is therefore a question of how much of the latter component to remove.

We study this choice in multivariate linear regression with isotropic random design. Given n independent training pairs, the covariates and responses satisfy

$$x_i \sim N(0, I_p), \qquad y_i = \Theta_\star x_i + \xi_i, \qquad \xi_i \sim N(0, \sigma^2 I_q), \tag{3}$$

where the covariates and noise vectors are mutually independent. To represent an approximately equivariant coefficient, we bound only the component that can violate the symmetry:

$$\mathcal{T}(\rho) = \{\Theta \in \mathcal{H} : \|P_{\mathcal{E}^\perp}\Theta\|_F \le \rho\}, \qquad \rho > 0. \tag{4}$$

This parameter class is a cylinder. The component in $\mathcal{E}$ is unrestricted, while the component in $\mathcal{E}^\perp$ has radius ρ. The value of ρ controls the strength of the possible departure: $\rho = 0$ gives exact equivariance, and larger values permit larger departures.

To compare the statistical cost of these choices, we measure prediction error on an independent test covariate $x \sim N(0, I_p)$.

$$\mathcal{R}(\widehat{\Theta}, \Theta_\star) = \mathbb{E}\left\|(\widehat{\Theta} - \Theta_\star)x\right\|_2^2 = \mathbb{E}\left\|\widehat{\Theta} - \Theta_\star\right\|_F^2. \tag{5}$$

The expectation includes the training data and any independent randomness used by the estimator. Since $\mathbb{E}[xx^\top] = I_p$, prediction risk is equivalent to squared error in the matrix space. We therefore study the worst-case prediction risk over the approximate-equivariance class:

$$\mathfrak{M}_n(\rho) = \inf_{\widehat{\Theta}} \sup_{\Theta_\star \in \mathcal{T}(\rho)} \mathcal{R}(\widehat{\Theta}, \Theta_\star), \tag{6}$$

where the infimum ranges over all measurable estimators, including randomized procedures. For the core finite-sample comparison, we take $\sigma > 0$ and $\rho > 0$ so that both observation noise and possible symmetry departure are nondegenerate. We also assume $n \ge 2p + 2$. This condition makes the full ordinary least-squares estimator well-defined, ensures the inverse-Wishart moment used for its expected risk is finite, and gives the clean universal constant comparison used in the main rate bound.

3 A MINIMAX RULE FOR APPROXIMATE EQUIVARIANCE

3.1 PROJECTED SHRINKAGE AS A CONTINUUM OF SYMMETRY ENFORCEMENT

Let $X = [x_1, \ldots, x_n]$ and $Y = [y_1, \ldots, y_n]$. Under $n > p$, ordinary least squares provides the unconstrained reference estimate $\widetilde{\Theta} = YX^\top(XX^\top)^{-1}$. Its projections onto $\mathcal{E}$ and $\mathcal{E}^\perp$ have different statistical roles. The first preserves the component that satisfies the specified symmetry. The second contains the possible departure and can be retained with a controlled amount. This motivates the family indexed by a.

$$\widehat{\Theta}_a = P_{\mathcal{E}}\widetilde{\Theta} + aP_{\mathcal{E}^\perp}\widetilde{\Theta}, \qquad 0 \le a \le 1. \tag{7}$$

The endpoint $a = 0$ gives the hard equivariant projection, whereas $a = 1$ recovers ordinary least squares. Intermediate values retain part of the estimated component outside the equivariant subspace. The parameter a therefore measures how strongly the estimator trusts the specified symmetry in the bounded directions. The construction keeps the equivariant component unchanged, so all shrinkage acts on the part of the coefficient that the cylinder bounds.

3.2 EXACT RISK DECOMPOSITION AND ORACLE RETENTION

The isotropic design makes the noise in the OLS estimate isotropic after averaging over the design. Consequently, the expected noise energy of a fixed orthogonal projection in Frobenius geometry depends on its rank. This property allows the two components in Equation 7 to be analyzed separately.

Proposition 1 (Projected shrinkage risk identity)**.** *Under Equation 3, for every fixed $a \in [0, 1]$ and every $\Theta_\star \in \mathcal{H}$,*

$$\mathcal{R}(\widehat{\Theta}_a, \Theta_\star) = \frac{\sigma^2}{n - p - 1}(r + a^2 s) + (1 - a)^2 \|P_{\mathcal{E}^\perp}\Theta_\star\|_F^2. \tag{8}$$

The first term is the cost of estimating the unrestricted equivariant component together with the retained part of the departure. The second term is the squared bias from suppressing that departure. Maximizing over the cylinder replaces its squared norm by ρ^2. Writing $d = n-p-1$ and $v = \sigma^2 s/d$, the resulting worst case risk within the family is

$$\sup_{\Theta_\star \in \mathcal{T}(\rho)} \mathcal{R}(\widehat{\Theta}_a, \Theta_\star) = \frac{\sigma^2 r}{d} + a^2 v + (1-a)^2 \rho^2.$$

The two terms that depend on a form a quadratic balance. Its minimizer is

$$a_\star = \frac{\rho^2}{\rho^2 + v}, \qquad v = \frac{\sigma^2 s}{n-p-1}, \tag{9}$$

with worst case value

$$\frac{\sigma^2 r}{n-p-1} + \frac{\rho^2 v}{\rho^2 + v}.$$

The rule moves continuously between the two endpoints. When the allowed departure is small relative to the estimation variance, $a_\star$ is close to zero and projection is preferred. When the departure is large, $a_\star$ approaches one and the estimator retains the complementary component. The transition occurs when ρ^2 is on the scale of $\sigma^2 s/(n-p-1)$.

3.3 The minimax price of approximate equivariance

The preceding calculation gives the optimal retention level within the projected shrinkage family. The main question is whether an estimator outside this family can improve the order of the resulting risk. The answer is captured by the following upper and lower bound.

Theorem 2 (Minimax rate for the cylinder of approximate equivariance)**.** *Fix a compact group G with normalized Haar measure μ and continuous orthogonal representations $\phi : G \to O(p)$ and $\psi : G \to O(q)$. Suppose $1 \le r, s, \sigma, \rho > 0$, and $n \ge 2p+2$. Then*

$$\frac{2}{27}\left[\frac{\sigma^2 r}{n} + \min\left\{\rho^2, \frac{\sigma^2 s}{n}\right\}\right] \le \mathfrak{M}_n(\rho) \le \frac{n}{n-p-1}\left[\frac{\sigma^2 r}{n} + \min\left\{\rho^2, \frac{\sigma^2 s}{n}\right\}\right]. \tag{10}$$

The upper bound is achieved by $\widehat{\Theta}_{a_\star}$ from Equation 7 with $a_\star$ given in Equation 9.

The theorem separates the price of the two components of the cylinder. Estimating the unrestricted r-dimensional equivariant component costs order $\sigma^2 r/n$. The bounded complement contributes the smaller of its squared radius and the variance required to estimate s directions. Hence, the estimator retains the departure when the data can estimate it efficiently and suppresses it when the radius is the cheaper scale.

