GRIT: OUT-OF-DISTRIBUTION ROBUSTNESS FROM FINITE AND NOISY INVARIANT PAIRS

Anonymous authors
Paper under double-blind review

ABSTRACT

Invariant pairs provide information about variation that should not affect prediction, but a finite, noisy collection can leave relevant directions unseen or suppress predictive information. We study out-of-distribution (OOD) robustness on fixed pretrained features, assuming that the label distribution given a linear projection of these features is invariant across environments. We propose Geometric Robustness Invariant Training (GRIT), which estimates a linear subspace from invariant pair differences and projects it out before predictor fitting, allowing the same constraint to accompany ERM or compatible robustness objectives. Our theoretical analysis bounds target risk through source prediction risk, a term related to covariate shift, and residual sensitivity along the unstable subspace. For GRIT, the bound depends on pair count, noise relative to the weakest clean signal, and selected rank. A parallel guarantee uses the fitted optimum to characterize quadratic prediction consistency under the same noise and shift assumptions. On ColoredMNIST-INV and Waterbirds-INV, reported means improve over each of eight evaluated objectives and exceed prediction consistency in 15 of 16 settings. On UrbanCars, combining pairs with background or object changes improves robustness to two spurious features.

1 INTRODUCTION

Spurious correlations challenge out-of-distribution (OOD) robustness: signals predictive during training may become unreliable after deployment, as when image backgrounds predict object class. We study predictors fitted on fixed pretrained features: these features support practical linear probing (Radford et al., 2021), but the fitted predictors may still exploit spurious correlations (Kirichenko et al., 2023). Learning from paired observations has driven advances in representation learning from augmented views (SimCLR, Chen et al., 2020; BYOL, Grill et al., 2020), image–text pairs (CLIP, Radford et al., 2021), and related views or regions (I-JEPA, Assran et al., 2023; LeJEPA, Balestriero & LeCun, 2025).

For OOD robustness, consistency regularization uses pairs to penalize differences in predictions, representations, or feature attributions. CoRe penalizes prediction or loss variation within groups sharing an identity and label (Heinze-Deml & Meinshausen, 2021). MatchDG learns cross-domain matches and regularizes their representations (Mahajan et al., 2021). LAM aligns feature contributions to class logits across pairs that share semantic content (Gao et al., 2024). IPG uses pairs supplied to the method in corrective updates that align class-dependent feature attributions and separate classes (Surner et al., 2026). We call pairs intended to preserve task-relevant content while varying factors that should not affect prediction **invariant pairs**. Examples include the same bird against different backgrounds, cars of the same model photographed from different angles, or similar tissue samples from a patient imaged using different microscopes. Such pairs can arise through augmentations, multiple views, matching, or targeted interventions. *Our motivating hypothesis is that invariant pairs can do for OOD robustness what labeled examples do for supervised learning: provide a common interface that decouples application-specific knowledge from application-agnostic learning.*

Despite these advances, a central gap remains in characterizing which invariances can be reliably enforced from **finite, noisy invariant pairs**. A finite pair collection may fail to capture desired invariances, while pair noise may induce undesired invariances that discard information that remains

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.

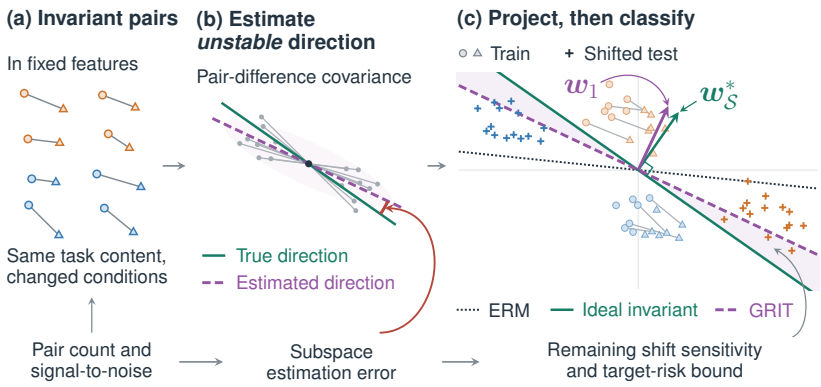

Figure 1: **GRIT uses noisy invariant pairs to approach the ideal invariant classifier.** **(a)** Eight pairs connect two training domains (shapes); colors denote classes. **(b)** Their difference covariance estimates an unstable direction; the red arc shows estimation error. **(c)** Projecting before fitting improves on ERM here, but GRIT retains sensitivity to shifted test points (+). Arrows show classifier normals; shading marks disagreement with the ideal boundary. Pair count and signal-to-noise affect estimation error; the target-risk bound also depends on shift magnitude and predictor norm.

predictive across environments. For consistency regularization, this means that penalizing differences between pairs can also suppress features that remain predictive across environments. Addressing this gap requires connecting pair count, noise, and geometric coverage to both the choice of invariance constraint and its effect on target risk. We ask: **How can we learn the right invariances from finite, noisy invariant pairs, and when are they sufficient for OOD robustness?**

We propose Geometric Robustness Invariant Training (GRIT) and characterize when finite, noisy invariant pairs suffice for OOD robustness. GRIT estimates a linear subspace from pair differences and projects it out before predictor fitting (Figure 1). This constraint can accompany ERM or another learning objective. For linear probing, our bounds separate source risk, changes in stable-feature distributions (bridge covariate shift), and sensitivity to unstable variation. GRIT's residual sensitivity depends on pair count, noise relative to the weakest clean signal, and rank. The paper makes three contributions:

1. **A geometric method for finite, noisy invariant pairs.** GRIT estimates a subspace from the leading eigenvectors of the pair covariance and projects features onto its orthogonal complement before predictor fitting. The selected rank controls which directions are removed, separating subspace estimation from downstream learning and allowing the same constraint to accompany different learning objectives.
2. **Theoretical conditions for OOD robustness from finite, noisy invariant pairs.** Under stated noise and distribution-shift assumptions, we give high-probability target-risk bounds for GRIT and quadratic prediction consistency. The GRIT bound depends on pair count, noise relative to the weakest clean signal, and the cost of removing extra directions. Fitted-predictor examples match its leading pair-count and noise dependence at the correct rank.
3. **Empirical improvements in OOD robustness with invariant pairs.** Because GRIT projects fixed features before predictor fitting, it can serve as a preprocessing step for ERM or compatible domain generalization methods. On ColoredMNIST-INV and Waterbirds-INV, GRIT improves mean performance with each of eight learning objectives, including ERM and four-group GroupDRO. Its mean exceeds prediction consistency in 15 of 16 settings, with the exception occurring on Waterbirds-INV. On UrbanCars, combining pairs with background or object changes improves robustness to two spurious features.

2 Problem Setup: Predictive Stability and Pair Information

To connect invariant pairs to prediction, we formalize predictive stability in a fixed representation. Let $\boldsymbol{z} = \phi(\boldsymbol{x}) \in \mathbb{R}^d$, where $\phi : \mathbb{X} \to \mathbb{R}^d$ is a fixed feature extractor. Let p denote the source distribution and q denote an unseen target distribution of $(\mathbf{z}, y)$. For a subspace $\mathcal{T}$, let $P_{\mathcal{T}}$ be its orthogonal projector and $\boldsymbol{z}_{\mathcal{T}} = P_{\mathcal{T}}\boldsymbol{z}$.

Assumption 1 (A1: Predictive stability)**.** *There is a subspace $\mathcal{S} \subseteq \mathbb{R}^d$ and a conditional distribution P_{inv} such that, in every source or target environment e,*

$$P^{(e)}(y \mid \boldsymbol{z}_{\mathcal{S}}) = P_{\text{inv}}(y \mid \boldsymbol{z}_{\mathcal{S}}) \quad P^{(e)}\text{-almost surely.} \tag{1}$$

We call $\mathcal{S}$ a *predictively stable subspace* and $\mathcal{U} = \mathcal{S}^{\perp}$ its *unstable complement*. Although the label distribution given the stable features is unchanged, the distribution of those features can vary across environments. Unstable features can also help predict labels in the source environment even when their relationship to labels changes at deployment. Removing them may therefore reduce source accuracy; predictive stability alone does not eliminate this tradeoff. A1 extends feature-subset conditional-invariance models to linear projections (Rojas-Carulla et al., 2018; Peters et al., 2016); full-feature covariate shift is the case $\mathcal{S} = \mathbb{R}^d$ (Sugiyama et al., 2007). Appendix E details the distinction from IRM.

Two sources of information. The learner receives labeled inputs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and invariant pairs $\{(\boldsymbol{x}_j, \boldsymbol{x}'_j)\}_{j=1}^k$, with all inputs in $\mathbb{X}$. Labels inform prediction, while differences $\boldsymbol{\delta}_j = \phi(\boldsymbol{x}_j) - \phi(\boldsymbol{x}'_j)$ inform the unstable geometry. To estimate the unstable subspace, a pair collection needs *coverage* of relevant unstable directions and *precision* to avoid removing stable predictive directions and unnecessarily reducing predictive performance. These requirements concern the collection as a whole. For coverage, many invariant pairs may repeat one variation and leave others unseen, whereas pooling complementary pair types can cover several variations that prediction should ignore. For precision, individual differences need not lie entirely in $\mathcal{U}$: the aggregate covariance can still distinguish unstable variation from residual variation outside that subspace.

These collection properties apply whether pairs are designed or sampled. Repeated measurements, multiple views, augmentations, and targeted interventions can supply candidate invariant pairs, provided the chosen relation preserves task-relevant information. Appendix E gives sufficient constructions and their limits. We study learning from these collections once they are supplied.

3 GRIT: Estimating and Removing Unstable Directions

If $\mathcal{U}$ were known, invariance to its directions would require a linear score $f_{\boldsymbol{w}}(\boldsymbol{z}) = \boldsymbol{w}^\top \boldsymbol{z}$ to satisfy $P_{\mathcal{U}}\boldsymbol{w} = 0$. We could impose this constraint by fitting on $P_{\mathcal{S}}\boldsymbol{z}$. GRIT substitutes an estimate of $\mathcal{U}$ obtained from the pairs into this construction. The constraint defines the chosen invariant class; its predictive value still depends on the information preserved by $\mathcal{S}$.

Estimate, select, project, fit. Define the *pair covariance* by giving each observed difference and its negative equal weight. The symmetrization is detailed in Appendix F.2. With eigenvalues in descending order and orthonormal eigenvectors $\widehat{\boldsymbol{u}}_a$, compute

$$\widetilde{\boldsymbol{\Sigma}}_k = \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{\delta}_j \boldsymbol{\delta}_j^\top, \qquad P_{k,r} = \sum_{a=1}^{r} \widehat{\boldsymbol{u}}_a \widehat{\boldsymbol{u}}_a^\top, \qquad \boldsymbol{z}^{\text{GRIT}} = (\boldsymbol{I} - P_{k,r})\boldsymbol{z}. \tag{2}$$

Define $P_{k,r,0} = \boldsymbol{I} - P_{k,r}$. The rank r counts removed directions; the predictor retains $d - r$ dimensions. Fit the chosen objective on projected features, using validation data to select r and training hyperparameters. For empirical risk minimization (ERM), this is equivalent to minimizing the empirical loss subject to $P_{k,r}\boldsymbol{w} = 0$. The same preprocessing can accompany objectives that use training environments or groups. Appendix F.2 gives the computation details.

Projection gives exact invariance to $\widehat{\mathcal{U}}_r = \text{range}(P_{k,r})$: every head receiving only projected features has the same output at $\boldsymbol{z}$ and $\boldsymbol{z} + \boldsymbol{u}$ for $\boldsymbol{u} \in \widehat{\mathcal{U}}_r$. Appendix B.2 gives a minimax justification under Gaussian pair noise. Projection removes differences along the estimated subspace, but members of a noisy invariant pair can still differ in the retained directions. Section 4 analyzes when the estimated subspace captures the unstable variation and how estimation error affects target risk.

Relationship to consistency. Quadratic prediction consistency penalizes differences in the scalar linear scores:

$$\frac{\lambda}{k}\sum_{j=1}^{k} (\boldsymbol{w}^\top \boldsymbol{\delta}_j)^2 = \lambda \boldsymbol{w}^\top \widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{w} = \lambda \sum_{a=1}^{d} \widetilde{\lambda}_a (\widehat{\boldsymbol{u}}_a^\top \boldsymbol{w})^2, \qquad \lambda > 0. \tag{3}$$

For logistic prediction these are logit differences, not probability differences. This follows the same consistency principle as penalties in CoRe, MatchDG, and LAM (Heinze-Deml & Meinshausen, 2021; Mahajan et al., 2021; Gao et al., 2024). Squared differences between view embeddings also appear in LeJEPA's prediction objective (Balestriero & LeCun, 2025). The key difference from GRIT is that quadratic consistency retains eigenvalue magnitudes as penalty weights, whereas GRIT selects a subspace and enforces invariance to it exactly.

Proposition 1 (Covariance changes preserving the GRIT constraint)**.** *Let $C, C' \succeq 0$ be pair covariance matrices with strict gaps between their rth and $(r+1)$st eigenvalues, for fixed $1 \le r < d$. Let $\mathcal{V}$ be the leading r-dimensional eigenspace of C. The rank-r GRIT constraints from C and C' are identical if and only if C' does not mix $\mathcal{V}$ with $\mathcal{V}^\perp$ and its eigenvalues within $\mathcal{V}$ remain larger than those within $\mathcal{V}^\perp$; equivalently,*

$$P_{\mathcal{V}} C' P_{\mathcal{V}^\perp} = 0, \qquad \lambda_{\min}(C'|_{\mathcal{V}}) > \lambda_{\max}(C'|_{\mathcal{V}^\perp}).$$

For fixed $\lambda > 0$, the consistency penalties $\lambda \boldsymbol{w}^\top C \boldsymbol{w}$ and $\lambda \boldsymbol{w}^\top C' \boldsymbol{w}$ are identical for every $\boldsymbol{w}$ if and only if $C = C'$.

Thus GRIT tolerates changes to eigenvalues and rotations of eigenvectors within either the removed or retained subspace, provided the two subspaces do not mix and their eigenvalues remain separated. For example, $\mathrm{diag}(4, 1)$ and $\mathrm{diag}(9, 1)$ both give the rank-one GRIT constraint $w_1 = 0$, but assign different consistency penalties to that coordinate. This insensitivity to covariance changes is one possible explanation for GRIT's empirical advantage over prediction consistency. Appendix A.2 proves the proposition and further characterizes the allowed perturbations. Section 4 connects subspace recovery to target risk.

4 Theoretical Analysis

We analyze binary logistic risk $R_p(\boldsymbol{w}) = \mathbb{E}_p \log(1 + e^{-y \boldsymbol{w}^\top \mathbf{z}})$, with $y \in \{-1, 1\}$; mean squared error analogues appear in Appendix A.7.

4.1 What invariance controls under distribution shift

Assumption 1 fixes the conditional label distribution given $\boldsymbol{z}_\mathcal{S}$, but allows the test distribution to change in two ways: the stable marginal can change from $p(\boldsymbol{z}_\mathcal{S})$ to $q(\boldsymbol{z}_\mathcal{S})$, and the unstable conditional can change from $p(\boldsymbol{z}_\mathcal{U} \mid \boldsymbol{z}_\mathcal{S}, y)$ to $q(\boldsymbol{z}_\mathcal{U} \mid \boldsymbol{z}_\mathcal{S}, y)$. Our focus is the second change, which invariance learned from pairs can control. We introduce a bridge distribution to isolate it while accounting separately for the change in the stable marginal.

Define $\breve{q}(\boldsymbol{z}_\mathcal{S}, y, \boldsymbol{z}_\mathcal{U}) = p(\boldsymbol{z}_\mathcal{S}) p(y \mid \boldsymbol{z}_\mathcal{S}) q(\boldsymbol{z}_\mathcal{U} \mid \boldsymbol{z}_\mathcal{S}, y)$. This bridge keeps the source stable features and labels jointly distributed as before, but replaces the unstable conditional with that of the target. Thus $p \to \breve{q}$ isolates the unstable change, while $\breve{q} \to q$ changes only the stable marginal under A1. We account for the latter through the *bridge covariate shift term* $C_{\mathrm{br}}(\boldsymbol{w}) = |R_q(\boldsymbol{w}) - R_{\breve{q}}(\boldsymbol{w})|$. We fix a measurable version of the target conditional on the source stable-label support and assume that $p, q, \breve{q}$ have finite second feature moments. Let Π_{cond} contain couplings of p and $\breve{q}$ that share $(\mathbf{z}_\mathcal{S}, y)$, and let $\Delta = \mathbf{z}' - \mathbf{z} \in \mathcal{U}$ under such a coupling.

Theorem 1 (Risk decomposition)**.** *For every $\boldsymbol{w} \in \mathbb{R}^d$,*

$$R_q(\boldsymbol{w}) \le R_p(\boldsymbol{w}) + C_{\mathrm{br}}(\boldsymbol{w}) + \inf_{\pi \in \Pi_{\mathrm{cond}}} \mathbb{E}_\pi |\boldsymbol{w}^\top \Delta|. \tag{4}$$

If a conditional coupling satisfies $\mathbb{E}_\pi[\Delta\Delta^\top] \preceq \rho_\Delta^2 P_\mathcal{U}$ for $\rho_\Delta \ge 0$, the last term is at most $\rho_\Delta \|P_\mathcal{U} \boldsymbol{w}\|_2$.

The three terms are source prediction risk, bridge covariate shift, and direct unstable shift. Stable predictors eliminate the last term, while changes to the stable marginal can still change their risk. Appendix A.1.1 explains when a common stable Bayes rule remains optimal, and Appendix A.1 bounds the exact bridge term by transport. Appendix C.1 sharpens the direct-shift bound for learned logistic predictors.

The direct-shift bound separates deployment-shift size ρ_Δ from predictor sensitivity $\|P_\mathcal{U} \boldsymbol{w}\|_2$. By estimating which directions a predictor should ignore, pairs can help reduce this sensitivity. They need not reproduce the target distribution. Here $\mathcal{U}$ is the unstable complement from Assumption 1; the bound on shift size additionally uses the conditional-displacement condition in Theorem 1.

4.2 When does a collection reveal the unstable subspace?

The coverage and precision requirements from Section 2 can be expressed through the pair covariance. The following result characterizes when its leading eigenspace identifies $\mathcal{U}$, even if individual differences also contain components in $\mathcal{S}$.

Proposition 2 (Exact spectral recovery). *Let $1 \le r_* = \dim(\mathcal{U}) < d$ and $\boldsymbol{M} \succeq 0$. Its unique leading r_*-dimensional eigenspace is $\mathcal{U}$ if and only if $P_{\mathcal{U}} \boldsymbol{M} P_{\mathcal{S}} = 0$ and $\lambda_{\min}(\boldsymbol{M}|_{\mathcal{U}}) > \lambda_{\max}(\boldsymbol{M}|_{\mathcal{S}})$.*

The first condition prevents the covariance from mixing stable and unstable directions. The second requires its smallest eigenvalue within $\mathcal{U}$ to exceed its largest within $\mathcal{S}$, so all unstable directions are selected first. Equality at this boundary would make the selected subspace nonunique. These conditions characterize leading-eigenspace selection, rather than every possible estimator.

For the remaining pair guarantees, fix $1 \le r_* := \dim(\mathcal{U}) < d$.

Assumption 2 (A2: conditional pair noise). *Conditional on designed signals $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$ spanning $\mathcal{U}$, the observed differences satisfy $\boldsymbol{\delta}_j = \boldsymbol{u}_j + \boldsymbol{\xi}_j$, where $\boldsymbol{u}_j \in \mathcal{U}$. The noises are independent, mean zero, have covariance $\sigma_\xi^2 \boldsymbol{I}$, and satisfy $\sup_{\|\boldsymbol{v}\|_2=1} \|\boldsymbol{v}^\top \boldsymbol{\xi}_j\|_{\psi_2} \le K\sigma_\xi$ for $\sigma_\xi \ge 0$ and fixed $K \ge 1$.*

Define $\boldsymbol{\Sigma}_k = k^{-1} \sum_j \boldsymbol{u}_j \boldsymbol{u}_j^\top$, with positive eigenvalues $\lambda_{1,k} \ge \cdots \ge \lambda_{r_*,k} > 0$. Its noisy conditional mean is $\boldsymbol{\Sigma}_k + \sigma_\xi^2 \boldsymbol{I}$, which recovers $\mathcal{U}$ by Proposition 2. Sampling fluctuation can rotate this eigenspace. The next theorem controls its prediction cost; Appendix A.3 also states a consistency criterion along growing signal collections.

4.3 Finite-pair guarantee for GRIT

We now connect recovery to prediction. Define $g_k = \lambda_{r_*,k}$ for the weakest clean signal, and let $L_z = \|\mathbb{E}_p[\mathbf{z}\mathbf{z}^\top]\|_{\mathrm{op}}^{1/2}$. Assume $\mathcal{W}_{\mathcal{S}}^* = \arg\min_{\boldsymbol{w} \in \mathcal{S}} R_p(\boldsymbol{w})$ is nonempty, with risk $R_{p,\mathcal{S}}^*$. All rank projectors use one ordered eigenbasis and a fixed tie rule.

Theorem 2 (GRIT with sufficient rank). *Assume A1–A2 and the moment and conditional-shift conditions of Theorem 1. For every $\eta \in (0,1)$, conditional on the signals, with probability at least $1-\eta$, simultaneously for $r_* \le r \le d$, every existing minimizer $\boldsymbol{w}_r \in \arg\min_{P_{k,r}\boldsymbol{w}=0} R_p(\boldsymbol{w})$ and every $\boldsymbol{w}_{\mathcal{S}}^* \in \mathcal{W}_{\mathcal{S}}^*$ satisfy*

$$R_q(\boldsymbol{w}_r) \le \underbrace{R_{p,\mathcal{S}}^*}_{(1)} + \underbrace{C_{\mathrm{br}}(\boldsymbol{w}_r)}_{(2)} + \underbrace{L_z \|(P_{k,r} - P_{k,r_*})\boldsymbol{w}_{\mathcal{S}}^*\|_2}_{(3)} + \underbrace{C_{r,*}\left[\left(\sqrt{\frac{\sigma_\xi^2}{g_k}} + \frac{\sigma_\xi^2}{g_k}\right)\sqrt{\frac{d+\log(1/\eta)}{k}} + \frac{\sigma_\xi^2}{g_k}\frac{d+\log(1/\eta)}{k}\right]}_{(4)}, \tag{5}$$

where $C_{r,*} = C_K(L_z\|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + \rho_\Delta\|\boldsymbol{w}_r\|_2)$ and C_K depends only on K. If $k \ge d + \log(1/\eta)$, the bracket is at most $2(\sqrt{\sigma_\xi^2/g_k} + \sigma_\xi^2/g_k)\sqrt{(d+\log(1/\eta))/k}$.

Terms (1) and (2) retain stable prediction risk and bridge covariate shift. Term (3) bounds the cost of extra-direction removal: it vanishes at $r = r_*$ or when the extra directions are orthogonal to the stable oracle. Term (4) covers source prediction cost and residual unstable sensitivity from subspace estimation, with task and shift dependence in $C_{r,*}$. Because the bound holds simultaneously for every stable source optimum, one may use the optimum giving the smallest sum of terms (3) and (4). The theorem studies fitting with population source risk; its probability accounts for the k pair noises, not for sampling the labeled training data.

The noise-to-signal ratio σ_ξ^2/g_k governs recovery. Balanced coverage helps by raising the weakest signal g_k at fixed total signal energy. The $1/k$ term keeps the bound valid when $k < d + \log(1/\eta)$. With the displayed ratio and predictor factors controlled, the larger-k regime gives a $1/\sqrt{k}$ rate. Appendix B proves the sharper subspace estimate, and Appendix D gives source-dependent and quadratic extra-removal refinements. The next proposition gives a tightness result for the finite-pair

term: some problem instances attain its dependence on pair count and noise, up to constant factors, even after optimal source fitting under GRIT's estimated constraint.

Proposition 3 (Tightness of the finite-pair rate)**.** *There exist constants $c, \delta_0 > 0$ and fixed source and target distributions satisfying A1 and the moment and conditional-shift conditions, with $d = 2$, $r = r_* = 1$, $\rho_\Delta = 1$, and zero bridge shift, such that the following holds. For every $g, \sigma > 0$ and integer $k \geq 1$ with $\delta_k := \sqrt{(\sigma^2/g + \sigma^4/g^2)/k} \leq \delta_0$, there is a Gaussian A2 design with weakest clean eigenvalue g and noise variance σ^2 for which the GRIT predictor $\boldsymbol{w}_k$, fitted to minimize population source risk under its estimated constraint, satisfies with probability at least $1/64$,*

$$R_q(\boldsymbol{w}_k) - R^*_{p,\mathcal{S}} \geq c\,\delta_k\big(L_z\|\boldsymbol{w}^*_{\mathcal{S}}\|_2 + \rho_\Delta\|\boldsymbol{w}_k\|_2\big). \tag{6}$$

The fitted optimum exists for every pair realization, with uniformly bounded source and target feature norms and fitted coefficient norms.

Here $R^*_{p,\mathcal{S}}$ is the risk of the best stable source predictor. With probability at least $1/64$, GRIT's target risk exceeds this reference by an amount of order δ_k times the displayed prediction and shift factors, even though its source objective is minimized exactly. Both the loss of source predictive information and the remaining sensitivity to unstable shift contribute to this excess. For fixed dimension and confidence and sufficiently small δ_k, this attains the pair-count and noise dependence of term (4) in Theorem 2, up to constants. Appendix D.2 proves the result, and Appendix D.3 gives a further construction with positive bridge covariate shift that attains a refined target-risk comparison asymptotically.

Proposition 4 (Underselection)**.** *If $r < r_*$, then $\dim(\mathcal{U} \cap \widehat{\mathcal{U}}_r^{\perp}) \geq r_* - r$. For every fixed feasible $\boldsymbol{w}$ with $P_{\mathcal{U}}\boldsymbol{w} \neq 0$ and every source p satisfying A1 with finite second feature moments, there are feature-space targets q_t sharing $p(\boldsymbol{z}_{\mathcal{S}}, y)$ such that $\mathbb{P}_{q_t}(y\boldsymbol{w}^\top\mathbf{z} < 0) \to 1$ and $R_{q_t}(\boldsymbol{w}) \to \infty$ as $t \to \infty$.*

Underselection leaves unstable directions available; failure requires the fitted predictor to use them. This construction has zero bridge shift and growing direct-shift radius.