The upper bound follows from the exact family identity and the comparison $n/2 \le n-p-1 < n$. For the lower bound, choose orthonormal bases of $\mathcal{E}$ and $\mathcal{E}^\perp$ and place independent sign hypercubes on the two subspaces. Neighboring vertices differ in one coordinate, so their KL divergences remain bounded at the signal scales $\sigma/\sqrt{n}$ and $\min\{\rho/\sqrt{s}, \sigma/\sqrt{n}\}$. A testing argument then forces estimation error in both components. The product construction makes the two contributions add, yielding the same dependence on r, s, ρ, and n as the upper bound.

4 Geometry and computation beyond the isotropic regime

The same decision changes when prediction geometry is anisotropic, when full ordinary least squares is unavailable, or when the Haar projector must be approximated. Covariance determines which departure directions are worth retaining, restricted designs determine which components are identifiable, and approximate projectors add a separate implementation cost.

4.1 COVARIANCE GEOMETRY AND SELECTIVE RETENTION

Suppose that the covariates have known covariance $\Sigma \succ 0$. The original cylinder is still defined with the Frobenius norm, whereas prediction loss is measured in the covariance metric. To compare these two geometries, whiten the covariates with $T_\Sigma A = A\Sigma^{1/2}$ and let $\mathcal{L}_\Sigma = T_\Sigma(\mathcal{E})$. The bounded component is then measured in the quotient space orthogonal to $\mathcal{L}_\Sigma$. Its geometry is determined by $K_\Sigma = P_{\mathcal{L}_\Sigma^\perp} T_\Sigma\big|_{\mathcal{E}^\perp}$, whose singular values satisfy $\tau_1 \geq \cdots \geq \tau_s > 0$. For $U \in \mathcal{E}^\perp$,

$$\|K_\Sigma U\|_F^2 = \inf_{A \in \mathcal{E}} \|(U - A)\Sigma^{1/2}\|_F^2.$$

Thus K_Σ measures the prediction energy of a direction outside the equivariant subspace after the equivariant component has been removed. Set $b_j = \rho\tau_j$, $b_{s+1} = 0$, and

$$\Gamma_n(b_1, \ldots, b_s) = \min_{0 \leq k \leq s} \left\{ \frac{\sigma^2 k}{n} + b_{k+1}^2 \right\}.$$

Theorem 3 (Covariance sensitive minimax rate)**.** *Under $x_i \sim N(0, \Sigma)$, suppose $\Sigma \succ 0$, $1 \leq r, s$, $\rho, \sigma > 0$, and $n \geq 2p + 2$. Define*

$$\mathcal{R}_\Sigma(\widehat{\Theta}, \Theta) = \mathbb{E}\|(\widehat{\Theta} - \Theta)\Sigma^{1/2}\|_F^2.$$

Then

$$\frac{1}{27}\left[\frac{\sigma^2 r}{n} + \Gamma_n(b_1, \ldots, b_s)\right] \leq \inf_{\widehat{\Theta}} \sup_{\Theta \in \mathcal{T}(\rho)} \mathcal{R}_\Sigma(\widehat{\Theta}, \Theta) \leq \frac{n}{n - p - 1}\left[\frac{\sigma^2 r}{n} + \Gamma_n(b_1, \ldots, b_s)\right]. \tag{11}$$

The upper bound is achieved by whitening the OLS estimate, retaining the equivariant subspace and the first $k_\star$ quotient directions, where $k_\star$ minimizes Γ_n.

The isotropic formula is the special case $\Sigma = I_p$, for which every $\tau_j = 1$ and $\Gamma_n = \min\{\rho^2, \sigma^2 s/n\}$. Under anisotropic design, dimension alone no longer determines the price. Two intertwiner spaces with the same values of r and s can have different minimax risks because their orientations relative to Σ produce different quotient spectra. The scalar rule from Section 3 therefore becomes selective retention across quotient directions. Hard truncation gives the minimax order, while continuous spectral filtering refines the finite sample allocation within the diagonal spectral family.

The same principle applies when the complement is partitioned into known orthogonal blocks $\mathcal{E}_1, \ldots, \mathcal{E}_m$ with dimensions s_ℓ and radii ρ_ℓ. Each block receives its own coefficient

$$a_\ell = \frac{\rho_\ell^2}{\rho_\ell^2 + \sigma^2 s_\ell/(n - p - 1)},$$

and the minimax order becomes

$$\frac{\sigma^2 r}{n} + \sum_{\ell=1}^{m} \min\left\{\rho_\ell^2, \frac{\sigma^2 s_\ell}{n}\right\}.$$

Different representation blocks can therefore cross the retention threshold at different sample sizes and departure radii.

4.2 RESIDUALIZATION BEYOND FULL ORDINARY LEAST SQUARES

The estimator in Section 3 uses $(XX^\top)^{-1}$, so its construction does not apply when the full design is underdetermined. The restricted equivariant problem remains meaningful in this regime. Let $U_1, \ldots, U_r$ be a Frobenius orthonormal basis of $\mathcal{E}$ and define the restricted Gram matrix by

$$(G_X)_{jk} = \langle U_j X, U_k X \rangle_F.$$

The matrix G_X determines whether the unrestricted equivariant component is identifiable from the observed design. When it is invertible, define the map linking the two subspaces and its integrated leakage by

$$\mathsf{A}_X = G_X^{-1} \mathsf{D}_X^* \mathsf{C}_X, \qquad \Lambda_n = \|\mathbb{E}(\mathsf{A}_X^* \mathsf{A}_X)\|_{\mathrm{op}},$$

where D_X and C_X are the design maps restricted to $\mathcal{E}$ and $\mathcal{E}^\perp$.

If G_X is singular with positive probability or $\mathbb{E}\,\mathrm{tr}(G_X^{-1}) = +\infty$, the exact equivariant submodel already has infinite minimax risk. Otherwise, put $V_n = \sigma^2 \mathbb{E}\,\mathrm{tr}(G_X^{-1})$. When $\Lambda_n < \infty$, restricted least squares has the exact cylinder risk

$$\sup_{\Theta \in \mathcal{T}(\rho)} \mathcal{R}(\widehat{\Theta}_{\mathrm{RLS}}, \Theta) = V_n + \rho^2(1 + \Lambda_n). \tag{12}$$

The first term is the variance of estimating the equivariant component. The second term contains the bounded departure and the additional error caused by its leakage into the restricted fit.

Residualization makes the corresponding decision problem explicit. Let P_X project onto the restricted equivariant design image, and let $\widehat{U} \in \mathcal{E}^\perp$ depend on the residual data $(I - \mathsf{P}_X)\,\mathrm{vec}(Y)$. After correcting the equivariant coefficient with this residual estimate, the conditional risk is

$$\mathbb{E}\left[\|\widehat{\Theta} - \Theta\|_F^2 \mid X\right] = \sigma^2\,\mathrm{tr}(G_X^{-1}) + \mathbb{E}\left[\|\widehat{U} - U\|_F^2 + \|\mathsf{A}_X(\widehat{U} - U)\|_2^2 \mid X\right]. \tag{13}$$

The equivariant coefficient is removed from the residual experiment, while leakage changes the loss used to estimate the complement. The retention principle from Method I survives with this design dependent metric.

For scalar output, the leakage has a closed form. If $q = 1$, $1 \le r \le n-2$, and $p \ge n+2$, then

$$\Lambda_n = \frac{r}{n-r-1}, \qquad w_n = \frac{\sigma^2(n-r)}{p-n-1}.$$

Writing $s = p - r$, $m = n - r$, and $c_n = (m/s)\rho^2$, the residual min norm family uses $a_{\mathrm{res}} = \frac{c_n}{c_n + w_n}$. Its worst case risk is strictly smaller than the restricted least squares risk for every $\rho, \sigma > 0$. This scalar result isolates two high dimensional effects. A fraction of the departure can remain unidentifiable, while the residual design determines the variance of the estimable fraction.

4.3 Approximate projectors and calibration

For a continuous group, sample independent actions $g_1, \ldots, g_M \sim \mu$ and define

$$Q_M = \frac{1}{M}\sum_{t=1}^{M} U_{g_t}, \qquad \widehat{\Theta}_{a,M} = Q_M\widetilde{\Theta} + a(I - Q_M)\widetilde{\Theta}. \tag{14}$$

Each action fixes $\mathcal{E}$ pointwise. This anchor is required by the cylinder geometry because any approximation error along the unbounded component would create unbounded worst case bias.

Let $d = n - p - 1$ and $v = \sigma^2 s/d$. Averaging over the regression data and the sampled actions gives

$$\mathcal{R}(\widehat{\Theta}_{a,M}, \Theta) = \frac{\sigma^2}{d}\left[r + a^2 s + \frac{(1-a)^2 s}{M}\right] + (1-a)^2\left(1 + \frac{1}{M}\right)\|P_{\mathcal{E}^\perp}\Theta\|_F^2. \tag{15}$$

The excess caused by Haar sampling is at most v/M. For a realized sample, the operator event

$$M \ge \frac{2(1+\varepsilon/3)}{\varepsilon^2}\log\frac{2s}{\delta} \quad \Longrightarrow \quad \mathbb{P}\left(\|Q_M|_{\mathcal{E}^\perp}\|_{\mathrm{op}} \le \varepsilon\right) \ge 1 - \delta$$

converts an action budget into a high probability approximation guarantee.

The noise scale can be estimated from regression residuals. With

$$H_X = I_n - X^\top(XX^\top)^{-1}X, \qquad \widehat{\sigma}^2 = \frac{\|YH_X\|_F^2}{q(n-p)},$$

the residual estimator is independent of the OLS error and satisfies $q(n-p)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{q(n-p)}$. Replacing σ in $a_\star$ by $\widehat{\sigma}$ gives, for $n \ge 2p + 2$,

$$\sup_{\Theta \in \mathcal{T}(\rho)} \mathcal{R}(\widehat{\Theta}, \Theta) \le \frac{\sigma^2 r}{d} + \left(1 + \frac{1}{2(q(n-p)-2)}\right)\frac{\rho^2 v}{\rho^2 + v}.$$

The radius has a different statistical status. No estimator independent of ρ can attain the perpendicular minimax term within one fixed constant for every $\rho > 0$, even when the noise scale, representation, and exact projector are known. The computable rule can therefore calibrate the noise level from residuals, while simultaneous adaptation to every radius is ruled out by the parameter geometry.

Table 1: Coefficient risk on the core isotropic Gaussian benchmark. Entries report means ± 1.96 standard errors, and lower values are better. Bold entries mark the smallest mean at each departure radius. The family oracle uses the true (ρ, σ) and optimizes within the scalar projected shrinkage family.

Method	**Departure radius ρ**					
	0.12	0.20	0.32	0.50	0.80	1.20
Family oracle	0.117 ± 0.024	0.122 ± 0.017	0.152 ± 0.016	0.205 ± 0.022	0.279 ± 0.032	0.286 ± 0.034
Projection	0.119 ± 0.024	0.129 ± 0.017	0.183 ± 0.016	0.323 ± 0.015	0.730 ± 0.021	1.524 ± 0.023
OLS	0.381 ± 0.043	0.360 ± 0.028	0.339 ± 0.047	0.328 ± 0.043	0.366 ± 0.050	0.321 ± 0.041
Ridge	$\mathbf{0.062} \pm 0.007$	$\mathbf{0.071} \pm 0.007$	$\mathbf{0.088} \pm 0.011$	$\mathbf{0.149} \pm 0.017$	$\mathbf{0.233} \pm 0.019$	$\mathbf{0.264} \pm 0.035$
Soft	0.128 ± 0.026	0.130 ± 0.017	0.163 ± 0.017	0.221 ± 0.020	0.286 ± 0.032	0.291 ± 0.037
EB/JS	0.142 ± 0.029	0.145 ± 0.020	0.185 ± 0.026	0.226 ± 0.022	0.300 ± 0.035	0.295 ± 0.035
Validation	0.129 ± 0.026	0.131 ± 0.017	0.159 ± 0.016	0.220 ± 0.021	0.281 ± 0.032	0.291 ± 0.038

5 EXPERIMENTS

We use experiments to answer the following four questions: ***(i)*** Does prediction risk follow the predicted tradeoff between bias and variance as the departure radius changes? ***(ii)*** How accurately can the retention level be selected with estimated quantities and limited validation data? ***(iii)*** How do covariance estimation and restricted design identifiability affect retention? ***(iv)*** How does the number of sampled group actions affect approximation error and prediction on the cyclic shift task?

5.1 EXPERIMENTAL SETUP

We use four families of simulated benchmarks. The core isotropic Gaussian problem varies the allowed departure from equivariance and tests the predicted risk surface. The calibration suite estimates the noise level, limits the validation budget, and changes the design, noise, or specified equivariant subspace. The geometry suite studies covariance estimation and high dimensional regression, where retention depends on covariance orientation or restricted identifiability. The final suite approximates Haar averaging with sampled group actions and evaluates the resulting prediction cost on cyclic shift regression (Chen et al., 2020; Lyle et al., 2020).

The core comparisons include the scalar family oracle, exact projection, ordinary least squares, ridge, a soft equivariance penalty (Finzi et al., 2021), empirical Bayes shrinkage (James & Stein, 1961), and validation selected shrinkage (Stone, 1974). The high dimensional benchmarks additionally compare restricted least squares, full minimum norm estimation, and residual shrinkage. The cyclic shift task compares exact and sampled averaging with ridge, soft equivariance, and ordinary least squares. Hyperparameters are selected on validation data, and comparisons use paired random seeds. For more detailed descriptions of the experimental settings, please refer to Appendix D.

5.2 THE PREDICTED TRADEOFF BETWEEN BIAS AND VARIANCE

We first test whether the preferred strength of symmetry changes with the departure radius. On the core Gaussian benchmark, we vary ρ while keeping the sample size, noise level, and dimensions of the two coefficient subspaces fixed. At each radius, we evaluate the projected shrinkage family, its theoretical choice $a_\star$, exact projection, ordinary least squares, and the tuned baselines. Table 1 reports the empirical coefficient risks, and Figure 2 shows the corresponding retention surface and validation choices.

The risk surface in Figure 2a forms a valley that moves from small to large retention as ρ increases. The theoretical $a_\star$ follows this movement, linking the location of the valley to the bias and variance terms in the risk identity. Validation selected coefficients in Figure 2b show the same direction of change, with the largest spread in the intermediate regime. Several neighboring coefficients have similar risk there, so coefficient variation need not imply a large prediction penalty.