4.4 Guarantee for quadratic prediction consistency

Quadratic consistency uses pair information through a penalty. Building on consistency-based robust prediction (Heinze-Deml & Meinshausen, 2021; Yang et al., 2023), we use the fitted optimum's first-order condition to obtain a guarantee. For $\eta \in (0,1)$, let $h = d + \log(1/\eta)$ and define the covariance-error radius

$$e_k = C_K\left[\sigma_\xi(\sqrt{\lambda_{1,k}} + \sigma_\xi)\sqrt{\frac{h}{k}} + \sigma_\xi^2\frac{h}{k}\right], \tag{7}$$

valid for every k under A2 (Lemma 1).

Theorem 3 (Quadratic prediction consistency)**.** *Under A1–A2 and the moment, shift, and stable-optimizer existence conditions above, with conditional probability at least $1 - \eta$, every existing minimizer $\boldsymbol{w}_\lambda \in \arg\min_{\boldsymbol{w}}\{R_p(\boldsymbol{w}) + \lambda\boldsymbol{w}^\top\widetilde{\boldsymbol{\Sigma}}_k\boldsymbol{w}\}$, $\lambda > 0$, satisfies, for every $\boldsymbol{w}^*_{\mathcal{S}} \in \mathcal{W}^*_{\mathcal{S}}$,*

$$R_q(\boldsymbol{w}_\lambda) \leq \underbrace{R^*_{p,\mathcal{S}}}_{(1)} + \underbrace{C_{\mathrm{br}}(\boldsymbol{w}_\lambda)}_{(2)} + \underbrace{\frac{\lambda\sigma_\xi^2}{2}\|\boldsymbol{w}^*_{\mathcal{S}}\|_2^2 + \frac{\rho_\Delta^2}{8\lambda(g_k + \sigma_\xi^2)}}_{(3)} + \underbrace{2\lambda e_k\|\boldsymbol{w}_\lambda\|_2\|\boldsymbol{w}^*_{\mathcal{S}} - \boldsymbol{w}_\lambda\|_2}_{(4)}. \tag{8}$$

Group (3) balances preservation of stable information against residual unstable sensitivity. Increasing λ reduces the second contribution but increases the first: isotropic pair noise adds a stable-direction penalty in the mean covariance. Using optimizer stationarity halves both coefficients relative to an objective-value comparison, without adding assumptions. Group (4) captures finite-pair covariance fluctuation, whose concentration depends on the largest clean eigenvalue.

Our analysis of quadratic prediction consistency applies independently of GRIT and may be of interest in its own right. Complementing existing consistency analyses, it uses the fitted optimizer's first-order condition to connect regularization strength and pair covariance to target risk. This perspective clarifies how noise in invariant pairs affects the balance between preserving stable predictive

information and reducing unstable sensitivity. Appendix C.2 develops this analysis further through a sharper comparison retaining signed terms and additional curvature conditions under which the stable-information cost becomes quadratic in small regularization strength.

5 Related Work

Robust learning from paired observations. CoRe, MatchDG, LAM, and IPG use paired information to constrain predictions, representations, or attributions (Heinze-Deml & Meinshausen, 2021; Mahajan et al., 2021; Gao et al., 2024; Surner et al., 2026). CoRe penalizes within-identity prediction or loss variation; MatchDG learns matches and regularizes their representations. LAM matches feature contributions to logits, while IPG uses an attribution-alignment loss and class separation to guide corrective updates. These choices differ from the scalar prediction penalty analyzed here. Consistency learning already has statistical and robustness guarantees, including intervention models and misspecified augmentations (Heinze-Deml & Meinshausen, 2021; Yang et al., 2023). GRIT uses aggregate geometry from finite, noisy invariant pairs to define a reusable projection; our analysis also gives a quadratic prediction-consistency guarantee in the same setting.

Predictive invariance and invariant representations. Invariant causal prediction and invariant transfer model persistent predictive relationships (Peters et al., 2016; Rojas-Carulla et al., 2018), whereas IRM asks for a common optimal predictor (Arjovsky et al., 2019); the connection depends on loss, predictor class, and realizability (Kamath et al., 2021). Distribution alignment constrains feature distributions: DANN uses domain-adversarial adaptation (Ganin et al., 2016), C-DANN uses conditional adversarial alignment for DG (Li et al., 2018), and CORAL aligns correlations (Sun & Saenko, 2016). GRIT constrains directions in fixed representations without requiring feature distributions to agree across environments.

Complementary approaches to OOD robustness. Group-balanced last-layer retraining (Kirichenko et al., 2023) and invariant-subspace recovery from class-conditional environment moments (Wang et al., 2022) provide fixed-feature and geometric precedents with different information requirements. Other methods control worst-group risk, cross-environment optimality, risk or gradient variation (Sagawa et al., 2020; Arjovsky et al., 2019; Krueger et al., 2021; Ramé et al., 2022), or use selective interpolation (Yao et al., 2022). Spectral Decoupling (SD) penalizes squared logits to counter gradient starvation (Pezeshki et al., 2021). Risk Distribution Matching (RDM) aligns the worst environment's distribution of individual losses with their pooled distribution using maximum mean discrepancy (Nguyen et al., 2024). Projection can precede these compatible objectives. We next evaluate this compatibility and compare GRIT with prediction consistency.

6 Experimental Evaluation

We augment ColoredMNIST (Arjovsky et al., 2019) and Waterbirds (Sagawa et al., 2020) with invariant pairs by construction, calling the resulting variants *ColoredMNIST-INV* and *Waterbirds-INV*. The suffix denotes the availability of invariant-pair information. ColoredMNIST-INV provides 256 pairs showing the same training digit in both colors. For Waterbirds-INV, we follow the original dataset's foreground–background compositing methodology to construct 240 pairs showing the same bird against land and water backgrounds. Appendix F.1 gives the exact constructions and how the endpoints enter training. For the objective comparison, we use frozen, unnormalized, 512-dimensional CLIP ViT-B/32 features (Radford et al., 2021) on both variants. These constructions preserve task content without establishing exact subspace geometry or feature-noise assumptions. The frozen-feature probes differ from the original benchmarks in representation and training protocol, so absolute accuracies are not directly comparable.

Each objective uses the same labeled dataset for vanilla, prediction consistency (PC), and GRIT. PC penalizes paired logit differences, representing output-consistency components of CoRe, MatchDG, and SelfReg (Heinze-Deml & Meinshausen, 2021; Mahajan et al., 2021; Kim et al., 2021).[1] GRIT

[1] MatchDG's reported second-phase architecture takes the predictor after its representation to be the identity, so its matching penalty acts on logits (Mahajan et al., 2021, App. C.1).

projects before fitting. Both pair-aware variants use the same pair bank and frozen CLIP features, comparing a representative consistency penalty with GRIT's geometric constraint.

Appendix F details information access, validation, and seed aggregation, and reports sample standard deviations.

6.1 Does GRIT improve robustness across objectives?

Table 1 shows higher GRIT performance than vanilla for all eight objectives on both datasets. ERM gains are 50.87 and 17.40 points, respectively. Four-group GroupDRO already reaches 79.83 on Waterbirds-INV; GRIT raises this to 81.91, a 2.08-point gain.

Table 1: Mean OOD accuracy (%, ten seeds): ColoredMNIST-INV test accuracy and Waterbirds-INV worst-group accuracy. Within each objective and dataset, **bold**/*italic* mark the best/second-best mean. Solid/dotted boxes mark the best/second-best mean among all entries for each dataset. PC: prediction consistency; + adds a pair intervention to the row objective. Appendix F.3 reports sample standard deviations.

	ColoredMNIST-INV (CLIP feat.)			Waterbirds-INV (CLIP feat.)		
Objective	Vanilla	+PC	+GRIT	Vanilla	+PC	+GRIT
ERM	17.02	*57.16*	**67.89**	52.69	*66.84*	**70.09**
SD	17.02	*56.68*	**67.77**	52.06	*66.82*	**70.09**
RDM	39.95	*59.55*	**67.60**	52.80	**78.59**	*76.29*
IRMv1	59.89	*61.28*	**70.27**	47.78	*77.86*	**78.78**
V-REx	54.83	*59.98*	**68.20**	48.27	*78.09*	**80.99**
Fishr	65.28	*65.33*	**68.91**	47.16	*76.72*	**80.84**
GroupDRO	30.90	*61.51*	**68.01**	79.83	*81.66*	**81.91**
LISA	36.68	*61.62*	**67.80**	79.65	*80.77*	**82.30**

GRIT exceeds prediction consistency in 15/16 settings; prediction consistency wins with RDM on Waterbirds-INV. Bold entries compare means: Waterbirds-INV GroupDRO's paired 95% interval for GRIT versus prediction consistency includes zero (Appendix F.3).

6.2 What controls recovery from noisy invariant pairs?

We test the pair-information dependencies in Theorem 2 on *CueMNIST*: binary MNIST (LeCun et al., 1998) with two independently controlled color cues. Its engineered 96-dimensional features give a known two-dimensional unstable subspace. Each cue matches the source label with probability 0.95; the target reverses both cues. All methods use the same 3,000 labeled examples and ridge-logistic fitting. GRIT removes rank two; quadratic prediction consistency (CR, a trace-normalized score counterpart of PC) selects its strength by worst-group validation accuracy. We vary the smallest clean eigenvalue, pair noise, and pair count separately. Geometry is shown at three secondary settings; accuracy retains the original defaults. The smallest-eigenvalue sweep fixes the largest eigenvalue at one; its total clean energy varies. Independent Gaussian noise is added to the constructed feature-space pair differences.

Figure 2 supports the predicted dependencies: stronger signal in the weakest direction, less noise, and more pairs improve recovery, and the corresponding accuracy trends give evidence that these factors affect robustness. Pair count alone therefore does not characterize collection quality. CR is competitive at small budgets; at 4,096 pairs, GRIT reaches 80.0% target accuracy versus 71.4% for CR under the tested selection protocol. Appendix H gives the complete construction, fitted-instability diagnostics, Gaussian replication, and comparison limitations.

6.3 Can two pair types encode complementary invariances?

Section 2 motivates combining pair types to cover different unstable variations. UrbanCars (Li et al., 2023) makes this need concrete: both the background and a co-occurring object correlate with the car label during training. One type of invariant pair changes only the background; the other changes only the object. We vary their allocation within a fixed budget of 256 pairs on frozen CLIP features. Hyperparameters and checkpoints are selected by validation worst-group accuracy.

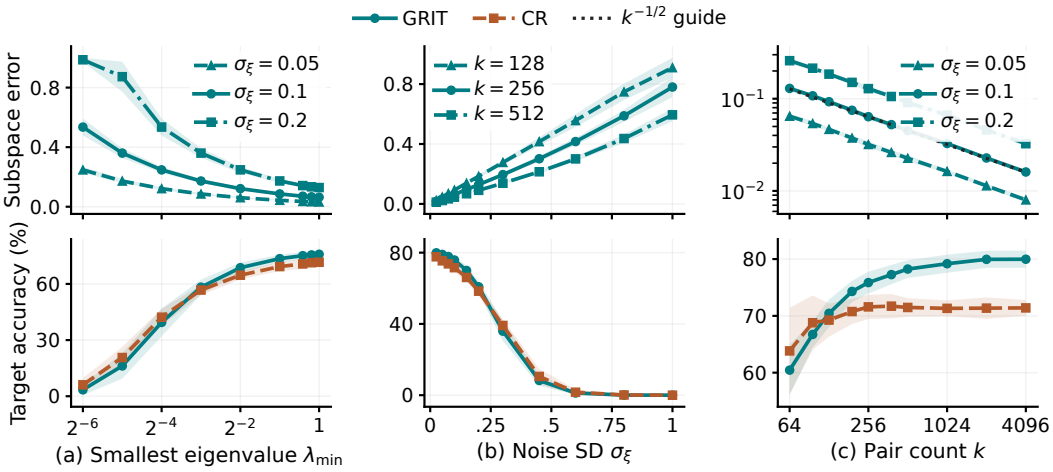

Figure 2: **Stronger signal in the weakest direction and less noise improve recovery and target accuracy; more pairs yield near-$k^{-1/2}$ subspace error and a GRIT advantage over CR at larger budgets.** On CueMNIST, the top row shows GRIT's subspace error; the bottom compares target accuracy. Defaults are $k = 256$, $\sigma_\xi = 0.1$, and clean spectrum $(1, 1)$. Top-panel legends vary a second setting; bottom panels use the defaults. Bands show sample SD across 10 replicates; the dotted line is a shape guide. Y-axis ranges differ by panel.

Table 2: Combining two pair types improves UrbanCars robustness. n_B, n_O count pairs that vary the background and object, respectively. B/O shift changes the indicated factor's label alignment, holding the other aligned; worst group is the minimum across eight label–background–object groups. Accuracy (%); bold marks each row's best method per metric. Solid/dotted boxes mark the highest/second-highest distinct mean across all allocations and methods within each metric. +PC/+GRIT add prediction consistency/GRIT to ERM.

UrbanCars (CLIP feat.)										
Pair counts		B shift			O shift			Worst group		
n_B	n_O	ERM	+PC	+GRIT	ERM	+PC	+GRIT	ERM	+PC	+GRIT
256	0	45.4	85.7	**88.5**	**76.0**	65.1	68.4	20.4	52.9	**56.7**
128	128	45.4	76.0	**79.1**	76.0	84.4	**85.1**	20.4	68.3	**71.9**
0	256	45.4	**54.7**	53.9	76.0	80.2	**94.1**	20.4	42.6	**43.8**

Table 2 shows the tradeoff: each single-type collection favors its corresponding shift, while combining both raises GRIT's worst-group accuracy from 56.7% or 43.8% to 71.9%. This is consistent with the weakest-signal dependence in Section 4.3: missing variations leave directions poorly covered. GRIT exceeds ERM and prediction consistency in worst-group mean accuracy for every allocation shown. Results average ten seeds and, for pair methods, three pair banks. The benefit of mixing pair types is smaller when validation is weighted toward the training distribution (Appendix G).

7 Discussion and Conclusion

GRIT learns which variations to ignore from finite, noisy invariant pairs and enforces the resulting constraint before predictor fitting. Our theory connects pair count, noise, coverage, and selected rank to target risk, and provides an independent analysis of quadratic prediction consistency. Experiments show improvements across the tested objectives and the value of combining complementary pair types. These results provide a foundation for extending both the predictors and the geometry learned from pairs. Nonlinear heads can use the same projection, while relaxing A1's linear stable subspace presents a deeper challenge (Appendix I.1). Progress also depends on acquiring pairs that preserve task information and cover relevant variation; broader studies are needed to assess these requirements across applications (Appendix I.2). By separating this application-specific knowledge from predictor fitting, GRIT offers a starting point for learning procedures that can turn diverse sources of invariant pairs into more robust predictions.

AI USE STATEMENT

We used generative AI in preparing this paper under the authors' direction. The work drew on a manually written draft, core theoretical ideas initially developed and checked by hand, and a concept document and outline carefully edited by the authors. AI assisted with drafting and editing, literature review, further development, checking, and contextualization of theory and proofs, and preparing tables, figures, and LaTeX. It also assisted with experimental design and interpretation, developing the codebase, implementing our method and baselines, and running experiments. The authors guided each stage, shaped the paper's ideas and structure, performed repeated manual checks and revisions, and retain responsibility for its content and conclusions.

REPRODUCIBILITY STATEMENT

Appendices A–D give the assumptions, refinements, and proofs, and Appendices F–G describe data construction, information access, optimization, and validation. Appendix H gives a reproduction procedure and the complete construction, selection, seeds, and uncertainty definitions for the controlled CueMNIST and Gaussian studies.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/pdf/1907.02893v3.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL https://arxiv.org/abs/2511.08544v1.

T. Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018. doi: 10.1214/17-AOS1541. URL https://arxiv.org/abs/1605.00353.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://jmlr.org/papers/v17/15-239.html.

Han Gao, Kaican Li, Weiyan Xie, Zhi Lin, Yongxiang Huang, Luning Wang, Caleb Cao, and Nevin Zhang. Consistency regularization for domain generalization with logit attribution matching. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244, pp. 1389–1407. PMLR, 2024. URL https://proceedings.mlr.press/v244/gao24a.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent—a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110:303–348, 2021. doi: 10.1007/s10994-020-05924-1. URL https://link.springer.com/article/10.1007/s10994-020-05924-1.

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 4069–4077. PMLR, 2021. URL https://proceedings.mlr.press/v130/kamath21a.html.

Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. SelfReg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9619–9628, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Kim_SelfReg_Self-Supervised_Contrastive_Regularization_for_Domain_Generalization_ICCV_2021_paper.html.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/pdf/2204.02937.

Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. doi: 10.1002/aic.690370209. URL https://doi.org/10.1002/aic.690370209.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5815–5826. PMLR, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://leon.bottou.org/papers/lecun-98h.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Ya_Li_Deep_Domain_Generalization_ECCV_2018_paper.html.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20071–20082, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_A_Whac-a-Mole_Dilemma_Shortcuts_Come_in_Multiples_Where_Mitigating_One_CVPR_2023_paper.html.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 7313–7324. PMLR, 2021. URL https://proceedings.mlr.press/v139/mahajan21b/mahajan21b.pdf.

Toan Nguyen, Kien Do, Bao Duong, and Thin Nguyen. Domain generalisation via risk distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2790–2799, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Nguyen_Domain_Generalisation_via_Risk_Distribution_Matching_WACV_2024_paper.html.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167. URL https://arxiv.org/pdf/1501.01332.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1256–1272, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0987b8b338d6c90bbedd8631bc499221-Paper.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.

Alexandre Ramé, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 18347–18377. PMLR, 2022. URL https://proceedings.mlr.press/v162/rame22a.html.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. URL https://jmlr.org/papers/v19/16-432.html.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. doi: 10.1162/089976698300017467. URL https://alex.smola.org/papers/1998/SchSmoMul98.pdf.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007. URL https://jmlr.org/papers/v8/sugiyama07a.html.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *Computer Vision—ECCV 2016 Workshops*, pp. 443–450. Springer, 2016. doi: 10.1007/978-3-319-49409-8_35. URL https://arxiv.org/abs/1607.01719.

Martin Surner, Abdelmajid Khelil, and Ludwig Bothmann. Invariance pair guidance: Robustness to spurious correlations via corrective gradients. *Machine Learning*, 115:188, 2026. doi: 10.1007/s10994-026-07092-0. URL https://link.springer.com/article/10.1007/s10994-026-07092-0.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, first edition, 2018. doi: 10.1017/9781108231596. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-1.pdf.

Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 23018–23033. PMLR, 2022. URL https://proceedings.mlr.press/v162/wang22x.html.

Shuo Yang, Yijun Dong, Rachel Ward, Inderjit S. Dhillon, Sujay Sanghavi, and Qi Lei. Sample efficiency of data augmentation consistency regularization. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 3825–3853. PMLR, 2023. URL https://proceedings.mlr.press/v206/yang23c.html.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 25407–25437. PMLR, 2022. URL https://proceedings.mlr.press/v162/yao22b.html.

A Proofs and Technical Details

The results below concern population prediction risk and randomness in the pairs. All shift radii $\rho_\Delta, \rho_{\mathrm{br}}, \rho_z, \rho_y$ are nonnegative. We do not assume that the clean pair signals are sampled from a common distribution. We state optimizer existence explicitly because an unregularized logistic objective can have an unattained infimum. An additive population objective gap ϵ_{opt} can be added to the hard-GRIT and objective-comparison CR bounds below. The stationarity-based CR and learned-logistic refinements use exact minimizers. Additional generalization arguments would be needed for empirical labeled-sample minimizers.

A.1 Bridge distributions, risk decomposition, and common stable prediction

All distributions are on finite-dimensional Euclidean feature spaces and the stated label space, so regular conditional distributions exist. Let μ_p denote the source distribution of $(\mathbf{z}_{\mathcal{S}}, y)$. Fix a measurable kernel $K_q(d\boldsymbol{z}_{\mathcal{U}} \mid \boldsymbol{z}_{\mathcal{S}}, y)$ that agrees with the target conditional on its almost-sure domain, and specify its values on any additional set charged by μ_p. The bridge is $\breve{q} = \mu_p K_q$. Its value, its moments, and the resulting certificate may depend on this extension when the stable-label supports differ. A conditional product coupling establishes that Π_{cond} is nonempty; a smaller displacement can be obtained from other conditional couplings. Finite second feature moments of $p, q, \breve{q}$ suffice for every fixed finite linear score used here.

Proof of Theorem 1.

Proof. The proof compares losses along a coupling that holds the stable features and label fixed, then separates the exact bridge risk difference. Put $\varphi(a) = \log(1 + e^{-a})$. Since $\varphi'(a) = -1/(1 + e^a)$, its absolute derivative is at most one. For every $\pi \in \Pi_{\mathrm{cond}}$,

$$R_q(\boldsymbol{w}) = R_p(\boldsymbol{w}) + [R_{\breve{q}}(\boldsymbol{w}) - R_p(\boldsymbol{w})] + [R_q(\boldsymbol{w}) - R_{\breve{q}}(\boldsymbol{w})], \tag{9}$$

$$R_{\breve{q}}(\boldsymbol{w}) - R_p(\boldsymbol{w}) = \mathbb{E}_\pi[\varphi(y\boldsymbol{w}^\top \mathbf{z}') - \varphi(y\boldsymbol{w}^\top \mathbf{z})], \tag{10}$$

$$\leq \mathbb{E}_\pi|\boldsymbol{w}^\top(\mathbf{z}' - \mathbf{z})|, \tag{11}$$

$$\leq \sqrt{\boldsymbol{w}^\top \mathbb{E}_\pi[\Delta\Delta^\top]\boldsymbol{w}}. \tag{12}$$

Equation (10) uses the two coupling marginals and shared label; (11) uses the derivative bound and $|y| = 1$; (12) is Cauchy–Schwarz. Bounding the last bracket in (9) by its absolute value and taking the infimum in (11) proves (4). Under the stated displacement bound, (12) is at most $\rho_\Delta \|P_{\mathcal{U}}\boldsymbol{w}\|_2$. □

Bridge transport. If γ couples q with $\breve{q}$, possibly with different labels, define $\boldsymbol{\eta}_\gamma = y^q\mathbf{z}^q - y^{\mathrm{br}}\mathbf{z}^{\mathrm{br}}$. The same derivative bound, applied to signed margins, gives

$$C_{\mathrm{br}}(\boldsymbol{w}) \leq \inf_\gamma \mathbb{E}_\gamma|\boldsymbol{w}^\top\boldsymbol{\eta}_\gamma| \leq \inf_\gamma \sqrt{\boldsymbol{w}^\top \mathbb{E}_\gamma[\boldsymbol{\eta}_\gamma\boldsymbol{\eta}_\gamma^\top]\boldsymbol{w}}. \tag{13}$$

In particular, the existence of a coupling with signed second moment at most $\rho_{\mathrm{br}}^2\boldsymbol{I}$ gives $C_{\mathrm{br}}(\boldsymbol{w}) \leq \rho_{\mathrm{br}}\|\boldsymbol{w}\|_2$. This is an optional upper bound on the exact bridge term.

A sharp direct-shift coefficient. The coefficient one in (11) cannot be uniformly reduced, even when the bridge term is zero and the direct displacement is fixed. Set $d = 2$, $\mathcal{S} = \mathrm{span}(\boldsymbol{e}_1)$, $\mathcal{U} = \mathrm{span}(\boldsymbol{e}_2)$, and $\boldsymbol{w} = \boldsymbol{e}_2$. For fixed $a > 0$ and each $b > 0$, take $y = 1$ and deterministic source and target features $(0, -b)$ and $(0, -b - a)$. A1 holds, $\breve{q} = q$, and $\mathbb{E}|\boldsymbol{w}^\top\Delta| = a$. Direct calculation gives

$$\frac{R_q(\boldsymbol{w}) - R_p(\boldsymbol{w})}{\mathbb{E}|\boldsymbol{w}^\top\Delta|} = \frac{\log(1 + e^{b+a}) - \log(1 + e^b)}{a} \longrightarrow 1. \tag{14}$$

Indeed, $\log(1 + e^c) = c + \log(1 + e^{-c})$, and the two remainders vanish as $b \to \infty$. Every member has finite moments and direct radius a. This establishes sharpness of the transport coefficient for arbitrary fixed predictors. For fitted logistic predictors, Appendix C.1 sharpens the transport bound; Appendix D.2 separately establishes matching target-risk dependence for fitted GRIT.

A.1.1 A COMMON BAYES RULE CAN HAVE DIFFERENT RISKS

Proposition 5 (Common stable logistic predictions)**.** *Suppose A1 gives a measurable conditional probability $\theta(\boldsymbol{z}_S) = \mathbb{P}(y = 1 \mid \mathbf{z}_S = \boldsymbol{z}_S) \in (0, 1)$ almost surely under both stable marginals. Suppose there is a finite $\boldsymbol{w}_S^* \in S$ with $\boldsymbol{w}_S^{*\top} \boldsymbol{z}_S = \log(\theta/(1-\theta))$ almost surely under both. Then this score minimizes logistic risk among all measurable stable scores in both distributions. If $q_S \ll p_S$, every stable source minimizer with finite risk has the same score q_S-almost surely and is also target optimal among stable scores.*

Proof. The shared conditional risk is minimized pointwise; strictness identifies source predictions, and absolute continuity transfers that identification. At a fixed stable feature value with conditional probability θ, define

$$L_\theta(a) = \theta \log(1 + e^{-a}) + (1 - \theta) \log(1 + e^a). \tag{15}$$

Its derivatives are $L'_\theta(a) = (1 + e^{-a})^{-1} - \theta$ and $L''_\theta(a) = e^{-a}/(1 + e^{-a})^2 > 0$. Thus its unique minimum is at $a^* = \log(\theta/(1 - \theta))$. The conditional excess $L_\theta(a) - L_\theta(a^*)$ is nonnegative and equals zero precisely at $a = a^*$. Integrating proves optimality under each stable marginal. A source minimizer has zero expected conditional excess, hence agrees with a^* p_S-almost surely. Absolute continuity then yields agreement q_S-almost surely. This proves equality of predictions; it does not require uniqueness of coefficient vectors. □

The shared optimum has conditional risk $-\theta \log \theta - (1-\theta) \log(1-\theta)$, whose expectation can differ across stable marginals. A1 therefore does not imply equal optimal risks. The realizability assumption is also material: a misspecified linear class can have different optimal approximations under different marginals. Absolute continuity is a sufficient condition for transferring source-almost-sure identification; topological support containment alone would not exclude a target atom on a source null set.

A.2 SPECTRAL EQUIVALENCE AND EXACT RECOVERY

Proof of Proposition 1.