The table also shows that tuned ridge has the lowest pointwise risk for every simulated radius. The evaluated coefficient has a zero equivariant component, so shrinking this component improves

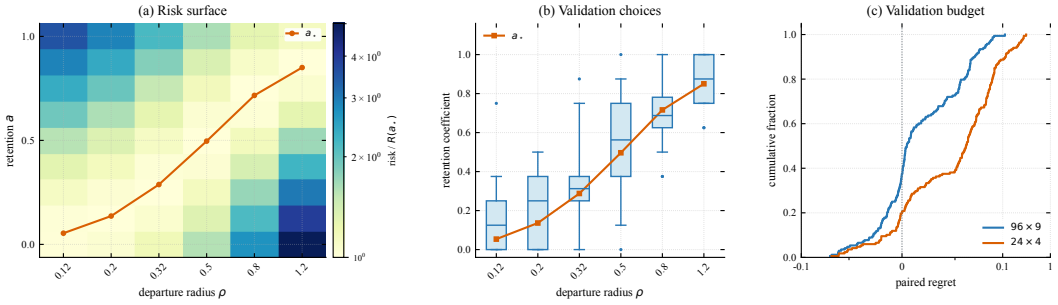

Figure 2: Core Gaussian risk and retention. (a) Risk relative to the scalar oracle over departure radius ρ and retention a; the orange curve marks $a_\star$. (b) Validation-selected retention with $a_\star$ overlaid. (c) Cumulative paired regret for full validation with 96 points and 9 candidates, and budgeted validation with 24 points and 4 candidates.

its pointwise estimate. The minimax class leaves that component unrestricted, which gives the projected shrinkage rule a different objective. This comparison separates performance at one coefficient from uniform protection over the full parameter class.

5.3 Data driven retention with estimated quantities

We next replace the inputs to the theoretical rule one at a time. The residual estimate of σ, perturbed radius values, a reduced validation budget, changes in the noise or design, and changes in the assumed equivariant dimension are evaluated on the same experimental family. Table 4 summarizes the resulting paired regret, while Figure 2c shows the pooled regret distributions for full and limited validation.

The residual estimate of σ produces a maximum mean regret of 0.00231 across the radius sweep. Radius perturbations have a larger effect, with the largest mean regret equal to 0.0309 for the smaller radius specification. The error is largest when the radius estimate changes the preferred side of the retention transition. A limited validation budget also moves selected coefficients toward the middle of the interval. Its maximum mean regret is 0.0692, and its largest 90% regret quantile is 0.250.

The distribution and subspace experiments give a directional interpretation of these changes. Under heteroskedastic noise, validation selected shrinkage and the soft penalty have negative regret relative to the Gaussian scalar reference, indicating lower pointwise risk for the simulated coefficient. The direction dependent noise variance changes the relative value of shrinking each component. Assuming $r = 4$ places two zero coefficient directions inside the shrinkable component and produces negative regret. Assuming $r = 8$ leaves two departure directions outside the shrinkable component and produces positive regret, reaching 0.0384 near $\rho = 0.32$. The dominant uncertainty is therefore the definition of the subspace on which retention acts.

5.4 Geometry and identifiability beyond full ordinary least squares

We then study covariance estimation and restricted identifiability beyond the core isotropic problem. The covariance experiment estimates a rotated covariance from independent unlabeled samples and records filter error together with a covariance diagnostic. The high dimensional experiment uses $p > n$ and compares restricted least squares, full minimum norm estimation, ridge, and residual shrinkage in scalar and two output regression. Figure 3 reports the high dimensional risk ratio relative to restricted least squares.

More unlabeled samples reduce the recorded covariance diagnostic from 1.90 to 0.62. The filter error varies from 0.0809 to 0.0909, so the diagnostic and pointwise coefficient error capture different aspects of covariance estimation. Residual shrinkage improves over restricted least squares as ρ grows, reaching risk reductions of 49.4% in the scalar experiment and 48.7% in the two output experiment. The value of residual information increases with the size of the departure.

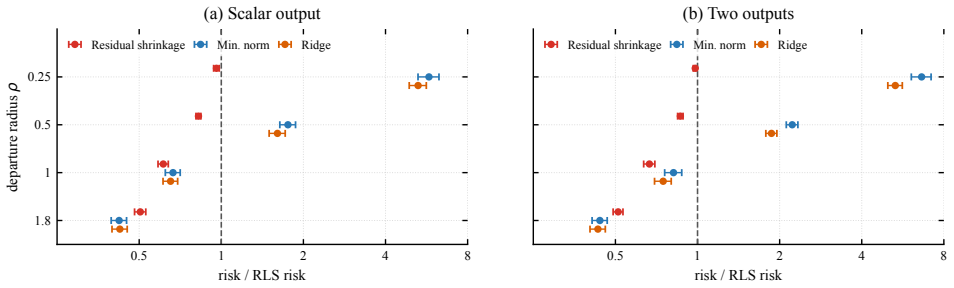

Figure 3: High-dimensional risk ratios relative to restricted least squares (RLS). The vertical line at one denotes equal risk with RLS, and horizontal bars show ± 1.96 standard errors. Panels (a) and (b) report scalar and two-output regression, respectively.

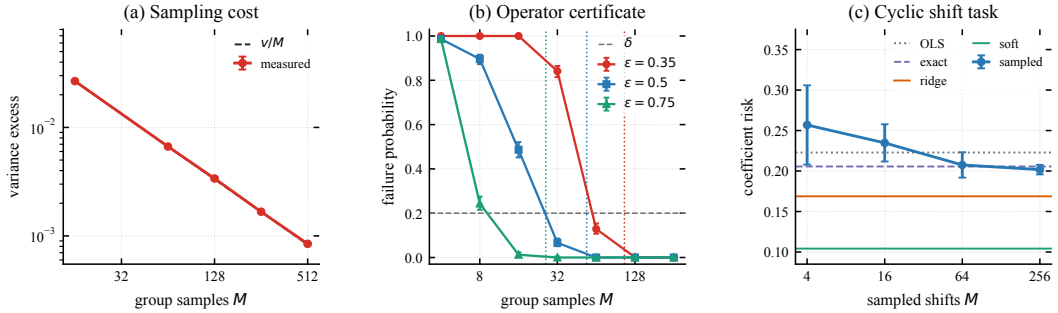

Figure 4: Approximate group averaging. (a) Variance excess compared with the predicted v/M scaling. (b) Operator failure probability; the horizontal line marks δ, and vertical dotted lines mark the theoretical sample thresholds. (c) Cyclic-shift coefficient risk versus the number of sampled shifts.

5.5 Approximate group averaging and downstream prediction

We finally connect projector accuracy to a downstream prediction task. In cyclic shift regression, the target coefficient contains a constant invariant component and a sinusoidal departure. We compare exact averaging, sampled averaging, ridge, soft equivariance, and ordinary least squares. The soft estimator reaches risk 0.104, compared with 0.206 for exact averaging and 0.223 for OLS. Sampled averaging decreases from 0.257 with four actions to 0.208 with 64 actions and 0.202 with 256 actions.

The controlled averaging audits explain this pattern. The measured variance excess follows the predicted v/M scale, and the operator failure rate decreases as M grows. Smaller tolerances require larger action budgets. Increasing M therefore improves the accuracy of the group projector, while the soft estimator retains the sinusoidal departure that exact averaging removes.

6 Conclusion

To determine how strongly a learner should enforce an approximate symmetry when labels are limited, we formulate multivariate linear regression with a known symmetry as a minimax estimation problem and introduce projected shrinkage between exact equivariance and an unconstrained fit. We derive an exact finite-sample risk identity for this family and prove matching upper and lower bounds that characterize the global minimax prediction rate up to universal constants. The result shows that the departure radius and its estimation variance jointly determine how much of the component outside the equivariant subspace should be retained. The experiments support this decision principle across the predicted crossover, high dimensional restricted designs, and sampled group averaging. These guarantees assume a fixed known representation, extending the theory beyond these conditions, which remain open.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing the writing. We take responsibility for the final content of this work.

REFERENCES

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, 2016.

David L Donoho, Richard C Liu, and Brenda MacGibbon. Minimax risk over hyperrectangles, and implications. *The Annals of Statistics*, pp. 1416–1437, 1990.

M.L. Eaton. *Group Invariance Applications in Statistics*. Regional conference series in probability and statistics. Institute of Mathematical Statistics, 1989.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969. PMLR, 2021.

Marc Finzi, Gregory Benton, and Andrew G. Wilson. Residual pathway priors for soft equivariance constraints. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30037–30049, 2021.

Gerald B Folland. *A course in abstract harmonic analysis*. CRC press, 2016.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2013.

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

Ningyuan Huang, Ron Levie, and Soledad Villar. Approximately equivariant graph networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Willard James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379. University of California Press, 1961.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2747–2755, 2018.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.

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

Omiros Papaspiliopoulos. High-dimensional probability: An introduction with applications in data science, 2020.

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mark Semenovich Pinsker. Optimal filtering of square-integrable signals in gaussian noise. *Problemy Peredachi Informatsii*, 16(2):52–68, 1980.

Siamak Ravanbakhsh, Jeff G. Schneider, and Barnabás Póczos. Equivariance through parameter-sharing. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2892–2901, 2017.

David W. Romero and Suhas Lohit. Learning partial equivariances from data. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36466–36478, 2022.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–147, 1974. doi: 10.1111/j.2517-6161.1974.tb00994.x.

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.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23078–23091, 2022.

A Extended Related Work

Exact equivariant representations. Exact equivariance constrains a linear map to an intertwiner subspace determined by the input and output group actions. Parameter sharing and group convolution implement this constraint for discrete and compact groups (Cohen & Welling, 2016; Ravanbakhsh et al., 2017; Kondor & Trivedi, 2018). The present paper starts after this representation-theoretic step. The group action and the intertwiner space are treated as known, and the statistical question is how much of the complementary component should be retained.

Statistical benefits of symmetry. Projection onto an exact equivariant space can reduce estimation variance when the symmetry is correct (Elesedy & Zaidi, 2021). Related work studies invariance through data augmentation, orbit averaging, and group-theoretic constructions (Lyle et al., 2020; Chen et al., 2020). These results establish gains from exact structure under their respective assumptions. Our parameter class allows a bounded component outside the equivariant space, so exact projection can introduce bias. The minimax problem therefore includes both the variance reduction from sharing and the cost of suppressing a possible departure.

Approximate symmetry. Partial equivariance, residual pathways, and approximately equivariant architectures relax a hard symmetry constraint (Romero & Lohit, 2022; Finzi et al., 2021; Wang et al., 2022; Huang et al., 2023). Broader approximation and generalization analyses also study the effect of imperfect group structure (Petrache & Trivedi, 2023). These works motivate treating symmetry strength as a design choice. The present analysis gives a finite-sample minimax formulation for that choice in multivariate linear regression, with an exact risk identity for the projected shrinkage family and a lower bound over all measurable estimators.

Minimax estimation. The isotropic parameter class becomes a product of an unrestricted subspace and a Euclidean ball. This geometry is related to classical minimax results for ellipsoids, hyperrectangles, and linear filters (Pinsker, 1980; Donoho et al., 1990). We use nonasymptotic testing arguments to obtain the lower bound (Tsybakov, 2009; Wainwright, 2019). The resulting rate is specific to the random-design regression experiment and the known equivariant subspace.

Covariance and high-dimensional regimes. Under nonisotropic design, prediction risk is measured in a covariance geometry and the relevant departure directions are determined by a quotient spectrum. Under underdetermined designs, the full OLS inverse is unavailable and the restricted Gram matrix controls identifiability. Standard inverse-Wishart calculations describe the isotropic OLS risk (Muirhead, 1982), while high-dimensional ridgeless regression shows why this formula

cannot be extended mechanically beyond the full-rank regime (Hastie et al., 2022). Our extensions track these effects through covariance-sensitive retention and leakage.

Finite group averaging. Haar averaging is a classical way to construct an orthogonal projector onto a fixed-point space (Folland, 2016; Eaton, 1989). A finite sample of group actions introduces an additional approximation term. Matrix concentration controls this term through operator-norm bounds for sums of random matrices (Tropp, 2012; Papaspiliopoulos, 2020). Our finite-Haar result includes an anchor on the unbounded equivariant component and translates the projector error into prediction risk.

B Extended Background

This section records the standard mathematical ingredients used by the proofs and places the paper's contribution relative to them. The core results below are specific to the random-design cylinder, even though each ingredient has a classical counterpart.

B.1 Compact Group Averaging

A compact group has a normalized Haar probability measure, which permits averaging a function or an operator over the group. For a representation on a finite-dimensional inner-product space, this average is the Reynolds operator. Compact harmonic analysis shows that it maps onto the fixed-point subspace and becomes an orthogonal projector when the action is unitary (Folland, 2016). In invariant statistical decision problems, averaging an equivariant procedure is a standard way to construct a procedure adapted to the group action (Eaton, 1989).

In the present setting, the induced action on coefficient matrices is

$$U_g(A) = \psi(g)A\phi(g)^{-1}.$$

The orthogonality of $\phi(g)$ and $\psi(g)$ makes U_g an isometry under the Frobenius inner product. Its fixed-point space is the intertwiner space $\mathcal{E}$, and the Haar average is therefore $P_{\mathcal{E}}$. Appendix C verifies these properties directly because the subsequent risk calculation uses both idempotence and self adjointness.

B.2 Gaussian Design Geometry

For the isotropic model, $S = XX^\top$ is Wishart. The inverse moment

$$\mathbb{E}[S^{-1}] = \frac{I_p}{n-p-1}$$

is the finite-sample fact that turns the OLS error into an isotropic random matrix after averaging over the design; standard multivariate treatments give this identity and its moment condition (Muirhead, 1982). Consequently, every fixed Frobenius-orthogonal projection pays a variance proportional to its rank.