Proof. The GRIT constraint fixes the retained subspace, while a quadratic penalty specifies the full covariance. Let P, P' project onto the leading rank-r eigenspaces of C, C'. Equality of the feasible sets means $\ker P = \ker P'$, or equivalently that both selected subspaces equal $\mathcal{V}$. We now characterize when C' selects $\mathcal{V}$. If it does, symmetry makes $\mathcal{V}^\perp$ invariant as well, so $P_{\mathcal{V}} C' P_{\mathcal{V}^\perp} = 0$. The strict rank-$r$ gap gives $\lambda_{\min}(C'|_{\mathcal{V}}) > \lambda_{\max}(C'|_{\mathcal{V}^\perp})$. Conversely, the zero cross block decomposes C' into its restrictions to the two subspaces. The strict ordering places all r eigenvalues from $\mathcal{V}$ first, so its leading subspace is exactly $\mathcal{V}$ and the GRIT constraints coincide. For fixed $\lambda > 0$, equality of quadratic penalties for every $\boldsymbol{w}$ implies $\boldsymbol{w}^\top \boldsymbol{D} \boldsymbol{w} = 0$ for $\boldsymbol{D} = \boldsymbol{C} - \boldsymbol{C}'$. Taking $\boldsymbol{w} = \boldsymbol{e}_i$ gives $D_{ii} = 0$, and taking $\boldsymbol{w} = \boldsymbol{e}_i + \boldsymbol{e}_j$ gives $2D_{ij} = 0$ for $i \neq j$, since $\boldsymbol{D}$ is symmetric. Therefore $\boldsymbol{C} = \boldsymbol{C}'$, which also suffices. The quadratic-form equivalence concerns the whole coefficient space. □

For any function h using only projected features and every $\boldsymbol{v} \in \operatorname{range}(P)$, $h((\boldsymbol{I} - P)(\boldsymbol{z} + \boldsymbol{v})) = h((\boldsymbol{I} - P)\boldsymbol{z})$ because $(\boldsymbol{I} - P)\boldsymbol{v} = 0$. For a linear score, the exact sensitivity identity is

$$\sup_{\boldsymbol{v} \in \operatorname{range}(P),\ \|\boldsymbol{v}\|_2 \leq \rho} |\boldsymbol{w}^\top \boldsymbol{v}| = \rho \|P\boldsymbol{w}\|_2. \tag{16}$$

Cauchy–Schwarz proves the upper bound. If $P\boldsymbol{w} \neq 0$, the vector $\rho P\boldsymbol{w}/\|P\boldsymbol{w}\|_2$ attains it; otherwise both sides vanish. The equalities concern the selected subspace, whether or not it equals $\mathcal{U}$.

Proof of Proposition 2.

Proof. The spectral theorem reduces the claim to two invariant blocks and their ordering. If the unique leading r_*-dimensional eigenspace is $\mathcal{U}$, symmetry implies that $\mathcal{S}$ is invariant too, so $P_{\mathcal{U}} \boldsymbol{M} P_{\mathcal{S}} = 0$. Its r_* selected eigenvalues must all exceed the remaining eigenvalues: a boundary tie would allow a different choice in the tied eigenspace. This proves the two conditions. Conversely, the zero cross block makes $\boldsymbol{M}$ the orthogonal direct sum of its restrictions to $\mathcal{U}$ and $\mathcal{S}$. Strict spectral

ordering places exactly the r_* eigenvalues from $\mathcal{U}$ above those from $\mathcal{S}$, giving the unique leading eigenspace. □

Proposition 6 (Perturbations preserving a selected projector)**.** *Let $\boldsymbol{M} \succeq 0$ have a strict gap $g = \lambda_r(\boldsymbol{M}) - \lambda_{r+1}(\boldsymbol{M}) > 0$, $1 \le r < d$, with leading projector P and range $\mathcal{V}$. Let $\boldsymbol{E}$ be symmetric, $\boldsymbol{M}' = \boldsymbol{M} + \boldsymbol{E} \succeq 0$, and suppose $\boldsymbol{M}'$ also has a strict rank-r gap. Then $P_r(\boldsymbol{M}') = P$ if and only if*

$$P\boldsymbol{E}(\boldsymbol{I} - P) = 0, \qquad \lambda_{\min}(\boldsymbol{M}'|_{\mathcal{V}}) > \lambda_{\max}(\boldsymbol{M}'|_{\mathcal{V}^\perp}). \tag{17}$$

Within the cross-block-preserving class, $\|\boldsymbol{E}\|_{\mathrm{op}} < g/2$ suffices; the strict radius $g/2$ cannot be increased uniformly. Without the cross-block restriction, no positive radius guarantees exact equality of the projector.

Proof. We first apply the exact recovery criterion to the selected range, then use boundary eigenvectors to locate the sharp structured threshold. Since $\boldsymbol{M}$ already preserves the two blocks, $\boldsymbol{M}'$ preserves them exactly when $P\boldsymbol{E}(\boldsymbol{I}-P) = 0$. Proposition 2, with $\mathcal{V}$ in place of $\mathcal{U}$, then proves (17). Rayleigh quotients give

$$\lambda_{\min}(\boldsymbol{M}'|_{\mathcal{V}}) \ge \lambda_r(\boldsymbol{M}) - \|\boldsymbol{E}\|_{\mathrm{op}}, \tag{18}$$

$$\lambda_{\max}(\boldsymbol{M}'|_{\mathcal{V}^\perp}) \le \lambda_{r+1}(\boldsymbol{M}) + \|\boldsymbol{E}\|_{\mathrm{op}}. \tag{19}$$

Their difference is positive when $2\|\boldsymbol{E}\|_{\mathrm{op}} < g$. Let unit eigenvectors $\boldsymbol{u} \in \mathcal{V}$ and $\boldsymbol{v} \in \mathcal{V}^\perp$ have eigenvalues λ_r and λ_{r+1}. The perturbation $\boldsymbol{E}_t = -t\boldsymbol{u}\boldsymbol{u}^\top + t\boldsymbol{v}\boldsymbol{v}^\top$ has norm t and closes their gap at $t = g/2$. The altered eigenvalues then equal $(\lambda_r + \lambda_{r+1})/2 > 0$, so the matrix remains positive semidefinite but loses uniqueness. For t slightly above $g/2$, positive semidefiniteness persists and these eigenvalues exchange order, changing the unique selected subspace. Finally, conjugating $\boldsymbol{M}$ by arbitrarily small nonzero rotations in the $(\boldsymbol{u}, \boldsymbol{v})$ plane preserves all eigenvalues and positive semidefiniteness, while changing the leading projector. The resulting perturbations converge to zero in norm, proving the last claim. □

A.3 Conditional concentration and subspace estimation

We use $\|X\|_{\psi_2} = \inf\{a > 0 : \mathbb{E}e^{X^2/a^2} \le 2\}$ and $\|X\|_{\psi_1} = \inf\{a > 0 : \mathbb{E}e^{|X|/a} \le 2\}$. All probabilities in this subsection condition on the complete signal design. Only the k original noises are independent observations; symmetrizing pair orientations does not double that count.

Lemma 1 (Pair covariance concentration)**.** *Under the conditional noise requirements of A2, for arbitrary designed signals $\boldsymbol{u}_j \in \mathbb{R}^d$ and $\eta \in (0,1)$, set $h = d\log 9 + \log(4/\eta)$ and $\lambda_{1,k} = \|\boldsymbol{\Sigma}_k\|_{\mathrm{op}}$ for $\boldsymbol{\Sigma}_k = k^{-1}\sum_j \boldsymbol{u}_j\boldsymbol{u}_j^\top$. There is a constant $C > 0$, independent of the design, dimension, and sample size, such that with probability at least $1-\eta$,*

$$\left\|\widetilde{\boldsymbol{\Sigma}}_k - (\boldsymbol{\Sigma}_k + \sigma_\xi^2\boldsymbol{I})\right\|_{\mathrm{op}} \le C\left[K\sigma_\xi\sqrt{\frac{\lambda_{1,k}h}{k}} + K^2\sigma_\xi^2\left(\sqrt{\frac{h}{k}} + \frac{h}{k}\right)\right]. \tag{20}$$

Consequently, the constant C_K in (7) can be chosen to make e_k a valid radius for every $k \ge 1$. No bound on the largest individual $\|\boldsymbol{u}_j\|_2$ is required.

Proof. The proof controls the signal–noise cross term by scalar sub-Gaussian concentration and the noise-covariance term by scalar Bernstein, uniformly over a finite net. This avoids imposing a maximum leverage condition on the deterministic design.

Step 1: scalar concentration. Expand the error into symmetric matrices

$$\boldsymbol{E} = \underbrace{\frac{1}{k}\sum_j(\boldsymbol{u}_j\boldsymbol{\xi}_j^\top + \boldsymbol{\xi}_j\boldsymbol{u}_j^\top)}_{\boldsymbol{A}} + \underbrace{\frac{1}{k}\sum_j(\boldsymbol{\xi}_j\boldsymbol{\xi}_j^\top - \sigma_\xi^2\boldsymbol{I})}_{\boldsymbol{B}}. \tag{21}$$

For a fixed unit vector $\boldsymbol{v}$, the summands $X_j = 2(\boldsymbol{v}^\top\boldsymbol{u}_j)(\boldsymbol{v}^\top\boldsymbol{\xi}_j)$ are independent and centered. Their sub-Gaussian norm squares satisfy

$$\sum_j \|X_j\|_{\psi_2}^2 \le 4K^2\sigma_\xi^2\sum_j(\boldsymbol{v}^\top\boldsymbol{u}_j)^2 = 4K^2\sigma_\xi^2 k\,\boldsymbol{v}^\top\boldsymbol{\Sigma}_k\boldsymbol{v} \le 4K^2\sigma_\xi^2 k\lambda_{1,k}. \tag{22}$$

General Hoeffding concentration therefore gives, for any $s > 0$ and a universal C_1,

$$\mathbb{P}\left(|\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}| > C_1 K \sigma_\xi \sqrt{\lambda_{1,k} s/k}\right) \le 2e^{-s}. \tag{23}$$

This is Theorem 2.6.2 of Vershynin (2018, p. 29), applied using (22). Zero-variance cases make the cross term zero and need no division by its scale.

For the second matrix, $Y_j = (\boldsymbol{v}^\top \boldsymbol{\xi}_j)^2 - \sigma_\xi^2$ are independent and centered by isotropy. Squaring a sub-Gaussian variable gives a sub-exponential variable, with $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$ (Vershynin, 2018, Lemma 2.7.6, p. 33). Centering changes this norm by at most a universal factor: the triangle inequality and Jensen's inequality give $\|\mathbb{E}X^2\|_{\psi_1} = \mathbb{E}X^2/\log 2 \le \|X^2\|_{\psi_1}$. Thus $\|Y_j\|_{\psi_1} \le 2K^2\sigma_\xi^2$. Scalar Bernstein, Theorem 2.8.1 of Vershynin (2018, p. 36), implies

$$\mathbb{P}\left(|\boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{v}| > C_2 K^2 \sigma_\xi^2 [\sqrt{s/k} + s/k]\right) \le 2e^{-s}. \tag{24}$$

Independence between the coordinates of an individual noise vector was not used.

Step 2: uniform matrix control. A $1/4$-net $\mathcal{N}$ of the unit sphere with $|\mathcal{N}| \le 9^d$ exists by the volumetric bound (Vershynin, 2018, Corollary 4.2.13, p. 84). For every symmetric matrix $\boldsymbol{H}$,

$$\|\boldsymbol{H}\|_{\mathrm{op}} \le 2 \max_{\boldsymbol{v}\in\mathcal{N}} |\boldsymbol{v}^\top \boldsymbol{H}\boldsymbol{v}|. \tag{25}$$

For completeness, take a maximizing unit vector $\boldsymbol{x}$ and $\boldsymbol{v} \in \mathcal{N}$ with $\|\boldsymbol{x} - \boldsymbol{v}\|_2 \le 1/4$. Expanding the difference of quadratic forms bounds it by $2\|\boldsymbol{H}\|_{\mathrm{op}}\|\boldsymbol{x} - \boldsymbol{v}\|_2 \le \|\boldsymbol{H}\|_{\mathrm{op}}/2$, proving (25); see also Exercise 4.4.3(b), pp. 89–90 of Vershynin (2018). Applying (23) and (24) at $s = h$ to each net point fails with probability at most $4 \cdot 9^d e^{-h} = \eta$. On the remaining event, (25) and the triangle inequality yield (20).

Step 3: the main-text radius. For $d \ge 1$ and $\eta \in (0, 1)$, $d \log 9 + \log(4/\eta) \le 4(d + \log(1/\eta))$. Absorbing these fixed factors and K, K^2 into C_K gives (7) for every k. When $\sigma_\xi = 0$, covariance zero and mean zero imply $\boldsymbol{\xi}_j = 0$ almost surely, so both radii are zero. □

Lemma 2 (Projector perturbation under A2 geometry)**.** *Assume the A2 signal geometry* $\mathrm{range}(\boldsymbol{\Sigma}_k) = \mathcal{U}$*, with* $1 \le r_* < d$*. Let* $\boldsymbol{M} = \boldsymbol{\Sigma}_k + \sigma_\xi^2 \boldsymbol{I}$*,* $g = \lambda_{r_*,k} > 0$*, and let* $\widetilde{\boldsymbol{\Sigma}}_k = \boldsymbol{M} + \boldsymbol{E}$ *be symmetric with* $\|\boldsymbol{E}\|_{\mathrm{op}} \le e < g/2$*. Then the leading rank-*r_* *projector is unique and*

$$\|P_{k,r_*} - P_{\mathcal{U}}\|_{\mathrm{op}} \le \frac{e}{g - e} \le \frac{2e}{g}. \tag{26}$$

For every $r \ge r_*$ *using the same ordered eigenbasis,* $\|P_{\mathcal{U}}(\boldsymbol{I} - P_{k,r})\|_{\mathrm{op}} \le 2e/g$*.*

Proof. The isotropic noise makes the stable block a scalar matrix, so projecting the perturbed eigenvalue equation onto that block directly controls leakage. Rayleigh quotients, or the min–max characterization of eigenvalues, give

$$\widetilde{\lambda}_{r_*} \ge \sigma_\xi^2 + g - e, \qquad \widetilde{\lambda}_{r_*+1} \le \sigma_\xi^2 + e. \tag{27}$$

The gap is at least $g - 2e > 0$. Let $\widehat{\boldsymbol{Q}}$ contain the leading r_* orthonormal eigenvectors and $\widehat{\boldsymbol{\Lambda}}$ their eigenvalues. Since $P_{\mathcal{S}}\boldsymbol{\Sigma}_k = 0$, their eigenvalue equation implies

$$P_{\mathcal{S}}(\boldsymbol{M} + \boldsymbol{E})\widehat{\boldsymbol{Q}} = P_{\mathcal{S}}\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{\Lambda}}, \tag{28}$$

$$P_{\mathcal{S}}\widehat{\boldsymbol{Q}}(\widehat{\boldsymbol{\Lambda}} - \sigma_\xi^2 \boldsymbol{I}) = P_{\mathcal{S}}\boldsymbol{E}\widehat{\boldsymbol{Q}}, \tag{29}$$

$$\|P_{\mathcal{S}}\widehat{\boldsymbol{Q}}\|_{\mathrm{op}} \le e\|(\widehat{\boldsymbol{\Lambda}} - \sigma_\xi^2 \boldsymbol{I})^{-1}\|_{\mathrm{op}} \le \frac{e}{g - e}. \tag{30}$$

The inverse exists by (27).

To relate leakage to projector error without an additional perturbation theorem, let P, Q be equal-rank orthogonal projectors with orthonormal bases V, W. Both $\|(\boldsymbol{I} - P)Q\|_{\mathrm{op}}^2$ and $\|P(\boldsymbol{I} - Q)\|_{\mathrm{op}}^2$ equal $1 - s_{\min}(V^\top W)^2$, by multiplying the compressed projectors and using the singular values

of the square matrix $V^\top W$. Furthermore, $(P-Q)^2$ has diagonal blocks $P(\boldsymbol{I}-Q)P$ and $(\boldsymbol{I}-P)Q(\boldsymbol{I}-P)$ relative to P. Its largest eigenvalue is the maximum of the two equal leakage squares. Consequently $\|P-Q\|_{\mathrm{op}} = \|(\boldsymbol{I}-P)Q\|_{\mathrm{op}} = \|P(\boldsymbol{I}-Q)\|_{\mathrm{op}}$. Applying this identity to (30) proves (26). Finally, $(\boldsymbol{I}-P_{k,r_*})(\boldsymbol{I}-P_{k,r}) = \boldsymbol{I}-P_{k,r}$ by nesting, so enlarging the removed subspace cannot increase missed-unstable leakage. □

Proposition 7 (Consistency along designed collections)**.** *Consider a sequence of A2 designs, with $1 \le r_* < d$, $g_k = \lambda_{r_*,k} > 0$, and $\boldsymbol{E}_k = \widetilde{\boldsymbol{\Sigma}}_k - (\boldsymbol{\Sigma}_k + \sigma^2_{\xi,k}\boldsymbol{I})$. Then $\|P_{k,r_*} - P_{\mathcal{U}}\|_{\mathrm{op}} \to 0$ in probability, under any fixed measurable tie rule, if either of the following holds:*

(i) $\|\boldsymbol{E}_k\|_{\mathrm{op}}/g_k \to 0$ in probability;

(ii) K is uniformly bounded and, for some $\eta_k \downarrow 0$, the weakest-signal factor $T_k(\eta_k)$ in (75), with $g = g_k$ and $\sigma = \sigma_{\xi,k}$, tends to zero.

A sufficient condition for (i) is that the radius in (20), evaluated at $\eta_k \downarrow 0$ and divided by g_k, tends to zero, with K uniformly bounded.

Proof. Both conditions give a vanishing projector bound on an event whose probability tends to one. Under (i), with probability tending to one, $\|\boldsymbol{E}_k\|_{\mathrm{op}} < g_k/2$; Lemma 2 bounds projector error by twice this vanishing ratio. Lemma 1 makes the stated covariance-radius condition sufficient. Under (ii), Theorem 6 bounds projector error by $C_K T_k(\eta_k) \to 0$ with probability at least $1-\eta_k \to 1$. Either argument proves convergence in probability. The matrices need not converge to one common covariance, and the designs need not be nested. □

Condition (ii) avoids dependence on the largest clean eigenvalue and does not require controlling the full covariance error relative to the gap. Isotropy of the mean alone does not control finite-pair rotation. More generally, mean-zero anisotropic noise has expected covariance $\boldsymbol{\Sigma}_k + k^{-1}\sum_j \mathrm{Cov}(\boldsymbol{\xi}_j)$. Proposition 2 can still apply when this matrix preserves the blocks and their ordering. Arbitrary anisotropy or bias need not do so, even with infinitely many pairs.

A.4 Shared proxy bound and the GRIT guarantee

The next result shows how source prediction, coverage, and the cost of a constraint combine before choosing a particular pair estimator. Define $\boldsymbol{\Sigma}_{z,p} = \mathbb{E}_p[\mathbf{z}\mathbf{z}^\top]$ and $\|\boldsymbol{v}\|_{\boldsymbol{H}} = \sqrt{\boldsymbol{v}^\top \boldsymbol{H}\boldsymbol{v}}$ for a positive semidefinite matrix $\boldsymbol{H}$.

Lemma 3 (A deterministic rank–budget bound)**.** *Let $\boldsymbol{C} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ be a rank-r positive semidefinite matrix, $1 \le r \le d$, where $\boldsymbol{Q}$ has orthonormal columns and $\boldsymbol{\Lambda} \succ 0$. Set $P = \boldsymbol{Q}\boldsymbol{Q}^\top$, $P_0 = \boldsymbol{I} - P$, $\alpha = \|P_{\mathcal{U}}P_0\|_{\mathrm{op}}$, and $\kappa = \|P_{\mathcal{U}}\boldsymbol{Q}\boldsymbol{\Lambda}^{-1/2}\|_{\mathrm{op}}$. For $\tau \ge 0$, every feasible $\boldsymbol{w}$ satisfies*

$$\|P_{\mathcal{U}}\boldsymbol{w}\|_2 \le \alpha\|P_0\boldsymbol{w}\|_2 + \kappa\tau, \qquad \|\boldsymbol{w}\|_{\boldsymbol{C}} \le \tau. \tag{31}$$

If $\boldsymbol{w}_\tau$ minimizes R_p over this feasible set and $\boldsymbol{w}^*_{\mathcal{S}}$ is a stable source minimizer, put $b = \|\boldsymbol{w}^*_{\mathcal{S}}\|_{\boldsymbol{C}}$ and $s = (b-\tau)_+/b$, defining $s = 0$ when $b = 0$. Then, under Theorem 1's assumptions,

$$R_q(\boldsymbol{w}_\tau) \le R^*_{p,\mathcal{S}} + C_{\mathrm{br}}(\boldsymbol{w}_\tau) + s\|P\boldsymbol{w}^*_{\mathcal{S}}\|_{\boldsymbol{\Sigma}_{z,p}} + \rho_\Delta(\alpha\|P_0\boldsymbol{w}_\tau\|_2 + \kappa\tau). \tag{32}$$

Proof. Decompose the learned coefficient into constrained and unconstrained components. Then construct a feasible stable-oracle approximation by shrinking only the constrained component. The first step is

$$P_{\mathcal{U}}\boldsymbol{w} = P_{\mathcal{U}}P_0\boldsymbol{w} + P_{\mathcal{U}}\boldsymbol{Q}\boldsymbol{\Lambda}^{-1/2}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\top\boldsymbol{w}), \tag{33}$$

$$\|P_{\mathcal{U}}\boldsymbol{w}\|_2 \le \alpha\|P_0\boldsymbol{w}\|_2 + \kappa\|\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}^\top\boldsymbol{w}\|_2 \le \alpha\|P_0\boldsymbol{w}\|_2 + \kappa\tau. \tag{34}$$

For the source comparison, let $\boldsymbol{v} = \boldsymbol{w}^*_{\mathcal{S}} - sP\boldsymbol{w}^*_{\mathcal{S}}$. Its proxy norm is $(1-s)b \le \tau$, including $b = 0$. The unit logistic derivative bound gives

$$R_p(\boldsymbol{w}_\tau) \le R_p(\boldsymbol{v}), \tag{35}$$

$$R_p(\boldsymbol{v}) - R_p(\boldsymbol{w}^*_{\mathcal{S}}) \le \mathbb{E}_p|s\,\mathbf{z}^\top P\boldsymbol{w}^*_{\mathcal{S}}|, \tag{36}$$

$$\le s\|P\boldsymbol{w}^*_{\mathcal{S}}\|_{\boldsymbol{\Sigma}_{z,p}}. \tag{37}$$

Equation (35) is source optimality; (37) is Cauchy–Schwarz. Combine these inequalities with (31) and Theorem 1. The result is deterministic once the proxy is fixed. □

A quadratic-penalty minimizer is also a source-risk minimizer in the budget set whose radius is its own proxy norm: any point in that set with smaller risk would have smaller penalized objective.

Coverage over an unbounded feasible class. For $\rho > 0$, $\sup_{\|\boldsymbol{w}\|_{\boldsymbol{C}} \le \tau} \rho \|P_{\mathcal{U}} \boldsymbol{w}\|_2$ is finite if and only if $\mathcal{U} \subseteq \operatorname{range}(\boldsymbol{C})$. If coverage holds, $P_{\mathcal{U}} P_0 = 0$ and (31) gives finiteness. If it fails, some $\boldsymbol{v} \in \ker \boldsymbol{C}$ has $P_{\mathcal{U}} \boldsymbol{v} \neq 0$; every multiple of $\boldsymbol{v}$ is feasible and has unbounded sensitivity. This concerns certification over the whole unbounded feasible class. A particular stable predictor has zero direct sensitivity even without coverage; a norm-bounded feasible class has finite sensitivity regardless of coverage.

Proof of Theorem 2.

Proof. Project a stable oracle into the GRIT-feasible space to control source risk, then use estimated-subspace leakage to control direct unstable shift. Work on the event of Theorem 6, common to all ranks and coefficients. Put $\widehat{P} = P_{k,r_*}$, $P = P_{k,r}$, and $\alpha = \|\widehat{P} - P_{\mathcal{U}}\|_{\mathrm{op}}$. Because $r \ge r_*$, $P - \widehat{P}$ is an orthogonal projector. The feasible comparator $(\boldsymbol{I} - P)\boldsymbol{w}^*_{\mathcal{S}}$ gives

$$
\begin{aligned}
R_p(\boldsymbol{w}_r) &\le R_p((\boldsymbol{I} - P)\boldsymbol{w}^*_{\mathcal{S}}), && (38)\\
&\le R^*_{p,\mathcal{S}} + \|P\boldsymbol{w}^*_{\mathcal{S}}\|_{\boldsymbol{\Sigma}_{z,p}}, && (39)\\
&\le R^*_{p,\mathcal{S}} + L_z\|(P - \widehat{P})\boldsymbol{w}^*_{\mathcal{S}}\|_2 + L_z\|\widehat{P}\boldsymbol{w}^*_{\mathcal{S}}\|_2, && (40)\\
&\le R^*_{p,\mathcal{S}} + L_z\|(P - \widehat{P})\boldsymbol{w}^*_{\mathcal{S}}\|_2 + \alpha L_z\|\boldsymbol{w}^*_{\mathcal{S}}\|_2. && (41)
\end{aligned}
$$

Equation (39) follows from the same Lipschitz comparison used in Lemma 3; (40) uses the source second moment and the triangle inequality. For (41), use $P_{\mathcal{U}}\boldsymbol{w}^*_{\mathcal{S}} = 0$ and the definition of α. Feasibility and $\operatorname{range}(\widehat{P}) \subseteq \operatorname{range}(P)$ also give

$$
\|P_{\mathcal{U}}\boldsymbol{w}_r\|_2 = \|P_{\mathcal{U}}(\boldsymbol{I} - P)\boldsymbol{w}_r\|_2 \le \|P_{\mathcal{U}}(\boldsymbol{I} - \widehat{P})\|_{\mathrm{op}}\|\boldsymbol{w}_r\|_2 \le \alpha\|\boldsymbol{w}_r\|_2. \tag{42}
$$

Substitute (41) and (42) into Theorem 1, and bound α by Theorem 6. Every comparison was pointwise on the same matrix event, proving the simultaneous rank and oracle quantifiers. At $r = d$ the only feasible coefficient is zero and the argument remains valid. □

The same proof permits the clipped radius $\min\{1, C_K T_k\}$ from Theorem 6 in place of $C_K T_k$, or the sharper block radii given there. Stable and constrained minimizers need not be unique; taking the infimum of the combined oracle-dependent terms is valid. An approximate constrained population optimum adds its objective gap to the source comparison. Appendix D refines that comparison using source stationarity and logistic curvature.