When the design covariance is $\Sigma \succ 0$, prediction error is measured in the Σ metric. Whitening converts this metric to Euclidean geometry, and singular-value decompositions then identify the quotient directions that remain after removing the transformed equivariant subspace. These are standard linear-algebra operations (Golub & Van Loan, 2013); the covariance-sensitive minimax statement in Section 4.1 specifies how they interact with the original Frobenius cylinder.

B.3 Cylinder Geometry

The orthogonal decomposition gives the parameter class the product geometry

$$\mathcal{T}(\rho) \simeq \mathbb{R}^r \times B_2^s(\rho).$$

The first factor is unrestricted, while the second is bounded. Classical minimax theory studies related product, ellipsoidal, and hyperrectangular parameter spaces through linear filters and least favorable finite subsets (Pinsker, 1980; Donoho et al., 1990). The scalar rule in this paper is adapted to the two

subspaces induced by the known representation. It retains the unrestricted factor and shrinks only the bounded factor.

The lower bound uses a finite sign hypercube inside this product space. Neighboring vertices have bounded KL divergence, so a testing inequality limits the probability of recovering each sign. This is the usual reduction from estimation to testing in nonasymptotic minimax analysis (Tsybakov, 2009; Wainwright, 2019). The constants in Theorem 2 are derived for this regression experiment. They are not imported from an asymptotic sequence model.

The projected shrinkage family also relates to classical shrinkage under quadratic loss. The present coefficient a acts only on $\mathcal{E}^{\perp}$ and is selected from the radius and the projected OLS variance. The James and Stein estimator targets a different multivariate normal mean problem (James & Stein, 1961). This distinction is why the family identity and the global minimax bracket are stated separately.

B.4 High-Dimensional Estimation

When $p \geq n$, the full OLS inverse is unavailable. Minimum-norm and ridgeless estimators provide a useful reference for this regime and exhibit the interpolation behavior of high-dimensional least squares (Hastie et al., 2022). Our restricted-Gram and leakage quantities identify the additional cost created by the known equivariant subspace.

For a continuous group, the Haar projector can be approximated by an average over sampled actions. Matrix concentration gives operator-norm control for sums of centered random matrices (Tropp, 2012; Papaspiliopoulos, 2020). The finite-Haar results in Section 4.3 add the fixed-point anchor and translate this operator control into a prediction-risk term. These ingredients are standard; the combination with an unbounded equivariant component is the boundary addressed here.

C Proofs for the isotropic result

This appendix proves the projection identity, the exact risk formula for the scalar family, and the two sides of the main minimax rate. The arguments use the fixed representation only through the projector that is orthogonal under the Frobenius inner product.

C.1 Projection Operators

Lemma 4 (Haar averaging)**.** *The operator*

$$Q(A) = \int_G \psi(g) A \phi(g)^{-1} \, d\mu(g)$$

is the projector onto $\mathcal{E}$ that is orthogonal under the Frobenius inner product.

Proof. For $g \in G$, define $U_g(A) = \psi(g) A \phi(g)^{-1}$. Orthogonality of the two representations gives

$$\langle U_g(A), U_g(B) \rangle_F = \operatorname{tr}(A^{\top} B),$$

so U_g is an orthogonal operator on $\mathcal{H}$ and $U_g^* = U_{g^{-1}}$. The fixed point space of this action is $\mathcal{E}$ because $U_g(A) = A$ is equivalent to $\psi(g)A = A\phi(g)$.

For $h \in G$, left invariance of Haar measure gives

$$U_h Q(A) = \int_G U_{hg}(A) \, d\mu(g) = Q(A).$$

The image of Q lies in $\mathcal{E}$. If $A \in \mathcal{E}$, every integrand equals A, so $Q(A) = A$. Invariance under inversion gives

$$Q^* = \int_G U_g^* \, d\mu(g) = \int_G U_{g^{-1}} \, d\mu(g) = Q.$$

The operator Q is self adjoint, maps onto $\mathcal{E}$, and is the identity on $\mathcal{E}$. It follows that $Q^2 = Q = Q^*$, which proves $Q = P_{\mathcal{E}}$. □

Lemma 5 (Projected OLS noise). *Let $S = XX^\top$ and $W = \Xi X^\top S^{-1}$ under Equation 3. If $n > p + 1$, then for every deterministic projector P that is orthogonal under the Frobenius inner product and has rank k,*

$$\mathbb{E}\|PW\|_F^2 = \frac{\sigma^2 k}{n - p - 1}. \tag{16}$$

Proof. Under the isotropic Gaussian design, S has a Wishart distribution with n degrees of freedom and scale I_p. Its inverse moment is

$$\mathbb{E}[S^{-1}] = \frac{I_p}{n - p - 1}. \tag{17}$$

Conditionally on X, each row of W is centered with covariance $\sigma^2 S^{-1}$. Under vectorization by rows,

$$\mathbb{E}\big[\mathrm{vec}(W)\,\mathrm{vec}(W)^\top \mid X\big] = I_q \otimes \sigma^2 S^{-1}.$$

Taking expectation over X and applying Equation 17 yields

$$\mathbb{E}\big[\mathrm{vec}(W)\,\mathrm{vec}(W)^\top\big] = \frac{\sigma^2}{n - p - 1} I_{pq}.$$

Represent P by its $pq \times pq$ matrix under the same vectorization. Since $P = P^\top = P^2$ and $\mathrm{tr}(P) = k$,

$$\mathbb{E}\|PW\|_F^2 = \mathrm{tr}\big(P\,\mathbb{E}[\mathrm{vec}(W)\,\mathrm{vec}(W)^\top]P\big) = \frac{\sigma^2}{n - p - 1}\,\mathrm{tr}(P^2) = \frac{\sigma^2 k}{n - p - 1}.$$

□

C.2 Family Risk Upper Bound

Proof of Proposition 1. Substituting $Y = \Theta_\star X + \Xi$ into the OLS estimator gives $\widetilde{\Theta} = \Theta_\star + W$, where $W = \Xi X^\top (XX^\top)^{-1}$. Let $\Theta_\perp = P_{\mathcal{E}^\perp}\Theta_\star$. Equation 7 then gives

$$\widehat{\Theta}_a - \Theta_\star = P_{\mathcal{E}} W + a P_{\mathcal{E}^\perp} W - (1 - a)\Theta_\perp. \tag{18}$$

The first term is orthogonal to the other two. The cross term between $P_{\mathcal{E}^\perp} W$ and $\Theta_\perp$ has zero expectation because $\mathbb{E}[W] = 0$. Applying Lemma 5 at ranks r and s, followed by Equation 5, gives

$$\mathcal{R}(\widehat{\Theta}_a, \Theta_\star) = \frac{\sigma^2}{n - p - 1}(r + a^2 s) + (1 - a)^2 \|P_{\mathcal{E}^\perp}\Theta_\star\|_F^2.$$

For fixed a, the coefficient of the last term is nonnegative. Since $s \geq 1$, the supremum over $\mathcal{T}(\rho)$ is attained at a point with $\|P_{\mathcal{E}^\perp}\Theta_\star\|_F = \rho$. Writing $d = n - p - 1$ and $v = \sigma^2 s/d$, the worst case family risk is

$$\frac{\sigma^2 r}{d} + a^2 v + (1 - a)^2 \rho^2.$$

Completing the square gives

$$a^2 v + (1 - a)^2 \rho^2 = (v + \rho^2)\left(a - \frac{\rho^2}{v + \rho^2}\right)^2 + \frac{\rho^2 v}{\rho^2 + v}.$$