Fixed-energy interpretation. At fixed $T = \operatorname{tr}(\boldsymbol{\Sigma}_k) > 0$ and r_*, the weakest signal obeys $g_k \le T/r_*$. Equality holds exactly when all positive clean eigenvalues equal T/r_*. Every noise-to-signal term in (5) decreases as g_k increases, so this spectrum minimizes the displayed pair factor at fixed energy.

A.5 Underselection and actual target-risk failure

Proof of Proposition 4.

Proof. Rank–nullity supplies an available unstable direction. A label-dependent translation along the predictor's unstable component then makes its signed margin tend to minus infinity while preserving the stable-label joint distribution. For the restriction $P_{k,r}|_{\mathcal{U}} : \mathcal{U} \to \widehat{\mathcal{U}}_r$, its rank is at most r, so

$$
\dim\ker(P_{k,r}|_{\mathcal{U}}) = r_* - \operatorname{rank}(P_{k,r}|_{\mathcal{U}}) \ge r_* - r. \tag{43}
$$

Its kernel is $\mathcal{U} \cap \widehat{\mathcal{U}}_r^{\perp}$, proving the first part. For a fixed feasible $\boldsymbol{w}$ with $c = \|P_{\mathcal{U}}\boldsymbol{w}\|_2 > 0$, let $\boldsymbol{u} = P_{\mathcal{U}}\boldsymbol{w}/c$ and define q_t as the pushforward of p through

$$(\boldsymbol{z}, y) \longmapsto (\boldsymbol{z} - ty\boldsymbol{u}, y), \qquad t \geq 0. \tag{44}$$

The stable projection and label are unchanged because $\boldsymbol{u} \in \mathcal{U}$, hence A1 is preserved and $\breve{q}_t = q_t$ for the corresponding conditional version. The signed margin satisfies

$$y\boldsymbol{w}^\top(\mathbf{z} - ty\boldsymbol{u}) = y\boldsymbol{w}^\top\mathbf{z} - tc. \tag{45}$$

The initial margin is finite almost surely. Therefore its probability of being strictly negative tends to one by bounded convergence of the corresponding indicators. The logistic loss in (45) increases pointwise to infinity; monotone convergence gives $R_{q_t}(\boldsymbol{w}) \to \infty$. Each finite-$t$ target has finite second feature moments. The explicit coupling has displacement $-ty\boldsymbol{u}$ and second moment $t^2\boldsymbol{u}\boldsymbol{u}^\top \preceq t^2 P_{\mathcal{U}}$, so a radius $\rho_\Delta = t$ suffices. The bridge term is zero for every t. □

The proposition constructs feature distributions. Realizing these translations through a restricted pretrained feature map is a separate condition; identity features suffice. Moreover, underselection alone does not force the source optimum to use an unstable component. The following construction shows that it can do so even when pair geometry is exact and the direct radius stays fixed.

Proposition 8 (A source optimizer can exploit an omitted direction)**.** *For every $0 \leq r < m$ and $\epsilon \in (0, 1/4)$, there exist distributions with $\dim\mathcal{U} = m$ satisfying A1 and arbitrarily large clean designed pair collections satisfying A2 such that rank-r GRIT has a unique population source logistic minimizer with source error ϵ and target error $1 - \epsilon$. The stable optimum has error $1/4$ on both distributions. The bridge term is zero and direct radius two suffices. As $\epsilon \downarrow 0$, the GRIT source logistic risk tends to zero while its target risk tends to infinity.*

Proof. We place one source-predictive unstable coordinate below the selected spectral cutoff, and reverse its association with the label at target. Take features $\mathbf{z} = (X, Z_1, \ldots, Z_{m-1}, T) \in \mathbb{R}^{m+1}$, stable space $\mathrm{span}(\boldsymbol{e}_X)$, and unstable complement. Let X be a fair sign and $\mathbb{P}(y = y \mid X = x) = (1 + e^{-ybx})^{-1}$ with $b = \log 3$. Let $V \in \{-1, 1\}$ be an independent sign with $\mathbb{P}(V = 1) = 1 - \epsilon$, and let all Z_i be independent fair signs, independent of (X, y, V). Source features use $T = yV$; target features use $T = -yV$. The stable-label distribution is identical.

Put $a = \log((1 - \epsilon)/\epsilon) > b$. Bayes' rule gives the source conditional log odds $bX + aT$, so the coefficient $\boldsymbol{w}_p = b\boldsymbol{e}_X + a\boldsymbol{e}_T$ realizes the Bayes logit. Source feature support is the complete sign cube, with positive probability at each point and label probabilities strictly between zero and one. For every finite coefficient and every nonzero direction $\boldsymbol{v}$, the logistic Hessian quadratic form is

$$\mathbb{E}_p\left[\frac{e^{\boldsymbol{w}^\top\mathbf{z}}}{(1 + e^{\boldsymbol{w}^\top\mathbf{z}})^2}(\boldsymbol{v}^\top\mathbf{z})^2\right] > 0. \tag{46}$$

The strict inequality follows because the cube spans the feature space. Hence $\boldsymbol{w}_p$ is the unique source minimizer.

Choose distinct positive numbers $\mu_1 > \cdots > \mu_m$, with the T coordinate assigned μ_m. At $k = Nm$, repeat each clean signal $\sqrt{m\mu_i}\boldsymbol{e}_i \in \mathcal{U}$ exactly N times and set all noise to zero. The pair covariance is diagonal with these unstable eigenvalues and a zero stable eigenvalue, for every N. Selecting its leading r eigenvectors removes r of the Z_i and leaves T. Thus $\boldsymbol{w}_p$ remains feasible and is the unique constrained source minimizer. Identity features and pairs with identical stable coordinates realize these designed differences.

Because $a > b$, the sign of $bX + aT$ equals T everywhere. Source and target errors are consequently ϵ and $1-\epsilon$. The stable minimizer is $b\boldsymbol{e}_X$, with error $1/4$ in both domains. To verify the surrogate-risk claim directly, evaluating the source risk at $a\boldsymbol{e}_T$ gives

$$R_p(\boldsymbol{w}_p) \leq (1 - \epsilon)\log(1 + e^{-a}) + \epsilon\log(1 + e^{a}) = -\epsilon\log\epsilon - (1 - \epsilon)\log(1 - \epsilon) \longrightarrow 0. \tag{47}$$

At target, the event $T = -y$ has probability $1 - \epsilon$ and margin at most $b - a$, giving

$$R_q(\boldsymbol{w}_p) \geq (1 - \epsilon)\log(1 + e^{a-b}) \geq (1 - \epsilon)(a - b) \longrightarrow \infty. \tag{48}$$

Coupling the domains with the same (X, y, V, Z_i) flips only T; the displacement second moment is $4\boldsymbol{e}_T\boldsymbol{e}_T^\top \preceq 4P_{\mathcal{U}}$. The bridge equals the target. At each fixed ϵ, moments and coefficients are finite, but $\|\boldsymbol{w}_p\|_2$ grows as $\epsilon \downarrow 0$; no fixed-norm claim is made. □

A.6 Quadratic consistency: main proof and an objective comparison

Proof of Theorem 3.

Proof. The proof combines covariance concentration with a deterministic bound that uses the penalized optimum's first-order condition. Lemma 1 gives $\widetilde{\boldsymbol{\Sigma}}_k = \boldsymbol{\Sigma}_k + \sigma_\xi^2 \boldsymbol{I} + \boldsymbol{E}$ with $\|\boldsymbol{E}\|_{\mathrm{op}} \le e_k$ on one event of conditional probability at least $1-\eta$. Proposition 13 applies on that event with $\boldsymbol{v} = \boldsymbol{w}_{\mathcal{S}}^*$ and $g = g_k$, proving (8). The event is common to every regularization strength, optimizer, and stable oracle. □

An alternative certificate from objective values. For comparison, an objective-value argument gives a second valid certificate. It also accommodates an additive optimization gap and supplies the MSE argument below. On the covariance event, abbreviate $\boldsymbol{w} = \boldsymbol{w}_\lambda$ and $g = g_k$. Source optimality yields

$$R_p(\boldsymbol{w}) \le R_{p,\mathcal{S}}^* + \lambda \boldsymbol{w}_{\mathcal{S}}^{*\top} \widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{w}_{\mathcal{S}}^* - \lambda \boldsymbol{w}^\top \widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{w}, \tag{49}$$

$$\le R_{p,\mathcal{S}}^* + \lambda\sigma_\xi^2 \|\boldsymbol{w}_{\mathcal{S}}^*\|_2^2 - \lambda \boldsymbol{w}^\top \boldsymbol{\Sigma}_k \boldsymbol{w} - \lambda\sigma_\xi^2 \|\boldsymbol{w}\|_2^2 + \lambda e_k(\|\boldsymbol{w}_{\mathcal{S}}^*\|_2^2 + \|\boldsymbol{w}\|_2^2), \tag{50}$$

$$\le R_{p,\mathcal{S}}^* + \lambda\sigma_\xi^2 \|\boldsymbol{w}_{\mathcal{S}}^*\|_2^2 - \lambda(g+\sigma_\xi^2)\|P_{\mathcal{U}}\boldsymbol{w}\|_2^2 - \lambda\sigma_\xi^2 \|P_{\mathcal{S}}\boldsymbol{w}\|_2^2 + \lambda e_k(\|\boldsymbol{w}_{\mathcal{S}}^*\|_2^2 + \|\boldsymbol{w}\|_2^2). \tag{51}$$

Equation (50) uses $\boldsymbol{\Sigma}_k \boldsymbol{w}_{\mathcal{S}}^* = 0$ and $|\boldsymbol{v}^\top \boldsymbol{E} \boldsymbol{v}| \le e_k \|\boldsymbol{v}\|_2^2$. Equation (51) uses $\boldsymbol{\Sigma}_k \succeq g P_{\mathcal{U}}$ and orthogonality of $\mathcal{U}$ and $\mathcal{S}$. Add the bridge term and $\rho_\Delta t$ from Theorem 1, where $t = \|P_{\mathcal{U}}\boldsymbol{w}\|_2$. The elementary identity

$$\rho_\Delta t - \lambda(g+\sigma_\xi^2)t^2 = \frac{\rho_\Delta^2}{4\lambda(g+\sigma_\xi^2)} - \lambda(g+\sigma_\xi^2)\left(t - \frac{\rho_\Delta}{2\lambda(g+\sigma_\xi^2)}\right)^2 \tag{52}$$

bounds their sum by the first term on the right. Dropping the nonpositive stable penalty gives the alternative bound

$$R_q(\boldsymbol{w}) \le R_{p,\mathcal{S}}^* + C_{\mathrm{br}}(\boldsymbol{w}) + \lambda\sigma_\xi^2 \|\boldsymbol{w}_{\mathcal{S}}^*\|_2^2 + \frac{\rho_\Delta^2}{4\lambda(g+\sigma_\xi^2)} + \lambda e_k(\|\boldsymbol{w}_{\mathcal{S}}^*\|_2^2 + \|\boldsymbol{w}\|_2^2). \tag{53}$$

Both (8) and (53) are valid on this event, so their minimum is valid too. The stationarity bound halves the two population-matrix coefficients, but its covariance remainder need not be smaller for every oracle and fitted coefficient. An approximate penalized minimizer adds its objective gap to (53); objective closeness alone does not provide the exact stationarity used by the stronger bound.

The nonnegative terms in either certificate can exceed the actual risk increase. Appendix C.2 retains the signed identity, gives a quadratic refinement under segment curvature, and distinguishes fixed-source small-regularization behavior from worst-case behavior over a changing source class.

A.7 Mean squared error regression

For real labels with finite second moments under $p, q, \breve{q}$, define $R_p^{\mathrm{MSE}}(\boldsymbol{w}) = \mathbb{E}_p(\boldsymbol{w}^\top \mathbf{z} - y)^2$ and $R_p^{\mathrm{RMSE}}(\boldsymbol{w}) = \sqrt{R_p^{\mathrm{MSE}}(\boldsymbol{w})}$. Use the same common conditional label distribution in A1 and the same conditional couplings, now for real labels. The exact bridge quantity is $C_{\mathrm{br}}^{\mathrm{RMSE}}(\boldsymbol{w}) = |R_q^{\mathrm{RMSE}}(\boldsymbol{w}) - R_{\breve{q}}^{\mathrm{RMSE}}(\boldsymbol{w})|$. The source moment constant remains $L_z = \|\mathbb{E}_p[\mathbf{z}\mathbf{z}^\top]\|_{\mathrm{op}}^{1/2}$; no global Lipschitz bound on squared loss is assumed.

Theorem 4 (Regression transport and GRIT)**.** *Suppose a conditional coupling satisfies* $\mathbb{E}[\Delta\Delta^\top] \preceq \rho_\Delta^2 P_{\mathcal{U}}$ *and the stated second moments are finite. Then*

$$R_q^{\mathrm{RMSE}}(\boldsymbol{w}) \le R_p^{\mathrm{RMSE}}(\boldsymbol{w}) + C_{\mathrm{br}}^{\mathrm{RMSE}}(\boldsymbol{w}) + \rho_\Delta \|P_{\mathcal{U}}\boldsymbol{w}\|_2. \tag{54}$$

Let $\boldsymbol{w}_{\mathcal{S}}^*$ *minimize source MSE over* $\mathcal{S}$*. Under A2, on the event of Theorem 6, with its radius* γ_k*, every source-MSE minimizer* $\boldsymbol{w}_r$ *subject to* $P_{k,r}\boldsymbol{w} = 0$*,* $r \ge r_*$*, satisfies*

$$R_q^{\mathrm{RMSE}}(\boldsymbol{w}_r) \le R_p^{\mathrm{RMSE}}(\boldsymbol{w}_{\mathcal{S}}^*) + C_{\mathrm{br}}^{\mathrm{RMSE}}(\boldsymbol{w}_r) + L_z\|(P_{k,r} - P_{k,r_*})\boldsymbol{w}_{\mathcal{S}}^*\|_2 + \gamma_k(L_z\|\boldsymbol{w}_{\mathcal{S}}^*\|_2 + \rho_\Delta\|\boldsymbol{w}_r\|_2). \tag{55}$$

Squaring the nonnegative right side gives an MSE bound.

Proof. Apply the triangle inequality in L^2 to coupled prediction residuals, then use the same feasible-oracle geometry as for logistic risk. Under a conditional coupling, labels are shared, so

$$R_{\check{q}}^{\mathrm{RMSE}}(\boldsymbol{w}) = \|\boldsymbol{w}^\top \mathbf{z} - y + \boldsymbol{w}^\top \Delta\|_{L^2}, \tag{56}$$

$$\leq \|\boldsymbol{w}^\top \mathbf{z} - y\|_{L^2} + \|\boldsymbol{w}^\top \Delta\|_{L^2}, \tag{57}$$

$$\leq R_p^{\mathrm{RMSE}}(\boldsymbol{w}) + \rho_\Delta \|P_{\mathcal{U}} \boldsymbol{w}\|_2. \tag{58}$$

Adding the exact RMSE bridge difference proves (54). The source minimizer of MSE also minimizes RMSE, because square root is increasing. With $P = P_{k,r}$, source optimality and the L^2 triangle inequality give

$$R_p^{\mathrm{RMSE}}(\boldsymbol{w}_r) \leq R_p^{\mathrm{RMSE}}((\boldsymbol{I} - P)\boldsymbol{w}_{\mathcal{S}}^*) \leq R_p^{\mathrm{RMSE}}(\boldsymbol{w}_{\mathcal{S}}^*) + \|P\boldsymbol{w}_{\mathcal{S}}^*\|_{\boldsymbol{\Sigma}_{z,p}}. \tag{59}$$

The decomposition of the final norm into extra-direction removal and rank-r_* leakage is exactly (40)–(41), and (42) controls unstable sensitivity. Substitution proves (55). All terms on its right are nonnegative, justifying squaring. □

Under a bridge coupling γ for real labels, the reverse L^2 triangle inequality also gives

$$C_{\mathrm{br}}^{\mathrm{RMSE}}(\boldsymbol{w}) \leq \left\|\boldsymbol{w}^\top(\mathbf{z}^q - \mathbf{z}^{\mathrm{br}}) - (y^q - y^{\mathrm{br}})\right\|_{L^2(\gamma)}. \tag{60}$$

If feature displacement second moment is at most $\rho_z^2 \boldsymbol{I}$ and label displacement second moment at most ρ_y^2, its right side is at most $\rho_z\|\boldsymbol{w}\|_2 + \rho_y$. The deterministic rank–budget result also extends to RMSE with the same $s\|P\boldsymbol{w}_{\mathcal{S}}^*\|_{\boldsymbol{\Sigma}_{z,p}}$ term, by the source comparison just proved.

Theorem 5 (Quadratic consistency for MSE)**.** *Under A2 and the regression moment and conditional-shift conditions above, let* $\boldsymbol{w}_\lambda \in \arg\min_{\boldsymbol{w}}\{R_p^{\mathrm{MSE}}(\boldsymbol{w}) + \lambda \boldsymbol{w}^\top \widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{w}\}$, $\lambda > 0$, *and let* $\boldsymbol{w}_{\mathcal{S}}^*$ *be a stable source-MSE minimizer. On the covariance event* $\|\widetilde{\boldsymbol{\Sigma}}_k - (\boldsymbol{\Sigma}_k + \sigma_\xi^2 \boldsymbol{I})\|_{\mathrm{op}} \leq e$, *put*

$$A = R_p^{\mathrm{MSE}}(\boldsymbol{w}_{\mathcal{S}}^*) + \lambda\sigma_\xi^2\|\boldsymbol{w}_{\mathcal{S}}^*\|_2^2 + \lambda e(\|\boldsymbol{w}_{\mathcal{S}}^*\|_2^2 + \|\boldsymbol{w}_\lambda\|_2^2). \tag{61}$$

Then

$$R_q^{\mathrm{MSE}}(\boldsymbol{w}_\lambda) \leq \left[C_{\mathrm{br}}^{\mathrm{RMSE}}(\boldsymbol{w}_\lambda) + \sqrt{\left(1 + \frac{\rho_\Delta^2}{\lambda(\lambda_{r_*,k} + \sigma_\xi^2)}\right) A}\right]^2. \tag{62}$$

In particular, $e = e_k$ *from Lemma 1 gives the finite-pair guarantee without a small-gap requirement.*

Proof. Source optimality provides a quadratic budget for the source residual and unstable coefficient. A two-dimensional Cauchy–Schwarz inequality then combines that budget with RMSE transport. Let $a = \lambda(\lambda_{r_*,k} + \sigma_\xi^2) > 0$ and $t = \|P_{\mathcal{U}}\boldsymbol{w}_\lambda\|_2$. Repeating (49)–(51) for the MSE objective and dropping the nonpositive stable penalty gives

$$R_p^{\mathrm{MSE}}(\boldsymbol{w}_\lambda) + at^2 \leq A. \tag{63}$$

Therefore, by (54),

$$R_q^{\mathrm{RMSE}}(\boldsymbol{w}_\lambda) \leq C_{\mathrm{br}}^{\mathrm{RMSE}}(\boldsymbol{w}_\lambda) + \sqrt{R_p^{\mathrm{MSE}}(\boldsymbol{w}_\lambda)} + \rho_\Delta t, \tag{64}$$

$$\leq C_{\mathrm{br}}^{\mathrm{RMSE}}(\boldsymbol{w}_\lambda) + \sqrt{1 + \rho_\Delta^2/a}\,\sqrt{R_p^{\mathrm{MSE}}(\boldsymbol{w}_\lambda) + at^2}, \tag{65}$$

$$\leq C_{\mathrm{br}}^{\mathrm{RMSE}}(\boldsymbol{w}_\lambda) + \sqrt{(1 + \rho_\Delta^2/a)A}. \tag{66}$$

The middle line is Cauchy–Schwarz for vectors $(\sqrt{R_p^{\mathrm{MSE}}(\boldsymbol{w}_\lambda)}, \sqrt{a}\,t)$ and $(1, \rho_\Delta/\sqrt{a})$; the last uses (63). Squaring proves (62). □

For completeness, these finite-dimensional population MSE minimizers exist under the stated second moments. Each objective is a nonnegative quadratic form minus a linear term plus a constant. Its linear coefficient is orthogonal to every null vector of the quadratic form: a null direction has zero source score almost surely, hence zero covariance with the square-integrable label. Restricting to the range gives a positive definite quadratic with an attained minimum; null directions leave its value unchanged. This also applies within the linear subspaces defining stable and hard-GRIT predictors.

The underselection obstruction persists for regression. For a fixed coefficient with $c = \|P_{\mathcal{U}}\boldsymbol{w}\|_2 > 0$, translate source features by $tP_{\mathcal{U}}\boldsymbol{w}/c$ while holding the label fixed. The target residual becomes $\boldsymbol{w}^\top\mathbf{z} - y + tc$, so $R^{\mathrm{MSE}}_{q_t}(\boldsymbol{w}) = R^{\mathrm{MSE}}_{p}(\boldsymbol{w}) + 2tc\,\mathbb{E}_p(\boldsymbol{w}^\top\mathbf{z} - y) + t^2c^2 \to \infty$. The stable-label joint distribution stays fixed and each finite-t target has finite moments. As in classification, a restriction on the feature-map image or deployment radius can restrict these constructed targets.

B Subspace Estimation and Statistical Optimality

B.1 Subspace estimation from relative covariance blocks

The pair covariance contains fluctuations within each subspace and across the two subspaces. Controlling these blocks relative to the clean signal gives a subspace bound that depends on the weakest clean eigenvalue. Throughout this subsection, let $m = r_*$, $s = d - m$, $\sigma = \sigma_\xi$, and $g = \lambda_{r_*,k} > 0$. Restrictions and inverse square roots of $\boldsymbol{\Sigma}_k$ act on $\mathcal{U}$ only. All probabilities condition on the complete clean signal collection.

Lemma 4 (Rotation from relative blocks)**.** *Let $\boldsymbol{H}$ be symmetric, with blocks $\left(\begin{smallmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{B}^\top & \boldsymbol{D}\end{smallmatrix}\right)$ relative to $\mathcal{U} \oplus \mathcal{S}$, where $\dim\mathcal{U} = m$. If $\delta = \lambda_{\min}(\boldsymbol{A}) - \lambda_{\max}(\boldsymbol{D}) > 0$, its leading rank-$m$ projector Q is unique and*

$$\|Q - P_{\mathcal{U}}\|_{\mathrm{op}} \le \sin\left[\frac{1}{2}\arctan\left(\frac{2\|\boldsymbol{B}\|_{\mathrm{op}}}{\delta}\right)\right]. \tag{67}$$

In particular, suppose $\boldsymbol{H} = \boldsymbol{\Sigma}_k + \sigma^2\boldsymbol{I} + \boldsymbol{E}$ and define

$$\begin{aligned} \varepsilon &= \|\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1/2}\boldsymbol{E}_{UU}\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1/2}\|_{\mathrm{op}}, \quad d_S = \|\boldsymbol{E}_{SS}\|_{\mathrm{op}},\\ b &= \|\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1/2}\boldsymbol{E}_{US}\|_{\mathrm{op}}, \qquad\qquad \zeta = 1 - \varepsilon - d_S/g, \end{aligned} \tag{68}$$

where $\boldsymbol{\Sigma}_{k,\mathcal{U}} = \boldsymbol{\Sigma}_k|_{\mathcal{U}}$. If $\zeta > 0$, then

$$\|Q - P_{\mathcal{U}}\|_{\mathrm{op}} \le \sin\left[\frac{1}{2}\arctan\left(\frac{2b}{\sqrt{g}\,\zeta}\right)\right]. \tag{69}$$

Proof. The proof compresses the eigenvector equation to a plane attaining the largest principal angle. Diagonal block ordering keeps that angle below $\pi/4$; retaining its directional signal energy gives the relative bound.

Min–max on $\mathcal{U}$ and $\mathcal{S}$ gives $\lambda_m(\boldsymbol{H}) \ge \lambda_{\min}(\boldsymbol{A})$ and $\lambda_{m+1}(\boldsymbol{H}) \le \lambda_{\max}(\boldsymbol{D})$. The rank-$m$ gap therefore remains positive along the continuous path replacing $\boldsymbol{B}$ by $t\boldsymbol{B}$, $0 \le t \le 1$. Its leading projector and largest principal angle θ vary continuously from $P_{\mathcal{U}}$ and zero. For $0 < \theta < \pi/2$, take unit $\boldsymbol{u} \in \mathcal{U}$ and $\boldsymbol{v} \in \mathcal{S}$ with

$$\boldsymbol{x} = (\cos\theta)\boldsymbol{u} + (\sin\theta)\boldsymbol{v} \in \mathrm{range}(Q), \qquad \boldsymbol{y} = -(\sin\theta)\boldsymbol{u} + (\cos\theta)\boldsymbol{v} \in \ker Q. \tag{70}$$

Such a pair follows by maximizing $\|P_{\mathcal{S}}\boldsymbol{x}\|^2$ over unit $\boldsymbol{x} \in \mathrm{range}(Q)$. The compression equation $QP_{\mathcal{S}}\boldsymbol{x} = (\sin^2\theta)\boldsymbol{x}$ then gives $Q\boldsymbol{u} = (\cos\theta)\boldsymbol{x}$ and $Q\boldsymbol{v} = (\sin\theta)\boldsymbol{x}$. Invariance of the two spectral spaces yields

$$(\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u} - \boldsymbol{v}^\top\boldsymbol{D}\boldsymbol{v})\sin\theta\cos\theta = \cos(2\theta)\,\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{v}. \tag{71}$$

The left side is positive at $\theta = \pi/4$, while the right side is zero. Continuity along the path thus gives $\theta < \pi/4$. Dividing (71) proves $\tan(2\theta) \le 2\|\boldsymbol{B}\|_{\mathrm{op}}/\delta$. Let V, W be orthonormal bases for $\mathcal{U}$ and $\mathrm{range}(Q)$. The equal-rank projector identity gives

$$\begin{aligned} \|Q - P_{\mathcal{U}}\|_{\mathrm{op}} &= \|P_{\mathcal{U}}(\boldsymbol{I} - Q)\|_{\mathrm{op}} = \|(\boldsymbol{I} - P_{\mathcal{U}})Q\|_{\mathrm{op}}\\ &= \sqrt{1 - s_{\min}(V^\top W)^2} = \sin\theta. \end{aligned}$$

Each leakage norm squared equals $1-s_{\min}(V^\top W)^2$. Relative to $\mathcal{U}\oplus\mathcal{S}$, $(Q-P_{\mathcal{U}})^2$ is block diagonal with blocks $P_{\mathcal{U}}(\boldsymbol{I}-Q)P_{\mathcal{U}}$ and $(\boldsymbol{I}-P_{\mathcal{U}})Q(\boldsymbol{I}-P_{\mathcal{U}})$, whose norms are those leakage squares. This proves (67), including $\theta=0$.

For the relative statement, put $t=\boldsymbol{u}^\top\boldsymbol{\Sigma}_k\boldsymbol{u}\geq g$. The definitions in (68) imply

$$\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}-\boldsymbol{v}^\top\boldsymbol{D}\boldsymbol{v}\geq(1-\varepsilon)t-d_S>0, \tag{72}$$

$$|\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{v}|\leq b\sqrt{t}, \tag{73}$$

$$\tan(2\theta)\leq\frac{2b\sqrt{t}}{(1-\varepsilon)t-d_S}\leq\frac{2b}{\sqrt{g}\,\zeta}. \tag{74}$$

The last inequality uses that $\sqrt{t}/[(1-\varepsilon)t-d_S]$ decreases on $t\geq g$ when $\varepsilon<1$ and $d_S\geq 0$. Taking the half-arctangent and sine proves (69). □

The ordered-block bound is exact for a two-dimensional matrix with diagonal gap δ and off-diagonal entry $\|\boldsymbol{B}\|_{\mathrm{op}}$. Adding a scalar multiple of the identity makes this matrix positive semidefinite without changing its eigenspaces. In particular, a zero cross block gives exact recovery whenever the diagonal blocks remain ordered.

Lemma 5 (Relative block concentration)**.** *Under A2, define*

$$T_k=\left(\frac{\sigma}{\sqrt{g}}+\frac{\sigma^2}{g}\right)\sqrt{\frac{d+\log(1/\eta)}{k}}+\frac{\sigma^2}{g}\frac{d+\log(1/\eta)}{k},\qquad\eta\in(0,1). \tag{75}$$

Let $\boldsymbol{E}=\widetilde{\boldsymbol{\Sigma}}_k-(\boldsymbol{\Sigma}_k+\sigma^2\boldsymbol{I})$ and define its block quantities by (68). There is a constant A_K, depending only on K, such that with conditional probability at least $1-\eta$,

$$\varepsilon+d_S/g\leq A_K T_k,\qquad b/\sqrt{g}\leq A_K T_k. \tag{76}$$

Proof. The proof rescales the clean unstable block to identity covariance, controls five scalar sums, and then takes finite nets. This retains the common signal energy in the cross block and its spectral denominator.

In orthonormal coordinates of $\mathcal{U}$ and $\mathcal{S}$, put $\boldsymbol{a}_j=\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1/2}\boldsymbol{u}_j$, $\boldsymbol{\zeta}_j=\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1/2}P_{\mathcal{U}}\boldsymbol{\xi}_j$, and $\boldsymbol{\nu}_j=P_{\mathcal{S}}\boldsymbol{\xi}_j$. Then

$$\frac{1}{k}\sum_j\boldsymbol{a}_j\boldsymbol{a}_j^\top=\boldsymbol{I}_m,\qquad\mathbb{E}\boldsymbol{\zeta}_j\boldsymbol{\zeta}_j^\top=\sigma^2\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1},\qquad\mathbb{E}\boldsymbol{\nu}_j\boldsymbol{\nu}_j^\top=\sigma^2\boldsymbol{I}_s,\qquad\mathbb{E}\boldsymbol{\zeta}_j\boldsymbol{\nu}_j^\top=0. \tag{77}$$

The directional ψ_2 bounds are $K\sigma/\sqrt{g}$ for $\boldsymbol{\zeta}_j$ and $K\sigma$ for $\boldsymbol{\nu}_j$. Expanding the required blocks gives

$$\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1/2}\boldsymbol{E}_{UU}\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1/2}=\frac{1}{k}\sum_j\left(\boldsymbol{a}_j\boldsymbol{\zeta}_j^\top+\boldsymbol{\zeta}_j\boldsymbol{a}_j^\top+\boldsymbol{\zeta}_j\boldsymbol{\zeta}_j^\top-\sigma^2\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1}\right), \tag{78}$$

$$\boldsymbol{E}_{SS}=\frac{1}{k}\sum_j(\boldsymbol{\nu}_j\boldsymbol{\nu}_j^\top-\sigma^2\boldsymbol{I}_s), \tag{79}$$

$$\boldsymbol{\Sigma}_{k,\mathcal{U}}^{-1/2}\boldsymbol{E}_{US}=\frac{1}{k}\sum_j(\boldsymbol{a}_j\boldsymbol{\nu}_j^\top+\boldsymbol{\zeta}_j\boldsymbol{\nu}_j^\top). \tag{80}$$

Isotropy centers the last noise product even when the coordinates of a noise vector are dependent.

For fixed unit $\boldsymbol{x}\in\mathbb{R}^m$ and $\boldsymbol{y}\in\mathbb{R}^s$, the squared sub-Gaussian scales of the two signal–noise sums satisfy

$$\sum_j\|2(\boldsymbol{x}^\top\boldsymbol{a}_j)(\boldsymbol{x}^\top\boldsymbol{\zeta}_j)\|_{\psi_2}^2\leq 4K^2\sigma^2k/g, \tag{81}$$

$$\sum_j\|(\boldsymbol{x}^\top\boldsymbol{a}_j)(\boldsymbol{y}^\top\boldsymbol{\nu}_j)\|_{\psi_2}^2\leq K^2\sigma^2k. \tag{82}$$

Both inequalities use $\sum_j(\boldsymbol{x}^\top\boldsymbol{a}_j)^2=k$. The centered noise squares in (78) and (79) have ψ_1 scales at most $CK^2\sigma^2/g$ and $CK^2\sigma^2$, respectively. The scalar noise product in (80) has scale at

most $CK^2\sigma^2/\sqrt{g}$. For the product bound, use $|XY|/(ab) \le X^2/(2a^2) + Y^2/(2b^2)$ and Cauchy–Schwarz on the exponential when the factors have ψ_2 bounds a, b; independence of the factors is unnecessary. Centering changes the ψ_1 scale by at most a factor two.

Scalar sub-Gaussian concentration bounds a centered sum with squared scale total V by $C\sqrt{Vh}$. Scalar Bernstein bounds the average of k centered variables of common ψ_1 scale L by $CL(\sqrt{h/k} + h/k)$. Each bound has failure probability at most $2e^{-h}$; these are the independent-sum inequalities in Vershynin (2018, Theorem 2.6.2, p. 29, and Theorem 2.8.1, p. 36). Apply them to the five scalar sums above and take $1/4$-nets of the two unit spheres, of sizes at most 9^m and 9^s. The operator norm is at most twice the maximal net quadratic form for a symmetric matrix, or twice the maximal net bilinear form for a rectangular matrix: approximating maximizing unit vectors changes the form by at most half the operator norm. With $h = d\log 9 + \log(12/\eta)$, the union bound fails with probability at most $(4\cdot 9^m + 2\cdot 9^s + 4\cdot 9^{m+s})e^{-h} \le 5\eta/6$. On the remaining event,

$$\begin{aligned} \varepsilon &\le C\left[K\frac{\sigma}{\sqrt{g}}\sqrt{\frac{h}{k}} + K^2\frac{\sigma^2}{g}\left(\sqrt{\frac{h}{k}} + \frac{h}{k}\right)\right], \\ d_S &\le CK^2\sigma^2\left(\sqrt{\frac{h}{k}} + \frac{h}{k}\right), \\ b &\le C\left[K\sigma\sqrt{\frac{h}{k}} + K^2\frac{\sigma^2}{\sqrt{g}}\left(\sqrt{\frac{h}{k}} + \frac{h}{k}\right)\right]. \end{aligned} \tag{83}$$

Since $d \ge 2$, $h \le C'[d + \log(1/\eta)]$. Absorbing the universal constants and K, K^2 proves (76). If $\sigma = 0$, all noises vanish almost surely and the same conclusion holds with zero radii. □

Theorem 6 (Finite-pair subspace rate)**.** *Under A2, let T_k be given by (75). There is a constant C_K, depending only on K, such that with conditional probability at least $1-\eta$,*

$$\begin{aligned} \|P_{k,r_*} - P_{\mathcal{U}}\|_{\mathrm{op}} \le \gamma_k, \quad \gamma_k = \min\{1, C_K T_k\}, \\ \|P_{\mathcal{U}}(\boldsymbol{I} - P_{k,r})\|_{\mathrm{op}} \le \gamma_k \qquad \text{for every } r_* \le r \le d. \end{aligned} \tag{84}$$

All projectors use one ordered eigenbasis and a fixed measurable tie rule.

Proof. Relative block concentration gives accurate rotation when its scale is small; the universal projector bound covers the remaining regime. On (76), if $A_K T_k < 1/2$, then $\zeta > 1/2$ and Lemma 4 gives

$$\|P_{k,r_*} - P_{\mathcal{U}}\|_{\mathrm{op}} \le \frac{b}{\sqrt{g}\,\zeta} \le 2A_K T_k, \tag{85}$$

using $\sin[\frac{1}{2}\arctan(2x)] \le x$ for $x \ge 0$. If $A_K T_k \ge 1/2$, the norm of the difference of two orthogonal projectors is at most one. Taking $C_K = 2A_K$ proves the first inequality in (84) in both regimes. Nesting gives $(\boldsymbol{I} - P_{k,r_*})(\boldsymbol{I} - P_{k,r}) = \boldsymbol{I} - P_{k,r}$, so

$$\|P_{\mathcal{U}}(\boldsymbol{I} - P_{k,r})\|_{\mathrm{op}} \le \|P_{\mathcal{U}}(\boldsymbol{I} - P_{k,r_*})\|_{\mathrm{op}} = \|P_{k,r_*} - P_{\mathcal{U}}\|_{\mathrm{op}}. \tag{86}$$

This proves the simultaneous conclusion. □

When $k \ge d + \log(1/\eta)$, the linear term in T_k is absorbed by its square-root term. The full expression applies to every A2 design, including $k < d$. The rescaling in the proof is an analysis of the unknown clean covariance; it does not alter the GRIT estimator.

Sharp control from a full perturbation norm.

Proposition 9 (Sharp projector envelope)**.** *Let $\boldsymbol{M} = \boldsymbol{\Sigma}_k + \sigma^2\boldsymbol{I}$ and let $\boldsymbol{E}$ be symmetric with $\|\boldsymbol{E}\|_{\mathrm{op}} \le e < g/2$. The leading rank-$r_*$ projector Q of $\boldsymbol{M} + \boldsymbol{E}$ is unique and*

$$\|Q - P_{\mathcal{U}}\|_{\mathrm{op}} \le \sin\left[\frac{1}{2}\arcsin(2e/g)\right]. \tag{87}$$

For every $e \in [0, g/2)$, a positive-semidefinite two-dimensional example attains equality.

Proof. The proof bounds the off-diagonal entry in a principal-angle plane and constructs the perturbation that exactly cancels it. Min–max gives a gap at least $g - 2e > 0$ along $\boldsymbol{M} + t\boldsymbol{E}$, $0 \le t \le 1$. For the principal pair (70), spectral invariance yields

$$0 = \boldsymbol{x}^\top(\boldsymbol{M} + \boldsymbol{E})\boldsymbol{y} = -(\sin\theta\cos\theta)\boldsymbol{u}^\top\boldsymbol{\Sigma}_k\boldsymbol{u} + \boldsymbol{x}^\top\boldsymbol{E}\boldsymbol{y}, \qquad g\sin\theta\cos\theta \le e. \tag{88}$$

The continuous angle path cannot reach $\pi/4$, because that would require $g/2 \le e$. Thus $\sin(2\theta) \le 2e/g$ on the increasing branch, proving (87).

For attainment, choose $0 \le \theta < \pi/4$, $c = \cos\theta$, $h = \sin\theta$, $\boldsymbol{x} = (c, h)^\top$, and $\boldsymbol{y} = (-h, c)^\top$. Take

$$\boldsymbol{M} = \sigma^2\boldsymbol{I} + \mathrm{diag}(g, 0), \qquad \boldsymbol{M} + \boldsymbol{E} = \sigma^2\boldsymbol{I} + gc^2\boldsymbol{x}\boldsymbol{x}^\top + gh^2\boldsymbol{y}\boldsymbol{y}^\top. \tag{89}$$

In the $(\boldsymbol{x}, \boldsymbol{y})$ basis the perturbation is $\left(\begin{smallmatrix} 0 & gch \\ gch & 0 \end{smallmatrix}\right)$, so $\|\boldsymbol{E}\|_{\mathrm{op}} = gch$. The perturbed matrix is positive semidefinite, its gap is $g\cos(2\theta) > 0$, and its projector error is $\sin\theta$. The map $\theta \mapsto g\sin\theta\cos\theta$ covers $[0, g/2)$, proving the claim. □

B.2 Minimax estimation of Gaussian pair geometry

Under independent isotropic Gaussian pair noise and known rank, GRIT's spectral estimator attains the minimax rate for expected squared subspace error over the clean-signal class specified below. The following corollary specializes the singular-space estimation results of Cai & Zhang (2018) to the pair matrix.

Corollary 7 (Gaussian pair-geometry minimax rate)**.** *Observe $\boldsymbol{D} = \boldsymbol{L} + \sigma\boldsymbol{G} \in \mathbb{R}^{d\times k}$, where $\boldsymbol{G}$ has independent standard normal entries and $\sigma, g > 0$. Suppose the rank r_* is known, with $1 \le r_* \le \min(d, k)/16$, and define*

$$\mathcal{L}_{k,g} = \{\boldsymbol{L} \in \mathbb{R}^{d\times k} : \mathrm{rank}(\boldsymbol{L}) = r_*,\ s_{r_*}(\boldsymbol{L})^2 \ge kg\}, \qquad \mathfrak{r}_{k,g,\sigma} = \min\left\{1, \frac{d}{k}\left(\frac{\sigma^2}{g} + \frac{\sigma^4}{g^2}\right)\right\}, \tag{90}$$

where s_{r_*} denotes the r_*th singular value. Let $P_{\boldsymbol{L}}$ project onto $\mathrm{col}(\boldsymbol{L})$, and let $\widehat{P}_{\mathrm{GRIT}}$ project onto the leading r_* eigenvectors of $\boldsymbol{D}\boldsymbol{D}^\top/k$. There are universal constants $c_0, C_0 > 0$ such that

$$c_0\mathfrak{r}_{k,g,\sigma} \le \inf_{\widehat{P}} \sup_{\boldsymbol{L}\in\mathcal{L}_{k,g}} \mathbb{E}\|\widehat{P} - P_{\boldsymbol{L}}\|_{\mathrm{op}}^2 \le \sup_{\boldsymbol{L}\in\mathcal{L}_{k,g}} \mathbb{E}\|\widehat{P}_{\mathrm{GRIT}} - P_{\boldsymbol{L}}\|_{\mathrm{op}}^2 \le C_0\mathfrak{r}_{k,g,\sigma}. \tag{91}$$

The infimum ranges over measurable rank-r_* orthogonal-projector estimators based on $\boldsymbol{D}$ and the supplied rank.

Proof. Rescale the pair matrix to unit noise and apply the upper and lower singular-space bounds of Cai & Zhang (2018, Theorems 3–4, pp. 70–71). Their left-space squared spectral-sine rate for a $d \times k$ rank-r_* matrix with least singular value at least t is $\min\{1, d(t^2 + k)/t^4\}$. Set $t = \sqrt{kg}/\sigma$ for $\boldsymbol{D}/\sigma$ to obtain

$$\frac{d(t^2 + k)}{t^4} = \frac{d}{k}\left(\frac{\sigma^2}{g} + \frac{\sigma^4}{g^2}\right). \tag{92}$$

The stated rank restriction satisfies the lower theorem's dimension conditions. The spectral-sine loss equals the operator norm of the difference of the corresponding rank-r_* projectors. Finally, the leading left singular space of $\boldsymbol{D}$ is exactly the leading eigenspace of $\boldsymbol{D}\boldsymbol{D}^\top/k$. These identifications give all three inequalities in (91). □

C Logistic Optimizer and Transport Refinements

The main bounds retain a simple direct-shift term. This appendix gives optional refinements specific to population logistic fitting, and derives the improved quadratic-consistency coefficients without additional curvature assumptions. Throughout, $\varphi(t) = \log(1 + e^{-t})$ and source feature second moments are finite. These moments justify differentiating under the expectation and give

$$\nabla^2 R_p(\boldsymbol{w}) = \mathbb{E}_p \frac{\mathbf{z}\mathbf{z}^\top}{4\cosh^2(\boldsymbol{w}^\top\mathbf{z}/2)}, \qquad 0 \preceq \nabla^2 R_p(\boldsymbol{w}) \preceq \tfrac{1}{4}\mathbb{E}_p[\mathbf{z}\mathbf{z}^\top]. \tag{93}$$

C.1 A SOURCE-DEPENDENT DIRECT-SHIFT COEFFICIENT

For a finite coefficient $\boldsymbol{w}$, define $\boldsymbol{u} = P_{\mathcal{U}}\boldsymbol{w}$, $m = y\boldsymbol{w}^\top \mathbf{z}$, and

$$\beta(m) = (1+e^m)^{-1}, \qquad B_p(\boldsymbol{w}) = \{\mathbb{E}_p \beta(m)^2\}^{1/2}. \tag{94}$$

Let $\mathcal{Q}_\rho(p)$ contain the feature-space targets admitting a coupling with p that preserves $(P_{\mathcal{S}}\mathbf{z}, y)$ and satisfies $\mathbb{E}\Delta\Delta^\top \preceq \rho^2 P_{\mathcal{U}}$. This is a class of distributions on features; it does not impose image constraints from a prescribed nonlinear feature extractor.

Proposition 10 (Locally sharp logistic transport)**.** *For $\rho \geq 0$, put $a = \rho\|\boldsymbol{u}\|_2$. Then*

$$aB_p(\boldsymbol{w}) \leq \sup_{q \in \mathcal{Q}_\rho(p)} \{R_q(\boldsymbol{w}) - R_p(\boldsymbol{w})\} \leq \min\{a,\ aB_p(\boldsymbol{w}) + a^2/8\}. \tag{95}$$

In particular, the right derivative of the supremum at $\rho = 0$ is $\|\boldsymbol{u}\|_2 B_p(\boldsymbol{w})$.

Proof. The proof bounds the second-order loss remainder and constructs a displacement aligned with the first-order loss derivative. For an admissible coupling put $h = y\boldsymbol{w}^\top\Delta$. Its preserved stable projection implies $\Delta \in \mathcal{U}$. Since $0 \leq \varphi'' \leq 1/4$,

$$R_q(\boldsymbol{w}) - R_p(\boldsymbol{w}) \leq -\mathbb{E}[\beta(m)h] + \tfrac{1}{8}\mathbb{E}h^2, \tag{96}$$

$$\mathbb{E}h^2 \leq \rho^2\|\boldsymbol{u}\|_2^2 = a^2, \tag{97}$$

$$-\mathbb{E}[\beta(m)h] \leq B_p(\boldsymbol{w})(\mathbb{E}h^2)^{1/2} \leq aB_p(\boldsymbol{w}). \tag{98}$$

The first line is Taylor's integral bound, the second is the displacement moment bound, and the third is Cauchy–Schwarz. The unit-Lipschitz bound supplies the alternative upper bound a.

If $a > 0$, finite margins give $B_p(\boldsymbol{w}) > 0$. Define the conditional displacement

$$\Delta = -\frac{y\, a\beta(m)}{B_p(\boldsymbol{w})\|\boldsymbol{u}\|_2^2}\boldsymbol{u}. \tag{99}$$

It preserves the stable features and label, has finite second moments, and satisfies $\mathbb{E}\Delta\Delta^\top = \rho^2\boldsymbol{u}\boldsymbol{u}^\top/\|\boldsymbol{u}\|_2^2 \preceq \rho^2 P_{\mathcal{U}}$. The margin becomes $m - a\beta(m)/B_p(\boldsymbol{w})$. Convexity gives a loss increase at least $a\beta(m)^2/B_p(\boldsymbol{w})$, whose expectation proves the lower bound. The bounds squeeze the right derivative. If $a = 0$, every admissible displacement preserves the score and both bounds are zero. □

Apply the upper bound to the source-to-bridge coupling and add the exact $C_{\mathrm{br}}(\boldsymbol{w})$ for a refinement of Theorem 1. The lower bound allows its target to depend on the fixed coefficient; it is not a lower bound for every specified target. No pair assumption is used here.

Proposition 11 (Radial stationarity reduces the local coefficient)**.** *Suppose $\boldsymbol{C} \succeq 0$, $\lambda \geq 0$, and*

$$\boldsymbol{w}^\top \nabla R_p(\boldsymbol{w}) + 2\lambda \boldsymbol{w}^\top \boldsymbol{C}\boldsymbol{w} = 0. \tag{100}$$

Then

$$B_p(\boldsymbol{w})^2 \leq \tfrac{1}{4} - \lambda \boldsymbol{w}^\top \boldsymbol{C}\boldsymbol{w} \leq \tfrac{1}{4}. \tag{101}$$

Every exact population hard-GRIT optimizer satisfies this condition with $\lambda = 0$, and every exact unconstrained quadratic-CR optimizer satisfies it with its penalty matrix and strength. Their direct-shift contribution is at most

$$\min\{a,\ a/2 + a^2/8\}, \qquad a = \rho_\Delta \|P_{\mathcal{U}}\boldsymbol{w}\|_2. \tag{102}$$

Proof. A scalar inequality converts the radial first-order condition into a moment bound for the logistic derivative. For $t \in (0,1)$ set $F(t) = t\log((1-t)/t) - 1/2 + 2t^2$. Direct differentiation gives

$$F(1/2) = F'(1/2) = 0, \qquad F''(t) = 4 - \frac{1}{t(1-t)^2} < 0, \tag{103}$$

where $t(1-t)^2 \leq 4/27$ proves strict negativity. Consequently $F(t) \leq 0$. Substitute $t = \beta(m)$, so $\log((1-t)/t) = m$, to obtain

$$\beta(m)^2 \leq \tfrac{1}{4} - \tfrac{1}{2}m\beta(m), \qquad \mathbb{E}_p[m\beta(m)] = -\boldsymbol{w}^\top \nabla R_p(\boldsymbol{w}) = 2\lambda \boldsymbol{w}^\top \boldsymbol{C}\boldsymbol{w}. \tag{104}$$

Taking expectations proves (101). For a hard optimizer, its feasible set is a linear subspace containing the entire line through $\boldsymbol{w}$, so differentiating risk along that line proves the radial condition. Unconstrained CR stationarity proves the other case. Proposition 10 gives (102); for CR its coefficient $1/2$ can be replaced by $\sqrt{1/4 - \lambda \boldsymbol{w}^\top \boldsymbol{C}\boldsymbol{w}}$. □

Sharpness and optimization scope. The coefficient $1/2$ cannot be decreased uniformly for learned CR predictors. Let source features be $(0, X)$ with X a fair sign, take $\boldsymbol{C} = \mathrm{diag}(0, c)$ with $c > 0$, and fix $\lambda > 0$. For small $b > 0$, generate $y = X$ with probability $s(b) + 2\lambda cb$, where $s(b) = (1 + e^{-b})^{-1}$, and otherwise set $y = -X$. The fitted second coefficient is b because its stationarity equation is $s(w_2) - \mathbb{P}(y = X) + 2\lambda c w_2 = 0$; choose first coefficient zero. The conditional map $\Delta = -y\rho\boldsymbol{e}_2$ preserves stable features and labels and has radius ρ. Since all source margins tend to zero,

$$\frac{R_q(b\boldsymbol{e}_2) - R_p(b\boldsymbol{e}_2)}{\rho b} \longrightarrow \tfrac{1}{2} \qquad (b \downarrow 0,\ \rho > 0 \text{ fixed}). \tag{105}$$

The penalty matrix is realized exactly by noiseless A2 differences $\sqrt{c}\,\boldsymbol{e}_2$. This local example does not establish joint sharpness of the full CR target-risk certificate. The quadratic correction in (102) cannot simply be omitted at large radius: $\varphi(-a) - \varphi(0)$ eventually grows with slope one. An empirical optimizer only supplies an empirical radial identity. If the population radial residual is $r = \boldsymbol{w}^\top \nabla(R_p + \lambda \boldsymbol{w}^\top \boldsymbol{C}\boldsymbol{w})$, the valid replacement in (101) adds $r/2$ on the right.

C.2 Quadratic consistency from exact optimizer stationarity

For this subsection, abbreviate $R = R_p$, $\rho = \rho_\Delta$, $\sigma = \sigma_\xi$, and $\boldsymbol{v} = \boldsymbol{w}^*_{\mathcal{S}}$. On a covariance event, let

$$\boldsymbol{C} = \widetilde{\boldsymbol{\Sigma}}_k = \boldsymbol{M} + \boldsymbol{E}, \qquad \boldsymbol{M} = \boldsymbol{\Sigma}_k + \sigma^2 \boldsymbol{I}, \qquad \|\boldsymbol{E}\|_{\mathrm{op}} \le e, \tag{106}$$

where $\boldsymbol{\Sigma}_k P_{\mathcal{S}} = 0$ and $\boldsymbol{\Sigma}_k \succeq g P_{\mathcal{U}}$, $g > 0$. Let $\boldsymbol{w}$ be an existing exact unconstrained minimizer of $R(\boldsymbol{w}) + \lambda \boldsymbol{w}^\top \boldsymbol{C}\boldsymbol{w}$, $\lambda > 0$. The pair covariance $\boldsymbol{C}$ is positive semidefinite. The mean matrix, clean gap, covariance error, and stable oracle are analysis quantities.

Proposition 12 (Exact optimizer comparison)**.** *For any differentiable convex risk R, positive semidefinite $\boldsymbol{C}$, and existing exact penalized optimizer $\boldsymbol{w}$, define $D_R(\boldsymbol{v}, \boldsymbol{w}) = R(\boldsymbol{v}) - R(\boldsymbol{w}) - \nabla R(\boldsymbol{w})^\top (\boldsymbol{v} - \boldsymbol{w})$. For every comparator $\boldsymbol{v}$,*

$$\begin{aligned} R(\boldsymbol{w}) - R(\boldsymbol{v}) &= 2\lambda \boldsymbol{w}^\top \boldsymbol{C}(\boldsymbol{v} - \boldsymbol{w}) - D_R(\boldsymbol{v}, \boldsymbol{w}) \\ &= \lambda\{\boldsymbol{v}^\top \boldsymbol{C}\boldsymbol{v} - \boldsymbol{w}^\top \boldsymbol{C}\boldsymbol{w} - (\boldsymbol{v} - \boldsymbol{w})^\top \boldsymbol{C}(\boldsymbol{v} - \boldsymbol{w})\} - D_R(\boldsymbol{v}, \boldsymbol{w}). \end{aligned} \tag{107}$$

Both $D_R(\boldsymbol{v}, \boldsymbol{w})$ and the omitted quadratic in ordinary objective comparison are nonnegative.