The minimizer is $a_\star = \rho^2/(\rho^2 + v)$, and substitution proves the proposition. □

Upper bound in Theorem 2. Use Proposition 1 with $a = a_\star$. The harmonic term satisfies

$$\frac{\rho^2 v}{\rho^2 + v} \leq \min\{\rho^2, v\}.$$

The condition $n \geq 2p + 2$ implies $d = n - p - 1 \geq n/2$, so $C_n = n/d$ satisfies $1 \leq C_n \leq 2$. Thus

$$\frac{\sigma^2 r}{d} = C_n \frac{\sigma^2 r}{n}, \qquad v = C_n \frac{\sigma^2 s}{n}.$$

For nonnegative u, t and $C_n \geq 1$, $\min\{u, C_n t\} \leq C_n \min\{u, t\}$. The preceding identities give

$$\sup_{\Theta_\star \in \mathcal{T}(\rho)} \mathcal{R}(\widehat{\Theta}_{a_\star}, \Theta_\star) \leq \frac{n}{n - p - 1}\left[\frac{\sigma^2 r}{n} + \min\left\{\rho^2, \frac{\sigma^2 s}{n}\right\}\right].$$

Taking the infimum over all measurable estimators proves the upper side of Equation 10. □

C.3 PRODUCT HYPERCUBE LOWER BOUND

Lemma 6 (KL divergence in the random design experiment)**.** *Let P_Θ denote the joint law of the training sample under parameter Θ. For every $\Theta, \Theta' \in \mathcal{H}$,*

$$\mathrm{KL}(P_\Theta \| P_{\Theta'}) = \frac{n}{2\sigma^2}\|\Theta - \Theta'\|_F^2. \tag{19}$$

Proof. The design marginal is the same under both parameters. Conditionally on x_i, the response laws are Gaussian with common covariance $\sigma^2 I_q$ and means Θx_i and $\Theta' x_i$. The chain rule for KL divergence gives

$$\mathrm{KL}(P_\Theta \| P_{\Theta'}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbb{E}\|(\Theta - \Theta')x_i\|_2^2.$$

Since $\mathbb{E}[x_i x_i^\top] = I_p$, this equals Equation 19. □

Lower bound in Theorem 2. Choose bases $U_1, \ldots, U_r$ of $\mathcal{E}$ and $V_1, \ldots, V_s$ of $\mathcal{E}^\perp$ that are orthonormal under the Frobenius inner product. Set

$$\alpha = \frac{2\sigma}{3\sqrt{n}}, \qquad \beta = \min\left\{\frac{\rho}{\sqrt{s}}, \frac{2\sigma}{3\sqrt{n}}\right\},$$

and define

$$\Theta_{\omega,\nu} = \alpha\sum_{j=1}^{r}\omega_j U_j + \beta\sum_{k=1}^{s}\nu_k V_k, \qquad (\omega,\nu) \in \{-1,1\}^r \times \{-1,1\}^s.$$

The perpendicular norm of every vertex is $\sqrt{s}\,\beta \le \rho$, so the hypercube lies in $\mathcal{T}(\rho)$.

If two vertices differ only in one ω_j coordinate, their difference is $2\alpha U_j$. Lemma 6 gives KL divergence $8/9$. A single ν_k flip has divergence $2n\beta^2/\sigma^2 \le 8/9$. Pinsker's inequality gives total variation at most $2/3$ for every neighboring pair. Under equal prior probabilities, every binary test has error at least

$$\frac{1 - 2/3}{2} = \frac{1}{6}.$$

Put the uniform prior on the hypercube. Given an arbitrary estimator $\widehat{\Theta}$, decode the signs by

$$\widehat{\omega}_j = \mathrm{sign}\langle\widehat{\Theta}, U_j\rangle_F, \qquad \widehat{\nu}_k = \mathrm{sign}\langle\widehat{\Theta}, V_k\rangle_F,$$

with a fixed convention at zero. Pairing neighboring vertices and averaging over nuisance signs yields

$$\mathbb{P}(\widehat{\omega}_j \ne \omega_j) \ge \frac{1}{6}, \qquad \mathbb{P}(\widehat{\nu}_k \ne \nu_k) \ge \frac{1}{6}.$$

Whenever a decoded sign is wrong, the coefficient error along the corresponding orthonormal direction is at least α or β. Bessel's inequality gives the Bayes risk lower bound

$$2^{-(r+s)}\sum_{\omega,\nu}\mathbb{E}_{\omega,\nu}\|\widehat{\Theta} - \Theta_{\omega,\nu}\|_F^2 \ge \frac{1}{6}(r\alpha^2 + s\beta^2).$$

The supremum over $\mathcal{T}(\rho)$ dominates this average. Substituting α and β gives

$$\mathfrak{M}_n(\rho) \ge \frac{1}{6}\left[\frac{4\sigma^2 r}{9n} + \min\left\{\rho^2, \frac{4\sigma^2 s}{9n}\right\}\right].$$

For nonnegative A, B, $\min\{A, 4B/9\} \ge (4/9)\min\{A, B\}$. This gives

$$\mathfrak{M}_n(\rho) \ge \frac{2}{27}\left[\frac{\sigma^2 r}{n} + \min\left\{\rho^2, \frac{\sigma^2 s}{n}\right\}\right],$$

which proves the lower side of Equation 10. The same argument applies to randomized estimators after appending their independent random seed to the observation. □

Table 2: Protocol for the reported experiments.

Block	Dimensions and grids	Repeats and accounting
Gaussian shrinkage	$n = 96, p = 24, r = 6, s = 18, q = 1$; $\rho \in \{0.12, 0.20, 0.32, 0.50, 0.80, 1.20\}$; $a \in \{0, 0.125, \dots, 1\}$	28 paired seeds and 36 trials per seed for the formula audit; ridge grid $\{0, 0.1, 1, 10, 100\}$; soft penalty grid $\{0, 0.25, 1, 4, 16\}$
Unknown calibration	Same Gaussian dimensions and grids; specified radius 0.7ρ or 1.4ρ	28 paired seeds; validation size 96; regret relative to the oracle with known (ρ, σ)
Robust calibration	Gaussian, t_5 noise, t_5 design, shell design, and heteroskedastic noise at $\rho = 0.50$	28 paired seeds; shared train and validation split across validation, ridge, soft, EB/JS, and OLS
Budgeted validation and wrong subspace	Same Gaussian dimensions; 24 validation points and four random a candidates; assumed $r \in \{4, 6, 8\}$	28 paired seeds; regret relative to the scalar oracle with the known subspace
Proof certificates	Hypercube $n = 128, r = 6, s = 18$; finite Haar $n = 128, p = 52, s = 32$; operator sweep $s = 32$, $M \in \{4, 8, \dots, 256\}$	600 trials for finite Haar and 800 operator trials; report failure rates, confidence intervals, and the downstream link
Covariance quotient	$p = 24, q = 3$ with rotated nondiagonal covariance; unlabeled $M \in \{48, 96, 192, 384\}$	24 seeds; effective rank 13.0; condition number 12.0; Loewner threshold 1.0
High dimension	Scalar $(q, p, n, r, s) = (1, 120, 80, 12, 108)$; two output case $(q, p, n, r, s) = (2, 120, 80, 12, 48)$; $\rho \in \{0.25, 0.50, 1.00, 1.80\}$	24 seeds and validation size 80; 96 Monte Carlo trials per (ρ, a) audit row; baselines are RLS, minimum norm, ridge, and residual shrinkage
Track D	Cyclic shifts of length 64; $n_{\mathrm{train}} = 96$, $n_{\mathrm{val}} = 96$; finite Haar $M \in \{4, 16, 64, 256\}$	28 paired seeds; record risk, runtime proxy, memory proxy, and action count

D Experimental Protocol

This appendix records the settings used for the experiments in Section 5. All tuning uses the training and validation portions only. Test risk is used for the final report and is not used to select a coefficient or a baseline hyperparameter.