Proof. Substitute the unconstrained first-order condition $\nabla R(\boldsymbol{w}) = -2\lambda \boldsymbol{C}\boldsymbol{w}$ into the definition of D_R. Expanding the quadratic difference gives the second line. Convexity of R and positive semidefiniteness of $\boldsymbol{C}$ give the stated signs. □

Proposition 13 (Positive CR bound from stationarity)**.** *Under (106), with a stable source optimum $\boldsymbol{v}$, and the moment and conditional-shift assumptions of Theorem 1,*

$$\begin{aligned} R_q(\boldsymbol{w}) \le R^*_{p,\mathcal{S}} + C_{\mathrm{br}}(\boldsymbol{w}) + \frac{\lambda\sigma^2}{2}\|\boldsymbol{v}\|_2^2 + \frac{\rho^2}{8\lambda(g + \sigma^2)} \\ + 2\lambda e \|\boldsymbol{w}\|_2 \|\boldsymbol{v} - \boldsymbol{w}\|_2. \end{aligned} \tag{108}$$

The statement holds simultaneously for every strength, existing optimizer, and stable optimum on the same covariance event.

Proof. Retain the first-order comparison in (107), split it into stable and unstable blocks, and complete two squares. Put $\boldsymbol{s} = P_{\mathcal{S}}\boldsymbol{w}$, $\boldsymbol{u} = P_{\mathcal{U}}\boldsymbol{w}$, and $t = \|\boldsymbol{u}\|_2$. Since $\boldsymbol{v} \in \mathcal{S}$, block orthogonality gives

$$\begin{aligned} R(\boldsymbol{w}) - R(\boldsymbol{v}) &\le 2\lambda\sigma^2 \boldsymbol{s}^\top(\boldsymbol{v} - \boldsymbol{s}) - 2\lambda \boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u} + 2\lambda \boldsymbol{w}^\top \boldsymbol{E}(\boldsymbol{v} - \boldsymbol{w}) \\ &\le 2\lambda\sigma^2 \boldsymbol{s}^\top(\boldsymbol{v} - \boldsymbol{s}) - 2\lambda(g + \sigma^2)t^2 + 2\lambda e\|\boldsymbol{w}\|_2\|\boldsymbol{v} - \boldsymbol{w}\|_2. \end{aligned} \tag{109}$$

The first line discards only the nonnegative remainder D_R; the covariance correction there may have either sign. The second uses the clean eigenvalue lower bound and the operator-norm covariance event. Complete the squares:

$$2\lambda\sigma^2 \boldsymbol{s}^\top(\boldsymbol{v} - \boldsymbol{s}) = \tfrac{1}{2}\lambda\sigma^2\|\boldsymbol{v}\|_2^2 - 2\lambda\sigma^2\|\boldsymbol{s} - \boldsymbol{v}/2\|_2^2, \tag{110}$$

$$\rho t - 2\lambda(g + \sigma^2)t^2 = \frac{\rho^2}{8\lambda(g + \sigma^2)} - 2\lambda(g + \sigma^2)\left(t - \frac{\rho}{4\lambda(g + \sigma^2)}\right)^2. \tag{111}$$

Add the bridge and direct-shift comparison, discard the displayed nonpositive squares, and use $R(\boldsymbol{v}) = R^*_{p,\mathcal{S}}$. The same covariance event gives the bound simultaneously for every strength and optimizer satisfying the stated conditions. The cases $\sigma = 0$ and $\rho = 0$ require no division by either quantity and are included. □

The covariance remainder vanishes at $\boldsymbol{w} = \boldsymbol{v}$; bounding the two quadratic forms separately would lose this cancellation. Its coefficient is not always smaller than that of ordinary objective comparison, so two valid *complete* certificates can be combined by taking their minimum. Halving the two displayed population tradeoff coefficients does not prove their joint logistic sharpness. If an approximate coefficient has residual $\boldsymbol{r} = \nabla R(\boldsymbol{w}) + 2\lambda \boldsymbol{C}\boldsymbol{w}$, identity (107) has the additional term $\boldsymbol{r}^\top(\boldsymbol{w} - \boldsymbol{v})$. A source-objective gap alone cannot be inserted into this stationarity proof as the same additive error.

C.3 Optional curvature and linear versus quadratic stable cost

Proposition 14 (One-segment curvature refinement)**.** *In addition to the preceding CR conditions, suppose*

$$D_R(\boldsymbol{v}, \boldsymbol{w}) \geq \tfrac{1}{2}\mu_{\mathcal{S}}\|P_{\mathcal{S}}(\boldsymbol{w} - \boldsymbol{v})\|_2^2 + \tfrac{1}{2}\mu_{\mathcal{U}}\|P_{\mathcal{U}}\boldsymbol{w}\|_2^2, \quad \mu_{\mathcal{S}}, \mu_{\mathcal{U}} \geq 0. \tag{112}$$

Then

$$\begin{aligned} R_q(\boldsymbol{w}) \leq R^*_{p,\mathcal{S}} + C_{\mathrm{br}}(\boldsymbol{w}) + \frac{2\lambda^2\sigma^4\|\boldsymbol{v}\|_2^2}{\mu_{\mathcal{S}} + 4\lambda\sigma^2} \\ + \frac{\rho^2}{2\{\mu_{\mathcal{U}} + 4\lambda(g + \sigma^2)\}} + 2\lambda e\|\boldsymbol{w}\|_2\|\boldsymbol{v} - \boldsymbol{w}\|_2. \end{aligned} \tag{113}$$

When $\mu_{\mathcal{S}} = \sigma = 0$, define the stable contribution to be zero.

Proof. Retain the assumed remainder in the exact comparison, thereby strengthening each quadratic before completing its square. With $\boldsymbol{a} = \boldsymbol{s} - \boldsymbol{v}$,

$$2\lambda\sigma^2\boldsymbol{s}^\top(\boldsymbol{v} - \boldsymbol{s}) - \tfrac{1}{2}\mu_{\mathcal{S}}\|\boldsymbol{a}\|_2^2 = -2\lambda\sigma^2\boldsymbol{v}^\top\boldsymbol{a} - \tfrac{1}{2}(\mu_{\mathcal{S}} + 4\lambda\sigma^2)\|\boldsymbol{a}\|_2^2 \tag{114}$$

$$\leq \frac{2\lambda^2\sigma^4\|\boldsymbol{v}\|_2^2}{\mu_{\mathcal{S}} + 4\lambda\sigma^2}, \tag{115}$$

$$\rho t - 2\lambda(g + \sigma^2)t^2 - \tfrac{1}{2}\mu_{\mathcal{U}}t^2 \leq \frac{\rho^2}{2\{\mu_{\mathcal{U}} + 4\lambda(g + \sigma^2)\}}. \tag{116}$$

Both inequalities maximize a concave quadratic. If the stable denominator is zero, its undivided expression is identically zero. Apply the same covariance and transport steps as in Proposition 13. □

Lemma 6 (A sufficient logistic curvature condition)**.** *If $\mathbb{E}_p[\mathbf{z}\mathbf{z}^\top] \succeq \kappa\boldsymbol{I}$, $\kappa > 0$, and $|\boldsymbol{v}^\top\mathbf{z}|, |\boldsymbol{w}^\top\mathbf{z}| \leq B$ source-almost surely, then* (112) *holds with*

$$\mu_{\mathcal{S}} = \mu_{\mathcal{U}} = \frac{\kappa}{4\cosh^2(B/2)}. \tag{117}$$

Proof. Bound every score on the segment between the two coefficients, and integrate the resulting Hessian lower bound. With $\boldsymbol{d} = \boldsymbol{w} - \boldsymbol{v}$, Taylor's integral identity in the required argument order is

$$D_R(\boldsymbol{v}, \boldsymbol{w}) = \int_0^1 t\,\boldsymbol{d}^\top\nabla^2 R(\boldsymbol{v} + t\boldsymbol{d})\boldsymbol{d}\,dt. \tag{118}$$

Every segment score has magnitude at most B. Equation (93) therefore lower-bounds its Hessian by $\kappa\boldsymbol{I}/(4\cosh^2(B/2))$. Integrating t over $[0, 1]$ and decomposing $\boldsymbol{d}$ orthogonally proves the assertion. More generally, a matrix lower bound $\mu_{\mathcal{S}}P_{\mathcal{S}} + \mu_{\mathcal{U}}P_{\mathcal{U}}$ along this segment suffices. Bounds on only the two diagonal Hessian blocks do not suffice, because the cross terms also affect the quadratic form. □

Bounded features and bounded endpoint coefficient norms imply the score bound. The moment lower bound excludes directions nearly absent from the source. These are optional sufficient conditions, not consequences of pair coverage. For finite endpoints and a positive-definite source second moment, the weaker source-dependent matrix

$$\mathbb{E}_p \frac{\mathbf{z}\mathbf{z}^\top}{4\cosh^2(\max\{|\boldsymbol{v}^\top\mathbf{z}|, |\boldsymbol{w}^\top\mathbf{z}|\}/2)} \tag{119}$$

is positive definite and lower-bounds the Hessian on the segment. Its positive weight gives a positive quadratic form in every nonzero direction. This does not supply a useful uniform eigenvalue lower bound: weak source directions and large, saturated logistic scores can make it arbitrarily small. Optimizing the penalized objective does not by itself prevent either phenomenon. Full-source stationarity of the stable oracle is unnecessary for Proposition 14.

Let $\tau = \lambda\sigma^2$. The stable contribution is $2\tau^2\|\boldsymbol{v}\|_2^2/(\mu_{\mathcal{S}} + 4\tau)$. A positive curvature lower bound uniform along the considered strengths makes it quadratic for small τ; setting $\mu_{\mathcal{S}} = 0$ recovers $\tau\|\boldsymbol{v}\|_2^2/2$. Only this stable contribution becomes quadratic; finite-pair and direct-shift contributions need not do so.

Proposition 15 (Fixed-source local penalty bias)**.** *Suppose a fixed differentiable convex risk is twice continuously differentiable near $\boldsymbol{v}$, with $\nabla R(\boldsymbol{v}) = 0$ and $\boldsymbol{H} = \nabla^2 R(\boldsymbol{v}) \succ 0$. For a fixed positive semidefinite $\boldsymbol{C}$, the penalized optimum is unique for all sufficiently small $\lambda > 0$ and*

$$\boldsymbol{w}_\lambda = \boldsymbol{v} - 2\lambda\boldsymbol{H}^{-1}\boldsymbol{C}\boldsymbol{v} + o(\lambda), \tag{120}$$

$$R(\boldsymbol{w}_\lambda) - R(\boldsymbol{v}) = 2\lambda^2(\boldsymbol{C}\boldsymbol{v})^\top\boldsymbol{H}^{-1}(\boldsymbol{C}\boldsymbol{v}) + o(\lambda^2). \tag{121}$$

Proof. Differentiate the optimizer equation and expand the risk at its stationary source optimum. The map $F(\boldsymbol{x}, \lambda) = \nabla R(\boldsymbol{x}) + 2\lambda\boldsymbol{C}\boldsymbol{x}$ vanishes at $(\boldsymbol{v}, 0)$ and has invertible derivative $\boldsymbol{H}$ in $\boldsymbol{x}$ there. The implicit function theorem gives a local stationary branch with derivative $-2\boldsymbol{H}^{-1}\boldsymbol{C}\boldsymbol{v}$. Convexity makes it a global optimum. The Hessian remains positive definite locally; a second global optimum would contradict strict convexity on the initial segment joining the two optima. Finally $R(\boldsymbol{v} + \boldsymbol{h}) - R(\boldsymbol{v}) = \boldsymbol{h}^\top\boldsymbol{H}\boldsymbol{h}/2 + o(\|\boldsymbol{h}\|_2^2)$ gives the displayed coefficient. It is strictly positive if $\boldsymbol{C}\boldsymbol{v} \neq 0$. □

A fixed logistic source with quadratic actual excess. Take independent fair signs X, T, features (X, T), and $\mathbb{P}(y = 1 \mid X, T) = s(bX)$ for fixed $b > 0$. Let $q = p$, $\mathcal{S} = \mathrm{span}(\boldsymbol{e}_1)$, and use the population pair matrix $\boldsymbol{C} = \mathrm{diag}(\sigma^2, g + \sigma^2)$ with $g, \sigma > 0$. This is an A2 conditional mean, rather than a claim of zero finite-sample fluctuation. The stable and full-source optimum is $(b, 0)$. Let $h = s(b)(1 - s(b)) > 0$. The source Hessian there is $h\boldsymbol{I}$, and Proposition 15 gives

$$R_p(\boldsymbol{w}_\lambda) - R^*_{p,\mathcal{S}} = \frac{2\sigma^4 b^2}{h}\lambda^2 + o(\lambda^2). \tag{122}$$

In contrast, the simple population certificate (108) adds $\lambda\sigma^2 b^2/2$, since both shift terms and the covariance error vanish. Its excess-to-actual-excess ratio therefore diverges. The full-source stationarity used here is stronger than stable optimality or A1. The fixed-source quantifier also matters for the following complementary example.

Proposition 16 (Linear stable cost over the allowed source class)**.** *There is a family of bounded-feature A1 source problems, with fixed nonzero stable oracle norm, zero shift, and an A2 population pair matrix, whose actual CR excess is of order λ. Consequently a source-independent $O(\lambda^2)$ replacement is not valid over this class.*

Proof. Let the source information weaken at the penalty scale. Fix $b > 0$ and, for $0 < \lambda \leq 1/8$, take independent fair signs V, T, set $X = \sqrt{8\lambda}\,V$, and use features (X, T) with $\mathbb{P}(y = 1 \mid X, T) = s(bX)$, $s(x) = (1 + e^{-x})^{-1}$. Set $q = p$, $\mathcal{S} = \mathrm{span}(\boldsymbol{e}_1)$, and use $\boldsymbol{C} = \mathrm{diag}(1, 2)$. This is the A2 conditional mean with noise variance one and clean covariance $\mathrm{diag}(0, 1)$; using the mean matrix does not claim that a random finite-pair covariance has zero fluctuation. The stable optimum is $(b, 0)$, and $\|\mathbf{z}\|_2 \leq \sqrt{2}$.

Symmetry in T, convexity, and the positive-definite penalty give a fitted coefficient $(c_\lambda, 0)$. Put $a = \sqrt{8\lambda}$ and $\pi = s(ab)$. The scalar source risk is $R(c) = \log(1 + e^{ac}) - \pi ac$, and stationarity gives

$$a\{s(ac_\lambda) - s(ab)\} + 2\lambda c_\lambda = 0. \tag{123}$$

Its left side is strictly increasing, negative at zero and positive at b, so $0 < c_\lambda < b$. The sigmoid expansion is uniform on that interval and gives

$$2(c_\lambda - b) + 2c_\lambda + O(\lambda) = 0, \qquad c_\lambda = b/2 + O(\lambda). \tag{124}$$

Uniform expansion of the scalar risk now gives

$$R(c) - R(b) = \tfrac{1}{8}a^2(c-b)^2 + O(a^4), \qquad R(c_\lambda) - R(b) = \tfrac{1}{4}b^2\lambda + O(\lambda^2). \tag{125}$$

This contradicts a $K\lambda^2$ bound with K independent of the source. The source varies with λ, so it does not contradict Proposition 15. □

Assumption scope. The optimizer comparison applies to differentiable convex risks with exact unconstrained population fitting. The CR target-risk bounds combine it with the stated pair-covariance and conditional-transport assumptions. The curvature refinement uses the additional one-segment condition in (112).

D FINITE-PAIR TIGHTNESS AND SHARPER RISK COMPARISONS

We connect subspace-estimation error to the risk of the predictor actually fitted by GRIT. The constructions use a fixed feature space and population source fitting. They distinguish an attained finite-pair rate, an asymptotically attained risk comparison, and the separate cost of removing additional directions.

D.1 A RISK COMPARISON RETAINING SOURCE GEOMETRY

Let $\varphi(t) = \log(1 + e^{-t})$. For a source distribution with finite second feature moments, define

$$\boldsymbol{C}_p = \mathbb{E}_p[\mathbf{z}\mathbf{z}^\top], \qquad L_z^2 = \|\boldsymbol{C}_p\|_{\mathrm{op}}, \qquad B_p(\boldsymbol{w}) = \left(\mathbb{E}_p(1 + e^{y\boldsymbol{w}^\top\mathbf{z}})^{-2}\right)^{1/2}. \tag{126}$$

These are distribution-dependent analysis quantities.

Lemma 7 (Source and direct-shift refinement)**.** *Let $\boldsymbol{v} \in \mathcal{W}_S^*$ be finite, let Q be an orthogonal projector, and let $\boldsymbol{w}$ be an existing source-risk minimizer on $\ker Q$. Suppose the bridge and conditional-displacement assumptions of Theorem 1 hold with radius ρ_Δ. Set $\boldsymbol{h} = Q\boldsymbol{v}$, $\boldsymbol{g}_* = \nabla R_p(\boldsymbol{v})$, and $b = \rho_\Delta \|P_{\mathcal{U}}\boldsymbol{w}\|_2$. Then*

$$\begin{aligned} R_q(\boldsymbol{w}) \le R_p(\boldsymbol{v}) + C_{\mathrm{br}}(\boldsymbol{w}) &+ \min\left\{\mathbb{E}_p|\mathbf{z}^\top\boldsymbol{h}|, -\boldsymbol{g}_*^\top\boldsymbol{h} + \tfrac{1}{8}\boldsymbol{h}^\top\boldsymbol{C}_p\boldsymbol{h}\right\} \\ &+ \min\left\{b,\ bB_p(\boldsymbol{w}) + \tfrac{1}{8}b^2\right\}. \end{aligned} \tag{127}$$

Proof. Project the stable oracle into the fitted feasible space. Compare its source risk using either logistic Lipschitzness or its Hessian upper bound, then retain the source-dependent first-order term in the direct-shift comparison.

Finite second moments justify differentiation under the expectation. Stable optimality and logistic curvature give

$$P_{\mathcal{S}}\boldsymbol{g}_* = 0, \qquad 0 \preceq \nabla^2 R_p(\boldsymbol{x}) \preceq \boldsymbol{C}_p/4 \quad \text{for every finite } \boldsymbol{x}. \tag{128}$$

The vector $\boldsymbol{v} - \boldsymbol{h} = (I - Q)\boldsymbol{v}$ is feasible. Thus

$$R_p(\boldsymbol{w}) - R_p(\boldsymbol{v}) \le R_p(\boldsymbol{v} - \boldsymbol{h}) - R_p(\boldsymbol{v}), \tag{129}$$

$$R_p(\boldsymbol{v} - \boldsymbol{h}) - R_p(\boldsymbol{v}) \le \mathbb{E}_p|\mathbf{z}^\top\boldsymbol{h}|, \tag{130}$$

$$\begin{aligned} R_p(\boldsymbol{v} - \boldsymbol{h}) - R_p(\boldsymbol{v}) &= -\boldsymbol{g}_*^\top\boldsymbol{h} + \int_0^1 (1-t)\boldsymbol{h}^\top\nabla^2 R_p(\boldsymbol{v} - t\boldsymbol{h})\boldsymbol{h}\,dt \\ &\le -\boldsymbol{g}_*^\top\boldsymbol{h} + \tfrac{1}{8}\boldsymbol{h}^\top\boldsymbol{C}_p\boldsymbol{h}. \end{aligned} \tag{131}$$

Equation (130) uses $|\varphi'| \le 1$, and the last line uses (128). The Taylor expression can be negative; a tilted feasible space can exploit source-predictive unstable information.

Under the stipulated source-to-bridge coupling, let $M = y\boldsymbol{w}^\top \mathbf{z}$, $T = y\boldsymbol{w}^\top \Delta$, and $\beta(M) = (1 + e^M)^{-1}$. Since $0 \le \varphi'' \le 1/4$,

$$R_{\tilde{q}}(\boldsymbol{w}) - R_p(\boldsymbol{w}) \le -\mathbb{E}[\beta(M)T] + \tfrac{1}{8}\mathbb{E}T^2, \tag{132}$$

$$\mathbb{E}T^2 \le b^2, \qquad -\mathbb{E}[\beta(M)T] \le B_p(\boldsymbol{w})\sqrt{\mathbb{E}T^2} \le bB_p(\boldsymbol{w}). \tag{133}$$

The first inequality in (133) uses the displacement second moment, and the second is Cauchy–Schwarz. The same risk difference is at most b by logistic Lipschitzness. Add the exact bridge bound and the smaller source and direct-shift bounds to obtain (127). □

D.2 Proof of finite-pair tightness

We first obtain a finite-probability adverse angle from pairs with Gaussian noise, then construct fixed source and target distributions for which source refitting preserves its first-order risk cost. In this subsection, $\sigma = \sigma_\xi > 0$.

Lemma 8 (An adverse angle under Gaussian pair noise)**.** *In dimension two, let $\mathcal{U} = \mathrm{span}(\boldsymbol{e}_2)$, $\boldsymbol{u}_j = \sqrt{g}\,\boldsymbol{e}_2$ for $g > 0$, and $\boldsymbol{\xi}_j = \sigma(X_j, W_j)^\top$, where all X_j, W_j are independent standard normal variables. Set*

$$\delta_k = \sqrt{\frac{1}{k}\left(\frac{\sigma^2}{g} + \frac{\sigma^4}{g^2}\right)}. \tag{134}$$

If $\delta_k^2 \le 1/2048$, then, with probability at least $1/64$, the leading rank-one projector of the observed pair covariance equals $Q_t = \boldsymbol{u}_t\boldsymbol{u}_t^\top$, where $\boldsymbol{u}_t = (\sin t, \cos t)^\top$ and

$$\delta_k/6 \le t \le 16\delta_k, \qquad \delta_k/12 \le \|Q_t - P_{\mathcal{U}}\|_{\mathrm{op}} \le 16\delta_k. \tag{135}$$

Proof. A fourth-moment bound gives a cross-covariance fluctuation of its natural size with positive probability. A variance bound keeps the diagonal gap near g. The exact two-dimensional eigenvector formula then gives the angle.

Define $B_k = (\widetilde{\boldsymbol{\Sigma}}_k)_{12}$, $D_k = (\widetilde{\boldsymbol{\Sigma}}_k)_{22} - (\widetilde{\boldsymbol{\Sigma}}_k)_{11}$, and $h^2 = \sigma^2 g + \sigma^4$. Independence and Gaussian moments give

$$B_k = k^{-1}\sum_j \sigma X_j(\sqrt{g} + \sigma W_j), \qquad \mathbb{E}B_k^2 = h^2/k, \tag{136}$$

$$D_k - g = k^{-1}\sum_j [2\sqrt{g}\sigma W_j + \sigma^2(W_j^2 - X_j^2)], \qquad \mathrm{Var}(D_k) = 4h^2/k. \tag{137}$$

For one centered cross-term summand $A = \sigma X(\sqrt{g} + \sigma W)$,

$$\mathbb{E}A^4 = 3\sigma^4(g^2 + 6g\sigma^2 + 3\sigma^4) \le 9h^4, \qquad \mathbb{E}B_k^4 = \frac{k\mathbb{E}A^4 + 3k(k-1)h^4}{k^4} \le \frac{9h^4}{k^2}. \tag{138}$$

For any nonnegative H with finite second moment,

$$\Pr(H \ge \theta\mathbb{E}H) \ge \frac{(1-\theta)^2(\mathbb{E}H)^2}{\mathbb{E}H^2}, \qquad 0 < \theta < 1. \tag{139}$$

Indeed, split $\mathbb{E}H$ below and above $\theta\mathbb{E}H$, then use Cauchy–Schwarz on the upper part. Apply (139) to $H = B_k^2$ with $\theta = 1/4$. The distribution of B_k is symmetric under $X_{1:k} \mapsto -X_{1:k}$. This symmetry, a second-moment tail bound, and (137) yield

$$\begin{aligned} \Pr(B_k \ge h/(2\sqrt{k})) &\ge 1/32, \\ \Pr(B_k > 8h/\sqrt{k}) &\le 1/128, \\ \Pr(|D_k - g| > g/2) &\le 16\delta_k^2 \le 1/128. \end{aligned} \tag{140}$$

For the second line, symmetry divides the two-sided second-moment tail bound by two. Therefore the event

$$E_k = \{h/(2\sqrt{k}) \le B_k \le 8h/\sqrt{k}, \quad g/2 \le D_k \le 3g/2\} \tag{141}$$

has probability at least $1/32 - 1/128 - 1/128 = 1/64$. No independence between B_k and D_k is required.

On E_k, the leading eigenspace is unique and its angle satisfies $t = \frac{1}{2}\arctan(2B_k/D_k)$. The arctangent argument is positive and at most $32\delta_k \le 1/\sqrt{2}$. For $0 \le z \le 1$, $z/2 \le \arctan z \le z$, so

$$\delta_k/6 \le B_k/(2D_k) \le t \le B_k/D_k \le 16\delta_k. \tag{142}$$

The projector error is $\sin t$. Since $0 < t < \pi/4$, $t/2 \le \sin t \le t$, proving (135). The Gaussian noises satisfy A2 with a fixed directional sub-Gaussian constant. Their k original draws remain the independent observations under the symmetrized pair-covariance convention. □

Proof of Proposition 3.

Proof. The source has a noisy unstable label signal. A small rotation in the adverse direction reduces the usable source information, and a fixed target shift reduces the same signed margins further. Giving both labels positive probability at each source feature point ensures an attained fitted optimum for every possible pair realization.

Set $\mathcal{S} = \mathrm{span}(\boldsymbol{e}_1)$ and $\mathcal{U} = \mathrm{span}(\boldsymbol{e}_2)$. Let $\Pr(y = 1) = 3/4$ and let $V \in \{-1, 1\}$ satisfy $\Pr(V = y \mid y) = 3/4$. Fix the source and target by the coupling

$$\mathbf{z} = (1, V)^\top, \qquad \mathbf{z}' = (1, V + y)^\top. \tag{143}$$

They satisfy A1 with constant stable feature and conditional label probability $3/4$. Their stable-label marginals agree, so $\breve{q} = q$ and $C_{\mathrm{br}} = 0$. Also $\Delta = y\boldsymbol{e}_2$, giving $\mathbb{E}\Delta\Delta^\top = P_{\mathcal{U}}$ and $\rho_\Delta = 1$. Both distributions are fixed before the pair noise and independent of g, σ, k. Their features are bounded. The stable optimum and source second moment are

$$\boldsymbol{v} = a\boldsymbol{e}_1, \quad a = \log 3, \qquad \boldsymbol{C}_p = \begin{pmatrix} 1 & 1/4 \\ 1/4 & 1 \end{pmatrix}, \qquad L_z = \sqrt{5/4}. \tag{144}$$

For a removed direction $\boldsymbol{u}_t = (\sin t, \cos t)^\top$, the retained unit direction is $\boldsymbol{v}_t = (\cos t, -\sin t)^\top$. Let β_t be the scalar coefficient of the fitted predictor $\boldsymbol{w}_t = \beta_t \boldsymbol{v}_t$. Its objective is

$$F(\beta, t) = \mathbb{E}\varphi\{\beta(y\cos t - yV\sin t)\}. \tag{145}$$

Each of the four pairs (y, V) has probability at least $1/16$. Using $\varphi(s) + \varphi(-s) \ge |s|$ gives, for every t,

$$F(\beta, t) \ge \frac{|\beta|}{16}\big(|\cos t - \sin t| + |\cos t + \sin t|\big) \ge \frac{|\beta|}{8\sqrt{2}}. \tag{146}$$

Hence F is coercive uniformly in t. It is strictly convex because the two feature values $\cos t \pm \sin t$ cannot both vanish and logistic curvature is positive at every finite margin. Thus β_t exists uniquely for every selected line, with $|\beta_t| \le 8\sqrt{2}\log 2$ by comparison with $F(0, t) = \log 2$.

At zero angle, $\beta_0 = a$, $F_\beta(a, 0) = 0$, and $F_{\beta\beta}(a, 0) = 3/16$. The implicit-function theorem gives a smooth branch $\beta_t = a + O(t)$, which is the actual unique optimizer. Differentiating its source risk removes the optimizing-coefficient derivative by stationarity. Since $\mathbb{E}[V \mid y] = y/2$,

$$\frac{d}{dt} R_p(\boldsymbol{w}_t)\Big|_{t=0} = a\mathbb{E}\frac{yV}{1 + e^{ay}} = \frac{3a}{16}, \tag{147}$$

$$\frac{d}{dt}\{R_q(\boldsymbol{w}_t) - R_p(\boldsymbol{w}_t)\}\Big|_{t=0} = a\mathbb{E}\frac{1}{1 + e^{ay}} = \frac{3a}{8}. \tag{148}$$

The second line follows because the target margin is the source margin minus $\beta_t \sin t$. Finite support and smoothness justify Taylor expansion, giving

$$R_p(\boldsymbol{w}_t) - R_p(\boldsymbol{v}) = \frac{3a}{16}t + O(t^2), \qquad R_q(\boldsymbol{w}_t) - R_p(\boldsymbol{w}_t) = \frac{3a}{8}t + O(t^2). \tag{149}$$

Define $S_t = L_z a \sin t$ and $T_t = \|\boldsymbol{w}_t\|_2 \sin t$. Since $\beta_t \to a > 0$, the ratios of the two expressions in (149) to S_t and T_t tend respectively to $3/(16L_z)$ and $3/8$. Thus constants $b_0 > 0$ and $t_0 \in (0, 1/4)$ exist such that, for $0 < t \le t_0$,

$$R_p(\boldsymbol{w}_t) - R_p(\boldsymbol{v}) \ge b_0 S_t, \qquad R_q(\boldsymbol{w}_t) - R_p(\boldsymbol{w}_t) \ge b_0 T_t. \tag{150}$$

Both increments are positive on the same angle interval.

Use the Gaussian pair design in Lemma 8. Let $\delta_0 = \min\{1/\sqrt{2048}, t_0/16\}$ and suppose $0 < \delta_k \le \delta_0$. On its event of probability at least $1/64$, the fitted output is $\boldsymbol{w}_t$ with $0 < t \le t_0$ and $\sin t \ge \delta_k/12$. Adding (150) proves the proposition with $c = b_0/12$. It also proves its componentwise interpretation. The deterministic Lipschitz comparison gives the matching upper inequality on this event:

$$R_q(\boldsymbol{w}_k) - R^*_{p,\mathcal{S}} \le 16\delta_k\big(L_z\|\boldsymbol{v}\|_2 + \rho_\Delta\|\boldsymbol{w}_k\|_2\big). \tag{151}$$

Indeed, projecting $\boldsymbol{v}$ onto the retained line costs at most $L_z\|\boldsymbol{v}\|_2 \sin t$ in source risk, while residual unstable sensitivity is $\|\boldsymbol{w}_k\|_2 \sin t$. The source and target feature moments and fitted coefficient norms are bounded independently of g, σ, k. □

Rate and probability scope. For fixed confidence $\eta \in (0,1)$, put $x = \sigma^2/g$ and $H = 2 + \log(1/\eta)$. The expanded pair factor satisfies

$$\sqrt{H}\,\delta_k \le (\sqrt{x} + x)\sqrt{H/k} + xH/k \le (\sqrt{2H} + H)\delta_k. \tag{152}$$

This follows from $\sqrt{x + x^2} \le \sqrt{x} + x \le \sqrt{2}\sqrt{x + x^2}$ and $k \ge 1$. The construction therefore covers both $\sigma/\sqrt{gk}$ and $\sigma^2/(g\sqrt{k})$ regimes when k makes δ_k small. It does not separately establish necessity of the lower-order $1/k$ concentration term. A uniform $o(\delta_k)$ target-excess bound at failure probability below $1/64$ would contradict the proposition. This is a lower bound for GRIT's fitted procedure at fixed dimension and confidence, rather than an infimum over all estimators. Negative-angle realizations can have negative first-order excess relative to the restricted stable comparator, so the statement does not give the same rate for expected signed excess.

D.3 Positive bridge shift and an asymptotically attained comparison

The next construction adds a positive bridge covariate-shift contribution of the same order. Its target sequence is specified before the pair noise. It also attains the refined risk comparison to first order, including its source and direct-shift terms.

Fix $a, c, \rho > 0$ and $\nu \in (1/2, 1)$, and put $\mu = 2\nu - 1$. Under the source, let X be uniform on $\{1, 2\}$, let $\Pr(y = 1 \mid X = x) = \pi_x = s(ax)$ with $s(t) = (1+e^{-t})^{-1}$, and let $J \in \{-1, 1\}$ be independent with $\Pr(J = 1) = \nu$. Set

$$\mathbf{z} = (X, cyJ)^\top, \quad \boldsymbol{v} = a\boldsymbol{e}_1, \quad \beta_*(X, y) = (1 + e^{yaX})^{-1}, \quad \kappa = \mathbb{E}_p\beta_*, \quad B_* = (\mathbb{E}_p\beta_*^2)^{1/2}. \tag{153}$$

Here $\boldsymbol{v}$ is the stable optimum and direct conditioning gives

$$\kappa = \pi_1(1 - \pi_1) + \pi_2(1 - \pi_2) > 0, \qquad B_*^2 = \kappa/2 > 0. \tag{154}$$

For $\varepsilon \in [0, 1/2)$ define a target q_ε by assigning masses $1/2 + \varepsilon$ and $1/2 - \varepsilon$ to $X = 1, 2$, retaining the conditional distribution of (y, J) given X, and setting

$$\mathbf{z}' = \left(X, cyJ + \rho y\frac{\beta_*(X, y)}{B_*}\right)^\top. \tag{155}$$

The bridge $\breve{q}$ has the source's equal stable masses and this target unstable conditional. Its coupling with the source fixes (X, y, J) and satisfies

$$\Delta = \rho y\frac{\beta_*}{B_*}\boldsymbol{e}_2, \qquad \mathbb{E}_p\Delta\Delta^\top = \rho^2 P_{\mathcal{U}}. \tag{156}$$

A1 holds, all stable supports coincide, and source, bridge, and target features are uniformly bounded. The shift uses the fixed stable oracle, not a fitted coefficient.

For $Q_\theta = \boldsymbol{u}_\theta\boldsymbol{u}_\theta^\top$, where $\boldsymbol{u}_\theta = (\sin\theta, \cos\theta)^\top$, let b_θ be the scalar coefficient of the fitted source predictor $\boldsymbol{w}_\theta = b_\theta(\cos\theta, -\sin\theta)^\top$. Let $\boldsymbol{g}_* = \nabla R_p(\boldsymbol{v})$, $\boldsymbol{h}_\theta = Q_\theta\boldsymbol{v}$, and $\ell_\theta = \rho\|P_{\mathcal{U}}\boldsymbol{w}_\theta\|_2$. Define

$$\begin{aligned}\mathcal{C}_\varepsilon(\theta) = R_p(\boldsymbol{v}) + C_{\mathrm{br},\varepsilon}(\boldsymbol{w}_\theta) - \boldsymbol{g}_*^\top\boldsymbol{h}_\theta + \tfrac{1}{8}\boldsymbol{h}_\theta^\top\boldsymbol{C}_p\boldsymbol{h}_\theta \\ + \ell_\theta B_p(\boldsymbol{w}_\theta) + \tfrac{1}{8}\ell_\theta^2.\end{aligned} \tag{157}$$

Lemma 7 makes this an upper bound on $R_{q_\varepsilon}(\boldsymbol{w}_\theta)$.

Proposition 17 (Joint attainment of a refined risk comparison)**.** *For the fixed source* (153)*, use the Gaussian pair design of Lemma 8 and the target $q_k = q_{\delta_k}$, with δ_k from* (134)*. For all sufficiently small positive δ_k, an event of probability at least $1/64$ has the following properties. The fitted optimizer exists for every pair realization and has uniformly bounded norm. On the event, each of*

$$R_p(\boldsymbol{w}_k) - R^*_{p,\mathcal{S}}, \qquad R_{\breve{q}}(\boldsymbol{w}_k) - R_p(\boldsymbol{w}_k), \qquad C_{\mathrm{br},k}(\boldsymbol{w}_k) \tag{158}$$

is bounded above and below by positive constants times δ_k, and

$$0 \le \mathcal{C}_{\delta_k}(\theta_k) - R_{q_k}(\boldsymbol{w}_k) = O(\delta_k^2), \qquad \frac{\mathcal{C}_{\delta_k}(\theta_k) - R^*_{p,\mathcal{S}}}{R_{q_k}(\boldsymbol{w}_k) - R^*_{p,\mathcal{S}}} \longrightarrow 1. \tag{159}$$

The remainder constants depend only on the fixed source and shift parameters. Convergence is uniform over all angles in (135) *as $\delta_k \to 0$. Source and target are specified before observing the pairs; the target depends on k, g, σ only through its stable marginal.*

Proof. Source stationarity makes refitting preserve the first-order rotation cost. The fixed shift in (155) attains the source-dependent transport coefficient at the stable oracle. Moving stable mass toward the higher-entropy label kernel adds a positive exact bridge term. The remaining certificate slack is quadratic in the angle.

Existence and smooth fitting. The four source feature atoms (x, cv), $x \in \{1, 2\}$, $v \in \{-1, 1\}$ each occur jointly with each label with probability at least $\zeta = \frac{1}{2}(1 - \pi_2)(1 - \nu) > 0$. Pairing the two label losses at each atom gives

$$R_p(\boldsymbol{w}) \ge \zeta \sum_{x\in\{1,2\},\ v\in\{-1,1\}} |xw_1 + cvw_2| \ge \zeta\sqrt{2}\min(1, c)\|\boldsymbol{w}\|_2. \tag{160}$$

The last inequality already follows from $(1, \pm c)$, whose absolute scores sum to $2\max(|w_1|, c|w_2|)$. The source risk is coercive and has positive-definite Hessian at every finite coefficient because its atoms span the plane and logistic curvature is positive. Every line-constrained optimum is therefore unique and finite, with norm at most $\log 2/[\zeta\sqrt{2}\min(1, c)]$.

Its scalar objective is

$$F(b, \theta) = \mathbb{E}_p\varphi\{b(yX\cos\theta - cJ\sin\theta)\}. \tag{161}$$

At zero angle, $b_0 = a$, $F_b(a, 0) = 0$, and $F_{bb}(a, 0) = \mathbb{E}[X^2\pi_X(1-\pi_X)] > 0$. The implicit-function theorem gives $b_\theta = a + O(\theta)$, a smooth branch equal to the global line-constrained optimum by uniqueness.

Three positive first-order contributions. Differentiating the optimized source risk and the direct-shift difference at zero yields

$$\frac{d}{d\theta}R_p(\boldsymbol{w}_\theta)\Big|_0 = ac\mathbb{E}[J\beta_*] = ac\mu\kappa, \tag{162}$$

$$\frac{d}{d\theta}\{R_{\breve{q}}(\boldsymbol{w}_\theta) - R_p(\boldsymbol{w}_\theta)\}\Big|_0 = a\rho\frac{\mathbb{E}\beta_*^2}{B_*} = a\rho B_*. \tag{163}$$

The first line uses $F_b(a, 0) = 0$ and independence of J. For the second, the target margin is the source margin minus $\rho b_\theta \sin\theta\, \beta_*/B_*$. Smooth finite sums give second-order remainders for both expressions.

Let $h_2(p) = -p\log p - (1-p)\log(1-p)$ and $H = h_2(\pi_1) - h_2(\pi_2) > 0$. The sign follows because $1/2 < \pi_1 < \pi_2 < 1$ and entropy is strictly decreasing on that interval. If $L_x(\theta)$ is the target conditional risk at $X = x$, then $L_x(0) = h_2(\pi_x)$ and L_x is smooth. Thus

$$R_{q_\varepsilon}(\boldsymbol{w}_\theta) - R_{\breve{q}}(\boldsymbol{w}_\theta) = \varepsilon[L_1(\theta) - L_2(\theta)] = \varepsilon[H + O(\theta)]. \tag{164}$$

It is positive for small positive θ, ε, so it equals the absolute bridge term. Combining gives

$$\begin{aligned} R_{q_\varepsilon}(\boldsymbol{w}_\theta) - R_p(\boldsymbol{v}) &= R_p(\boldsymbol{w}_\theta) - R_p(\boldsymbol{v}) + R_{\breve{q}}(\boldsymbol{w}_\theta) - R_p(\boldsymbol{w}_\theta) + C_{\mathrm{br},\varepsilon}(\boldsymbol{w}_\theta) \\ &= a(c\mu\kappa + \rho B_*)\theta + H\varepsilon + O(\theta^2 + \varepsilon\theta). \end{aligned} \tag{165}$$

All three leading coefficients are strictly positive.

The refined certificate has the same first derivatives. Direct differentiation gives $\boldsymbol{g}_* = (0, -c\mu\kappa)^\top$. Also

$$-\boldsymbol{g}_*^\top \boldsymbol{h}_\theta = ac\mu\kappa \sin\theta\cos\theta, \quad \|\boldsymbol{h}_\theta\|_2 = a\sin\theta, \quad \ell_\theta = \rho b_\theta \sin\theta \tag{166}$$

for small positive θ. Since $B_p(\boldsymbol{w}_\theta) = B_* + O(\theta)$, the source and direct terms in (157) equal $ac\mu\kappa\theta + O(\theta^2)$ and $a\rho B_*\theta + O(\theta^2)$, respectively. The exact bridge term appears unchanged in the actual risk and the certificate. Their difference is therefore $O(\theta^2)$ and is nonnegative by Lemma 7.

Finally apply Lemma 8 and set $\varepsilon = \delta_k$. On its event, θ_k is between $\delta_k/6$ and $16\delta_k$, so all local expansions apply for small enough δ_k. Each contribution is bounded above and below by a positive constant times δ_k, the slack is $O(\delta_k^2)$, and the actual excess in the ratio denominator is at least a positive constant times δ_k. This proves (159) uniformly on the stated angle interval. □

Setting $\varepsilon = 0$ instead gives one target fixed for every pair count, with zero bridge term, while preserving the ratio-one conclusion for the refined source and direct-shift comparison. With $\varepsilon = \delta_k$, the bridge, source, and direct effects are simultaneously positive. The simpler source and direct terms remain jointly sharp up to constants because

$$\frac{R_p(\boldsymbol{w}_\theta) - R_p(\boldsymbol{v})}{L_z\|\boldsymbol{v}\|_2 \sin\theta} \longrightarrow \frac{c\mu\kappa}{L_z} > 0, \qquad \frac{R_{\check{q}}(\boldsymbol{w}_\theta) - R_p(\boldsymbol{w}_\theta)}{\rho\|\boldsymbol{w}_\theta\|_2 \sin\theta} \longrightarrow B_* > 0. \tag{167}$$

The ratio-one result concerns the complete refined comparison at the realized fitted projector, with $r = r_* = 1$. Replacing its angle by a concentration radius can add slack.

D.4 Quadratic cost of removing additional directions

Source stationarity also permits a sharper treatment of extra-direction removal without a Hessian lower bound. Let $P = P_{\mathcal{U}}$, let Q be a rank-r_* orthogonal projector, and let Q_r be an orthogonal projector containing its range. Define

$$\alpha = \|Q - P\|_{\mathrm{op}}, \quad H = Q_r - Q, \quad \boldsymbol{v} = \boldsymbol{w}_{\mathcal{S}}^*, \quad a = \|\boldsymbol{v}\|_2, \quad \delta = \|H\boldsymbol{v}\|_2, \quad G_U = \|P\nabla R_p(\boldsymbol{v})\|_2. \tag{168}$$

Proposition 18 (Quadratic extra-direction comparison)**.** *Suppose the source has finite second feature moments, $\boldsymbol{v}$ is an attained stable logistic optimum, and $\boldsymbol{w}_r$ is an attained source optimum on* $\ker Q_r$. *Then*

$$R_p(\boldsymbol{w}_r) - R_p(\boldsymbol{v}) \le G_U\alpha a + \frac{L_z^2}{8}(\alpha^2 a^2 + \delta^2). \tag{169}$$

Under the bridge and conditional-displacement assumptions, this implies

$$R_q(\boldsymbol{w}_r) \le R_p(\boldsymbol{v}) + C_{\mathrm{br}}(\boldsymbol{w}_r) + G_U\alpha a + \frac{L_z^2}{8}(\alpha^2 a^2 + \delta^2) + \rho_\Delta \alpha \|\boldsymbol{w}_r\|_2. \tag{170}$$

Proof. Compare with the projected stable oracle using the logistic Taylor bound. Stable stationarity confines the linear term to unstable directions; nesting controls it by the estimation angle. Orthogonality separates the two squared removal costs.

Set $\boldsymbol{h} = Q_r\boldsymbol{v}$. By feasibility, (128), and (131),

$$\begin{aligned} R_p(\boldsymbol{w}_r) - R_p(\boldsymbol{v}) &\le -\nabla R_p(\boldsymbol{v})^\top \boldsymbol{h} + \tfrac{1}{8}\boldsymbol{h}^\top \boldsymbol{C}_p \boldsymbol{h} \\ &\le G_U\|PQ_r\boldsymbol{v}\|_2 + \tfrac{1}{8}L_z^2\|Q_r\boldsymbol{v}\|_2^2. \end{aligned} \tag{171}$$

Because $P\boldsymbol{v} = 0$, $PQ_r\boldsymbol{v} = -P(I - Q_r)\boldsymbol{v}$. Nesting yields $(I - Q_r) = (I - Q)(I - Q_r)$, so $\|P(I - Q_r)\|_{\mathrm{op}} \le \|P(I - Q)\|_{\mathrm{op}} \le \alpha$. Moreover, Q and H are orthogonal projectors with orthogonal ranges. Hence

$$\|PQ_r\boldsymbol{v}\|_2 \le \alpha a, \qquad \|Q_r\boldsymbol{v}\|_2^2 = \|Q\boldsymbol{v}\|_2^2 + \|H\boldsymbol{v}\|_2^2 \le \alpha^2 a^2 + \delta^2. \tag{172}$$

Insert these into (171). Finally $\|P\boldsymbol{w}_r\|_2 \le \alpha\|\boldsymbol{w}_r\|_2$ because $\boldsymbol{w}_r \in \ker Q_r$; Theorem 1 proves (170). □

When $Q = P$, the source bound reduces to $L_z^2\delta^2/8$. Thus the linear cost $L_z\delta$ can be locally loose for pure extra-direction removal with bounded moments. The quadratic coefficient is attained to leading order by an actual noiseless GRIT example.

Attaining example. Take $d = 3$, $\mathcal{S} = \mathrm{span}(\boldsymbol{e}_1, \boldsymbol{e}_2)$, $\mathcal{U} = \mathrm{span}(\boldsymbol{e}_3)$, and $q = p$ with

$$\mathbf{z} = (1, X, 0)^\top, \quad \Pr(X = 1) = \Pr(X = -1) = 1/2, \quad \Pr(y = 1 \mid X) = s(bX), \qquad b > 0. \tag{173}$$

The stable optimum is uniquely $\boldsymbol{v} = (0, b, 0)^\top$, since its score is the conditional log odds and the stable feature vectors span the stable plane. Here $\boldsymbol{C}_p = \mathrm{diag}(1, 1, 0)$ and $L_z = 1$. For every k, take noiseless clean differences $\boldsymbol{u}_j = \sqrt{g}\,\boldsymbol{e}_3$. The observed pair covariance is $gP_{\boldsymbol{e}_3}$, satisfying A2 with $\sigma_\xi = 0$. Fix the zero-eigenvalue tie rule to select $\boldsymbol{e}_2$ before $\boldsymbol{e}_1$; it is specified before labels. Then $Q = P_{\boldsymbol{e}_3}$ and $Q_2 = Q + P_{\boldsymbol{e}_2}$. Only the intercept remains, and its fitted optimum is $\boldsymbol{w}_2 = 0$ because the marginal label probability is $1/2$. With $\delta = \|(Q_2 - Q)\boldsymbol{v}\|_2 = b$,

$$\begin{aligned} R_q(\boldsymbol{w}_2) - R_p(\boldsymbol{v}) &= \log 2 - h_2(s(b)) \\ &= \int_0^b t\, s(t)(1 - s(t))\, dt = \frac{b^2}{8} + O(b^4) = \frac{\delta^2}{8} + O(\delta^4). \end{aligned} \tag{174}$$

The integral follows by differentiating the entropy expression; $s(t)(1 - s(t)) = 1/4 + O(t^2)$ gives the expansion. Thus the quadratic coefficient $1/8$ cannot be reduced uniformly. This example uses an explicit tie rule, since the leading rank-two eigenspace is not unique without one.

E Predictive Stability and Sources of Pair Information

This appendix relates A1 to established invariance criteria, then gives sufficient feature-space conditions under which latent models, repeated observations, and augmentations supply pair information.

E.1 Conditional invariance and common optimal prediction

A1 posits a single conditional label distribution P_{inv} on projected features, valid almost surely in each environment. The notation $P^{(e)}(y \mid \boldsymbol{z}_\mathcal{S}) = P_{\mathrm{inv}}(y \mid \boldsymbol{z}_\mathcal{S})$ denotes equality of conditional distributions, without assuming a density or discrete labels. This is the subspace analogue of the invariant feature-subset condition in Rojas-Carulla et al. (2018, Section 2, assumptions A1 and A1′). It does not assert that $\mathcal{S}$ is unique or that it preserves all label information available in the original features. The zero subspace, for example, satisfies A1 whenever label marginals agree. Predictions based on this projection use only the common label marginal.

Common conditional distributions and common optimal predictors are related through the loss and predictor class. Ideal IRM requires one predictor to minimize each environment's risk on the chosen representation (Arjovsky et al., 2019, Section 3, Definition 3). Under squared loss, unrestricted optimal predictions determine conditional means but leave higher moments free. For binary logistic loss with attained finite optima, unrestricted optimal logits determine conditional label probabilities. Restrictions on the head change this relationship (Kamath et al., 2021, Section 2.1). In particular, even a common conditional distribution need not yield the same best misspecified linear predictor when the feature marginal changes.

Example 1 (A common conditional distribution with different optimal linear heads)**.** *Let $Z \in \{-2, -1, 1, 2\}$ have a symmetric distribution with $\Pr_e(|Z| = 2) = t_e \in (0, 1)$, and set $Y = Z^3$ in every environment. The conditional distribution of Y given Z is common, but the squared-loss optimal slope is*

$$w_e = \frac{1 + 15t_e}{1 + 3t_e}. \tag{175}$$

It differs for distinct values of t_e, with or without a fitted intercept.

Proof. Sketch. Compute the population normal equation; symmetry removes the intercept. Symmetry gives $\mathbb{E}_e Z = \mathbb{E}_e Y = 0$, so the optimal intercept is zero. Moreover, $\mathbb{E}_e Z^2 = 1 + 3t_e > 0$ and $\mathbb{E}_e ZY = \mathbb{E}_e Z^4 = 1 + 15t_e$. Differentiating the quadratic risk gives $w_e \mathbb{E}_e Z^2 = \mathbb{E}_e ZY$. The displayed ratio has derivative $12/(1 + 3t_e)^2 > 0$, proving the claim. □

The same distinction limits causal interpretations of A1. Invariant causal prediction uses invariance to identify candidate causal predictors under suitable intervention assumptions (Peters et al., 2016). Here A1 is a statistical condition on a selected projection, and pairs supply geometric information

about its complement. Neither conditional invariance nor observed pair agreement alone identifies a causal graph.

E.2 A SUFFICIENT LATENT CONSTRUCTION

An orthogonal latent decomposition provides one transparent way to satisfy the pair model. Let A and B have orthonormal columns with $A^\top B = 0$ and $[A\ B]$ an orthogonal basis of $\mathbb{R}^d$. Suppose features have the form $Z = AC + BV$, with a common conditional distribution of Y given C across environments. The joint distribution of (C, V) and the conditional distribution of V given (C, Y) may change. Choose $\mathcal{S} = \mathrm{range}(A)$ and $\mathcal{U} = \mathrm{range}(B)$.

Proposition 19 (Task-preserving repeated observations)**.** *Under this construction A1 holds. Two observations sharing C have feature difference in $\mathcal{U}$. For a finite collection of such pairs, its clean pair covariance has range $\mathcal{U}$ if and only if the latent differences span the column-coordinate space of B.*

Proof. Sketch. The stable projection recovers C, and differencing cancels it; the remaining covariance range is the span of the differences. Since $P_{\mathcal{S}} Z = AC$ and $A^\top AC = C$, conditioning on $P_{\mathcal{S}} Z$ is equivalent to conditioning on C. The common conditional distribution therefore gives A1. For shared-C observations, $Z - Z' = B(V - V') \in \mathcal{U}$. Let D be the matrix with columns $V_j - V_j'$. The pair covariance is $BDD^\top B^\top / k$. Its range is $\mathrm{range}(BD)$ because a Gram matrix and its factor have the same range. Since B is injective on its column-coordinate space, this range equals $\mathcal{U}$ precisely when D has full row rank. □

Interventions that change V while preserving C supply one causal realization of this sufficient construction. For a nonorthogonal mixing model, a stable linear readout may recover C, but its orthogonal complement must be checked against the actual pair-difference span.