The reported coefficient risk is the squared Frobenius error in the isotropic experiments and the corresponding covariance metric in the anisotropic experiment. Confidence intervals in the main text use paired seeds or paired bootstrap intervals as stated in the captions. The scalar family oracle uses the known (ρ, σ) only as a reference. Validation selection, residual calibration, and baseline tuning use quantities available from the training and validation data.

The protocol above specifies the dimensions, distributions, tuning grids, repeats, and evaluation rules needed to interpret the reported results. The experimental claims do not require private directories, private data locations, or internal records used to manage experiments.

E Boundary Conditions

Smallest admitted sample size. At $n = 2p + 2$, we have $d = n - p - 1 = p + 1 = n/2 > 0$. Thus $XX^\top$ is invertible almost surely, the inverse Wishart moment is finite, and the upper bound expressions in Theorem 2 are defined.

Nondegenerate subspaces. The assumptions $r \geq 1$ and $s \geq 1$ ensure that both orthogonal subspaces contain a nonzero direction. The unrestricted hypercube supplies the r term in the lower bound, while the perpendicular hypercube can reach the cylinder boundary of radius ρ.

Positive radius and noise. Theorem 2 assumes $\rho, \sigma > 0$. Under this condition, the hypercube amplitudes are positive and $a_\star \in (0, 1)$. The equality case $\rho^2 = \sigma^2 s/n$ and both sides of the crossover are included.

Table 3: Summary of the simulated benchmarks and selected quantitative checks.

Benchmark	Configuration	Repeats	Selected result
Gaussian regression	$q = 1, p = 24, r = 6, s = 18, n = 96$; six values of ρ	28 seeds, 36 trials per seed	Formula error 1.35% max; tuned regret 0.0152 max
Calibration	Residual $\widehat{\sigma}$, perturbed radius, and four candidates with 24 validation points	28 paired seeds	Budgeted q90 regret 0.250; plug-in σ regret 0.00231
Covariance geometry	$p = 24$, $q = 3$; rotated covariance; 48 to 384 unlabeled samples	24 seeds	Loewner gap reduction 67.5%; final pass rate 1.00
High dimensional regression	Scalar $(q, p, n, r, s) = (1, 120, 80, 12, 108)$; multi-output $(q, p, n, r) = (2, 120, 80, 12)$	24 seeds; 96 trials per audit row	Scalar shrinkage gain 49.4%; formula error 3.46% max
Group averaging and cyclic shift	Certificate sweep to $M = 256$; length 64 cyclic shift with $M \in \{4, 16, 64, 256\}$	600, 800, and 28 trials or seeds	Operator-linked excess reduction 98.4%; soft task risk reduction 53.2%

Table 4: Paired regret relative to the corresponding scalar oracle. Entries are absolute regret in units of 10^{-2} with ± 1.96 standard errors over 28 seeds.

Rule / ρ	0.12	0.2	0.32	0.5	0.8	1.2
Estimated noise and specified radius						
Residual $\widehat{\sigma}$	-0.00 ± 0.01	-0.00 ± 0.06	-0.04 ± 0.09	0.11 ± 0.17	0.23 ± 0.17	-0.07 ± 0.12
Radius 0.7ρ	-0.01 ± 0.03	0.10 ± 0.13	0.45 ± 0.33	1.14 ± 0.56	3.09 ± 0.84	1.91 ± 1.24
Radius 1.4ρ	0.11 ± 0.05	0.34 ± 0.24	0.80 ± 0.38	1.53 ± 0.63	0.71 ± 0.64	0.86 ± 0.83
Validation	0.86 ± 0.49	0.47 ± 0.45	0.67 ± 0.51	0.64 ± 0.48	0.34 ± 0.40	0.23 ± 1.30
Noise and design at $\rho = 0.50$						
Condition	Validation	Soft	Ridge	EB/JS		
Gaussian	1.90 ± 1.25	1.99 ± 1.20	-7.71 ± 2.89	2.22 ± 1.63		
t_5 noise	1.22 ± 0.90	1.94 ± 1.33	-6.23 ± 1.52	2.65 ± 1.30		
t_5 design	1.06 ± 0.54	1.52 ± 0.88	-5.92 ± 2.20	1.29 ± 0.50		
Shell	0.53 ± 0.53	0.61 ± 0.62	-8.20 ± 1.92	1.98 ± 0.95		
Heteroskedastic	-4.60 ± 1.98	-3.73 ± 2.36	-25.09 ± 4.11	1.04 ± 3.39		
Assumed equivariant dimension						
Assumed r / ρ	0.12	0.2	0.32	0.5	0.8	1.2
4	-2.09 ± 1.09	-2.38 ± 1.16	-2.38 ± 1.18	-0.52 ± 1.00	-0.61 ± 1.10	-0.20 ± 0.96
6	0.48 ± 0.27	0.68 ± 0.50	0.90 ± 0.68	1.21 ± 0.73	0.48 ± 0.89	0.21 ± 0.81
8	3.30 ± 1.12	2.49 ± 0.95	3.84 ± 1.85	2.93 ± 1.28	1.19 ± 1.09	1.22 ± 0.89

Unbounded equivariant component. The class $\mathcal{T}(\rho)$ is unbounded along $\mathcal{E}$. The OLS error is translation invariant in that component, so the upper risk does not depend on the magnitude of $P_{\mathcal{E}}\Theta_\star$. The lower bound uses only a finite subset of this unbounded component.

Representation dependence. The upper proof uses the representation only through an orthogonal projector of rank r. The lower proof uses arbitrary orthonormal bases of its range and orthogonal complement. No irreducible decomposition or coordinate alignment is required.

Endpoint interpretations. The projected endpoint $a = 0$ has worst case family risk $\sigma^2 r/d + \rho^2$. The unconstrained endpoint $a = 1$ has risk $\sigma^2(r + s)/d$. The scalar coefficient $a_\star$ interpolates between these endpoints and selects the lower of the two departure scales up to the theorem's universal factors.

Constants and theorem status. The constant $2/27$ comes from the amplitude $2/3$, the Pinsker testing error $1/6$, and the comparison factor $4/9$. The displayed family value in Proposition 1 is exact within the scalar family. Theorem 2 is a two sided rate statement over all measurable estimators, with universal lower and upper factors that depend on the sample size.