E.3 MULTIPLE VIEWS, AUGMENTATIONS, AND POOLED COLLECTIONS

Two camera views or repeated measurements can instantiate the shared-C construction when the measured state determining the task remains fixed and only the viewing or measurement variable changes. Even then, coverage depends on which changes are observed. Repeated observations along one direction cannot identify a higher-dimensional unstable space, irrespective of the number of repetitions. Changes to the task-relevant state instead challenge pair validity.

An augmentation T yields an invariant pair with exact geometric compatibility if $P_{\mathcal{S}}\phi(Tx) = P_{\mathcal{S}}\phi(x)$. The difference then lies in $\ker P_{\mathcal{S}} = \mathcal{U}$. This identity is a sufficient feature-space condition, stronger than simply assigning both endpoints the same class label. Augmented views in representation learning (Chen et al., 2020; Grill et al., 2020) motivate the use of related observations but need not satisfy this condition for a downstream prediction task. Residual geometric differences outside $\mathcal{U}$ require a noise or misspecification model, not merely the label "augmentation."

For several compatible sources, let $\widetilde{\mathbf{\Sigma}}^{(a)}$ be the pair covariance of source a, based on k_a pairs, and let $k = \sum_a k_a$. Pooling gives the exact identity

$$\widetilde{\mathbf{\Sigma}}_k = \sum_a \frac{k_a}{k} \widetilde{\mathbf{\Sigma}}^{(a)}. \tag{176}$$

Complementary clean spans can enlarge coverage, but source frequency and difference magnitude also change spectral ordering. Biased residuals or incompatible invariance relations can remove task information. The theoretical noise assumptions must therefore be checked for the pooled collection rather than inferred from its sources' names.

F EXPERIMENTAL PROTOCOLS AND ADDITIONAL RESULTS

This appendix describes the core comparison on ColoredMNIST-INV and Waterbirds-INV, our variants augmented with invariant-pair information. Appendix H documents the controlled CueMNIST and Gaussian studies.

F.1 Data, features, and information access

ColoredMNIST-INV. Following the task construction of Arjovsky et al. (2019), the clean binary target indicates whether an MNIST digit is at least five. Independently flipping this target with probability 0.25 defines the supervised target. Color is sampled relative to the supervised target, with disagreement probabilities 0.1 and 0.2 in the two training environments and 0.9 at test. The official training sources are partitioned into two disjoint sets of 25,000 digits and a validation set of 10,000 digits. The 10,000 official test digits are reserved for final evaluation. Partitioning precedes rendering. Validation uses three renderings of its sources, at color disagreement rates 0.1, 0.2, and 0.3. These are repeated measurements of 10,000 sources. The 0.4 and 0.5 held-out validation rates belong to separately selected sensitivity studies.

The 256 pairs are sampled without replacement from training sources and show the same digit in both colors. Both endpoints preserve its noisy target as well as its underlying digit. Their canonical orientation is red minus green. The auxiliary endpoints enter pair estimation or penalties, but are not added as supervised records. An anchor may also occur in the labeled training set. Source IDs are available to V-REx, IRMv1, Fishr, and RDM; training colors are not. GroupDRO and LISA use four groups defined by noisy target label crossed with training source. SD uses no source IDs. Consequently, comparing interventions within a base objective holds its annotation access fixed.

Waterbirds-INV. The released Waterbirds construction combines bird foregrounds with land or water backgrounds (Sagawa et al., 2020). Waterbirds-INV follows this compositing methodology to construct invariant pairs while retaining the training-set size. Our controlled-pair training set preserves its 4,795-record size and official validation and test splits. We select 184 landbirds on land and 56 waterbirds on water, create an opposite-background version of each, and replace the 240 unrelated minority training records with these generated endpoints. The training set therefore contains 4,315 unpaired records and both endpoints of 240 pairs. Every method uses the same labeled dataset; pair-aware variants add only the relations between endpoints. The group counts are 3,498 landbirds on land, 184 landbirds on water, 56 waterbirds on land, and 1,057 waterbirds on water. Validation and test contain 1,199 and 5,794 records, respectively.

Each pair reuses the source bird and compositing geometry while changing the background category. The retained endpoint is a released JPEG and the generated endpoint is a PNG composite, so the construction does not assert byte-identical foregrounds after image encoding. Pair orientation is land minus water. Construction and feature manifests identify the selected bird, mask, background, and both endpoint files. The reconstruction has its own seed and artifact identity. V-REx, IRMv1, Fishr, and RDM use each training example's background as its environment label. GroupDRO and LISA use four bird-label–background groups. ERM and SD use bird labels; their pair-based variants additionally receive the 240 relations. These are different forms of information, so comparisons across base objectives are not comparisons under identical supervision.

Frozen features and predictors. All matrix variants use deterministic evaluation preprocessing and the official CLIP ViT-B/32 image encoder (Radford et al., 2021). Its unnormalized 512-dimensional output defines $\boldsymbol{z} = \phi(\boldsymbol{x})$. Normalization is a separately prepared sensitivity, applied consistently to every feature and pair endpoint when used. Vanilla, GRIT, and prediction consistency fit a two-class affine classifier.

F.2 Interventions, optimization, and selection

Computing the GRIT projection. Assign mass $1/(2k)$ to each difference $\boldsymbol{\delta}_j$ and its opposite $-\boldsymbol{\delta}_j$. This symmetric distribution has mean zero and covariance

$$\frac{1}{2k}\sum_j \left[\boldsymbol{\delta}_j\boldsymbol{\delta}_j^\top + (-\boldsymbol{\delta}_j)(-\boldsymbol{\delta}_j)^\top\right] = \frac{1}{k}\sum_j \boldsymbol{\delta}_j\boldsymbol{\delta}_j^\top.$$

Symmetrization preserves the pair covariance and the number of pairs, k.

For computation, collect the differences in $\boldsymbol{D} \in \mathbb{R}^{d\times k}$. Since $\widetilde{\boldsymbol{\Sigma}}_k = \boldsymbol{D}\boldsymbol{D}^\top/k$, a singular-value decomposition supplies the same directions without forming the covariance. With $\widehat{\boldsymbol{U}}_r = [\widehat{\boldsymbol{u}}_1, \ldots, \widehat{\boldsymbol{u}}_r]$,

project as $\boldsymbol{z} - \widehat{\boldsymbol{U}}_r(\widehat{\boldsymbol{U}}_r^\top \boldsymbol{z})$. The projection stays fixed during fitting and at test time. The experimental protocol caps removal at numerical rank; the mathematical definition allows orthonormal completion.

The three interventions. Let $\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$ denote the two class logits and define $\boldsymbol{\delta}_j = \phi(\boldsymbol{x}_j) - \phi(\boldsymbol{x}'_j)$ for pair j. Vanilla fits the base objective on original features. GRIT uses $P_{k,r,0}\boldsymbol{z}$ throughout training and inference. The implementation computes a deterministic full SVD of the original oriented difference matrix without subtracting its mean. This estimates the pair covariance defined in the method section. The SVD uses float64 and a relative singular-value tolerance of 10^{-12}; the effective removal rank is capped by numerical rank.

Prediction consistency adds the penalty

$$\mathcal{L}_{\text{pred}} = \frac{1}{k}\sum_{j=1}^{k} \|\boldsymbol{W}\boldsymbol{\delta}_j\|_2^2. \tag{177}$$

The sum over coordinates is not divided by the number of logits. Biases cancel. The whole pair bank is evaluated at each update, including base-objective warm-up. The two-logit experimental penalty is related to, but not identical to, the scalar-score penalty analyzed in the theory. For logits with rows $\boldsymbol{w}_0, \boldsymbol{w}_1$, it penalizes both their contrast and their common component. Accuracy and cross-entropy depend only on the contrast. The scalar guarantee therefore does not directly certify the full experimental training procedure.

Connection to paired consistency methods. CoRe penalizes within-identity prediction variability and applies this penalty to logits for classification (Heinze-Deml & Meinshausen, 2021, Sec. 3.2). For a two-member group, its variance penalty with exponent one is proportional to squared paired logit differences. MatchDG first learns cross-domain matches; its reported second-phase implementation penalizes squared distances between their logits (Mahajan et al., 2021, Sec. 4 and App. C.1). SelfReg includes logit penalties for randomly paired same-class examples, with additional learned projection and interpolation components (Kim et al., 2021, Sec. 3). Those pairs need not share an underlying object. Our PC intervention compares a representative output-consistency penalty using supplied invariant pairs and frozen CLIP features, rather than reproducing the complete acquisition and training procedures of these methods.

Base objectives and optimization. V-REx penalizes variation in source risks (Krueger et al., 2021); IRMv1 uses the IRM gradient penalty (Arjovsky et al., 2019); Fishr matches gradient variances across environments (Ramé et al., 2022). SD adds the mean squared raw logits, averaged over examples and classes (Pezeshki et al., 2021). RDM matches the worst environment's distribution of individual cross-entropy losses to their pooled distribution using a Gaussian-kernel maximum mean discrepancy (Nguyen et al., 2024). Its implementation also regularizes the two loss variances, with coefficient 0.004. GroupDRO reweights supervised risks over four label–environment groups; its adversarial weights use supervised cross-entropy only (Sagawa et al., 2020). Feature-space LISA selectively interpolates same-label examples across environments or different labels within an environment, using $\text{Beta}(2, 2)$ mixing weights (Yao et al., 2022). GRIT projects the resulting mixed features; prediction consistency uses the original invariant pair bank independently of the mixing partners. Their supervised statistics are computed separately from the pair penalties. For V-REx and IRMv1, a base penalty coefficient α scales the combined objective as $(\mathcal{L}_{\text{risk}} + \alpha\mathcal{L}_{\text{base}} + \lambda\mathcal{L}_{\text{pair}})/\max(1, \alpha)$. Fishr computes per-example gradients of the classifier. Its moving-average coefficient is 0.95, with 1,500 warm-up updates and an Adam reset when the base penalty activates.

Every variant uses Adam with batch size 256 and at most 40 epochs on ColoredMNIST-INV or 100 on Waterbirds-INV. The shared learning-rate grid is $\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}\}$ and weight-decay grid is $\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$. The matrix searches removal ranks $\{2, 8, 16, 24\}$ and prediction-consistency strengths $\{0.01, 0.1, 1, 10\}$. V-REx and Fishr use base coefficients $\{10, 100, 1000, 10000\}$; IRMv1 uses $\{100, 1000, 10000, 100000\}$. V-REx and IRMv1 activate their penalties at updates 100 and 190, respectively. GroupDRO searches adversarial step sizes $\{0.001, 0.01, 0.1\}$; SD searches $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1\}$; RDM searches $\{0.1, 1, 5, 10\}$; LISA searches mixing probabilities $\{0, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$. SD acts from the first update.

RDM uses the same 1,500-update warm-up and Adam reset convention as Fishr, with its pair penalty active throughout.

Validation and uncertainty. ColoredMNIST-INV's primary selector maximizes the minimum accuracy across its three validation environments. This uses a held-out color environment in addition to source-like validation; a separate source-only selector uses the first two. Waterbirds-INV maximizes validation worst-group accuracy over the four combinations of bird label and background. Thus validation group metadata is available to all methods through selection. Test metrics do not select ordinary runs. Each candidate uses tuning seeds 1101–1103. The top three candidates receive confirmation seeds 2101–2102, and their combined five-seed validation means select one configuration. Final evaluation uses fresh seeds 3101–3110 shared across methods. Each final run selects its checkpoint with validation under that frozen configuration. Restoring this checkpoint precedes final test evaluation.

The reporting protocol calls for a mean, sample standard deviation, and 95% Student-t interval over final seeds, plus intervals for differences joined by explicit seed identity. For a verified set of ten differences d_s, the latter is $\bar{d} \pm t_{0.975,9} s_d/\sqrt{10}$. Such intervals summarize training variability conditional on the dataset and pair bank; they do not quantify uncertainty over new pair collections or populations. Table 3 reports sample standard deviations over ten final seeds. These are not standard errors; the standard error is $s/\sqrt{10}$.

F.3 Expanded core results

Table 3 expands the main comparison with sample standard deviations over the ten final training seeds.

Table 3: Expanded core comparison: means and sample standard deviations (s.d.) over ten final training seeds. +PC and +GRIT add prediction consistency and GRIT to each objective.

	Vanilla		+PC		+GRIT	
Objective	Mean	s.d.	Mean	s.d.	Mean	s.d.
ColoredMNIST-INV (CLIP feat.)						
ERM	17.02	2.68	57.16	0.47	67.89	0.55
SD	17.02	2.68	56.68	0.31	67.77	0.62
RDM	39.95	3.47	59.55	0.89	67.60	0.58
IRMv1	59.89	2.71	61.28	1.19	70.27	0.41
V-REx	54.83	3.45	59.98	1.52	68.20	0.13
Fishr	65.28	1.74	65.33	1.61	68.91	0.18
GroupDRO	30.90	2.46	61.51	0.51	68.01	0.43
LISA	36.68	3.92	61.62	0.68	67.80	0.22
Waterbirds-INV (CLIP feat.)						
ERM	52.69	0.49	66.84	1.84	70.09	1.19
SD	52.06	1.12	66.82	1.82	70.09	1.19
RDM	52.80	1.65	78.59	0.91	76.29	1.56
IRMv1	47.78	1.26	77.86	1.04	78.78	0.82
V-REx	48.27	0.67	78.09	1.39	80.99	0.94
Fishr	47.16	0.80	76.72	0.90	80.84	0.61
GroupDRO	79.83	1.75	81.66	0.80	81.91	0.54
LISA	79.65	1.28	80.77	0.89	82.30	0.73

Paired contrasts. For four-group GroupDRO on Waterbirds-INV, the reported GRIT gains and paired 95% Student-t intervals are $2.08\,[1.04, 3.12]$ against vanilla, $0.25\,[-0.44, 0.95]$ against prediction consistency. The latter interval includes zero; the highest mean does not establish a separation from prediction consistency. For ColoredMNIST-INV four-group GroupDRO, the corresponding intervals are $37.11\,[35.35, 38.86]$ and $6.50\,[5.99, 7.02]$. These intervals measure training-seed variability conditional on the fixed inputs.

The main table ranks displayed means: bold and italic identify the top two within each objective and dataset; solid and dotted boxes identify the top two among all entries within each dataset. Unrounded

paired contrasts can differ from differences of the displayed means. No statistical significance is inferred from these rankings.

G URBANCARS: COMBINING BACKGROUND AND OBJECT INVARIANCES

G.1 DATA, PAIR TYPES, AND ALLOCATION

UrbanCars composes car, background, and co-occurring-object images to study multiple spurious correlations (Li et al., 2023). We classify urban versus country car body types with frozen, unnormalized, 512-dimensional CLIP ViT-B/32 features. Background and object categories each agree with the label with probability 0.95 in training. There are 8,000 training, 1,000 validation, and 1,000 test images. The eight groups cross label, background alignment, and object alignment. Each label has 3,610, 190, 190, and 10 training images in the both-aligned, background-only-misaligned, object-only-misaligned, and both-misaligned groups. Validation and test each contain 125 examples per group. We use the standard compositing construction.

Each invariant pair of the first type preserves the car and co-object and replaces the background with one drawn from the opposite category's training pool. Each invariant pair of the second type preserves the car and background and replaces the co-object analogously. Neither pair endpoint supplies an additional supervised training record. Labeled examples and features are shared across methods and allocations. Construction and partner sampling use seed 1729. We reserve 1,024 training anchors for energy calibration, 2,048 for reference geometry, and three disjoint banks of 1,024 anchors. Every anchor has both partner types. An allocation uses the first 256 anchors of a bank; a fixed random ordering chooses which n_O receive object partners and which $n_B = 256 - n_O$ receive background partners. Thus changing allocation preserves the anchors and couples the collections within each bank. Both partner types use the same rendering and feature pipeline.

Equal count and equal energy. The main comparison uses the two single-type endpoints and the equal-count allocation $(n_B, n_O) = (128, 128)$. The full experiment also includes $(114, 142)$ and $(54, 202)$. For calibration, let $\bar{E}_B$ and $\bar{E}_O$ be the average squared Euclidean lengths of background and object differences in CLIP feature space. Their reported values are 61.0 and 16.2, with ratio 3.76. A target object-energy share s determines

$$n_O = \mathrm{round}\left(\frac{256 s \bar{E}_B}{s\bar{E}_B + (1-s)\bar{E}_O}\right), \qquad n_B = 256 - n_O.$$

The frozen counts are 142 for $s = 0.25$ and 202 for $s = 0.5$. The latter approximately equalizes total background and object difference energy, so it contains more pairs that vary the object because each contributes less energy on average. Equal count instead gives approximately 21% of its energy to object differences. These allocations change counts, not image intensity or pair strength; equal energy is a diagnostic choice, not an optimality condition for GRIT.

G.2 TRAINING, VALIDATION, AND METRICS

All methods fit a two-class linear probe using Adam, batch size 256, for 100 epochs. GRIT searches removal ranks $\{1, 2, 4, 8, 16, 32, 64, 128\}$, capped at the numerical rank of the pair matrix. Prediction consistency uses $\alpha \operatorname{tr}(W\widetilde{\mathbf{\Sigma}}_k W^\top)/\operatorname{tr}(\widetilde{\mathbf{\Sigma}}_k)$, where W contains the two logit coefficient vectors, and searches $\alpha \in \{10^{-3}, 10^{-2}, \ldots, 10^{6}\}$. Normalization gives a common strength grid across collections with different average difference energies.

Before the collection comparison, a validation optimizer search freezes one learning rate and weight decay per method. For GRIT and PC, the criterion averages each collection's best validation score equally over the background-only, 142-object, 202-object, and object-only collections on bank 0. This chooses settings that perform well across those collections; each allocation does not receive a separate optimizer search. The learning-rate grid was extended on validation to $\{10^{-5}, 3\cdot10^{-5}, 10^{-4}, 3\cdot10^{-4}, 10^{-3}, 3\cdot10^{-3}, 10^{-2}\}$, with weight decay in $\{0, 10^{-4}\}$. Frozen learning rates are 10^{-5} for ERM, 10^{-2} for PC, and 10^{-3} for GRIT; weight decay is zero for ERM and 10^{-4} for PC and GRIT. These optimizer choices are shared across both validation selectors. The ERM learning rate is the grid's lower edge; PC's is the upper edge, on a reported validation plateau after grid extension.

Validation selects hyperparameters and checkpoints. The primary selector maximizes the minimum validation accuracy over all eight groups. It therefore uses group metadata and examples of every shifted group. The secondary selector maximizes the same validation groups' accuracy weighted by their training frequencies: within each label the weights are $0.9025, 0.0475, 0.0475, 0.0025$ in the order above, and labels receive equal weight. Both selectors see the same balanced validation examples; only their objectives differ. For each method, allocation, bank, and selector, candidates use tuning seeds 1101–1103. The top three receive confirmation seeds 2101–2102, and their five-seed validation means choose the configuration. Lower rank or penalty strength breaks ties before candidate identity. Final fits use seeds 3101–3110 and select their own checkpoints by the specified validation criterion under the frozen configuration. Reported test metrics use the validation-selected checkpoints. The execution protocol freezes and hashes the candidate winners before unsealing test data, and checks that final refits reproduce their recorded validation curves.

Test metrics. B-shift accuracy averages the two labels' groups where background is misaligned and object aligned; O-shift reverses these roles. Both-shift averages the two labels' groups where both are misaligned. Worst-group accuracy takes the minimum of all eight group accuracies *for each fitted model*, before averaging across runs. Training-weighted test accuracy reweights the same eight test groups by the training probabilities above. It is not accuracy only on the both-aligned groups, and is distinct from choosing a checkpoint with the training-weighted validation selector.

G.3 Complete allocation results and selection sensitivity

Tables 4 and 5 retain all five allocations and five metrics under each validation selector. The primary selector favors collections covering both factors: GRIT's equal-count worst-group mean is 71.9%, compared with 56.7% and 43.8% for the single-type collections. The 142-object collection is similar at 71.7%. The equal-energy collection allocates more pairs to objects and raises O-shift accuracy to 87.5%, but its worst-group accuracy is lower at 67.9%. GRIT has higher worst-group means than PC throughout, while PC or ERM wins some individual-shift metrics.

Weighting validation toward the training distribution changes both the magnitude and ordering of these results. Equal-count GRIT reaches 51.7% worst-group accuracy, above background-only 49.3% and object-only 44.0%, but by smaller margins than under worst-group validation. The equal-energy collection drops to 42.8%, below both endpoints. Under training-weighted validation, PC selects its smallest penalty ($\alpha = 10^{-3}$) in 12 of 15 bank-level fits (balanced bank 1 selects $\alpha = 0.1$; object-only banks 0 and 1 select $\alpha = 0.01$), with worst-group means between 29.9% and 32.5%. Under primary validation, GRIT usually selects checkpoints at epochs 1–7, except when only the object varies; secondary validation typically selects epochs 35–100. Each fit runs for 100 epochs; validation selects which saved weights to retain. Per-epoch test curves are diagnostics and do not select checkpoints. Thus the observed benefit combines pair information with a validation objective that rewards performance on the shifted groups.

ERM optimizer control and large-pair reference. Table 6 separates the primary ERM baseline from plain ERM fitted with GRIT's optimizer (10^{-3} learning rate, 10^{-4} weight decay). This second row is an optimizer control, not GRIT: it applies no projection and uses no pairs. With primary selection, the two ERM worst-group means are similar (20.4% and 20.6%), although their background-shift and training-weighted accuracies differ substantially. GRIT remains above this control in worst-group accuracy for every allocation under both selectors. The reference projection uses 2,048 pairs that vary the background and 2,048 that vary the object, with rank selected on validation. Its learning rate is $3 \cdot 10^{-4}$ and weight decay zero. The larger pair budget makes it a diagnostic rather than a matched comparison or guaranteed accuracy ceiling.

Table 4: Mixed pair types improve worst-group accuracy under worst-group validation. Test accuracy (%), mean ± sample SD, for all five allocations, sorted by n_O. PC denotes prediction consistency. "Train-weighted" reweights test groups by training frequencies; it is a test metric, distinct from the validation selector. ERM uses no pairs and repeats for comparison. Bold marks the highest mean for each allocation and metric.

n_B	n_O	Method	B shift	O shift	Both shift	Worst group	Train-weighted
256	0	ERM	45.4 ± 5.4	**76.0** ± 11.7	28.5 ± 12.4	20.4 ± 9.5	83.1 ± 14.9
		PC	85.7 ± 1.4	65.1 ± 0.9	55.1 ± 1.0	52.9 ± 1.6	90.2 ± 0.5
		GRIT	**88.5** ± 0.6	68.4 ± 0.8	**60.3** ± 1.9	**56.7** ± 1.5	**92.3** ± 0.8
128	128	ERM	45.4 ± 5.4	76.0 ± 11.7	28.5 ± 12.4	20.4 ± 9.5	83.1 ± 14.9
		PC	76.0 ± 0.8	84.4 ± 0.5	71.7 ± 1.3	68.3 ± 1.3	87.5 ± 0.6
		GRIT	**79.1** ± 0.4	**85.1** ± 0.8	**74.8** ± 0.9	**71.9** ± 0.9	**89.3** ± 0.9
114	142	ERM	45.4 ± 5.4	76.0 ± 11.7	28.5 ± 12.4	20.4 ± 9.5	83.1 ± 14.9
		PC	75.9 ± 0.9	**86.0** ± 0.5	70.7 ± 1.0	67.9 ± 1.4	**89.9** ± 0.7
		GRIT	**78.2** ± 0.4	85.2 ± 0.3	**74.6** ± 0.8	**71.7** ± 1.0	89.1 ± 0.8
54	202	ERM	45.4 ± 5.4	76.0 ± 11.7	28.5 ± 12.4	20.4 ± 9.5	83.1 ± 14.9
		PC	68.8 ± 1.1	85.3 ± 1.3	65.7 ± 0.7	63.3 ± 0.6	87.7 ± 0.8
		GRIT	**73.7** ± 1.1	**87.5** ± 0.7	**70.4** ± 1.1	**67.9** ± 1.6	**89.5** ± 0.7
0	256	ERM	45.4 ± 5.4	76.0 ± 11.7	28.5 ± 12.4	20.4 ± 9.5	83.1 ± 14.9
		PC	**54.7** ± 3.2	80.2 ± 3.1	**46.9** ± 2.8	42.6 ± 2.4	82.4 ± 2.4
		GRIT	53.9 ± 0.4	**94.1** ± 0.4	45.5 ± 0.7	**43.8** ± 0.8	**95.5** ± 0.1

Table 5: Training-weighted validation reduces the benefit of mixing pair types and changes the collection ordering. Test accuracy (%), mean ± sample SD, for all five allocations, sorted by n_O. PC denotes prediction consistency. "Train-weighted" reweights test groups by training frequencies; it is a test metric, distinct from the validation selector. ERM uses no pairs and repeats for comparison. Bold marks the highest mean for each allocation and metric.

n_B	n_O	Method	B shift	O shift	Both shift	Worst group	Train-weighted
256	0	ERM	46.4 ± 2.9	**86.5** ± 1.7	16.4 ± 1.8	15.3 ± 1.5	94.6 ± 0.7
		PC	70.8 ± 1.4	81.7 ± 0.6	32.0 ± 1.8	30.0 ± 2.9	**97.2** ± 0.1
		GRIT	**87.3** ± 0.6	69.0 ± 0.7	**51.8** ± 0.4	**49.3** ± 1.3	95.6 ± 0.2
128	128	ERM	46.4 ± 2.9	**86.5** ± 1.7	16.4 ± 1.8	15.3 ± 1.5	94.6 ± 0.7
		PC	70.7 ± 1.4	81.8 ± 0.7	32.0 ± 1.9	29.9 ± 2.9	**97.2** ± 0.1
		GRIT	**80.8** ± 0.6	78.8 ± 0.2	**53.2** ± 0.5	**51.7** ± 0.8	95.5 ± 0.2
114	142	ERM	46.4 ± 2.9	**86.5** ± 1.7	16.4 ± 1.8	15.3 ± 1.5	94.6 ± 0.7
		PC	70.7 ± 1.3	81.7 ± 0.7	32.0 ± 2.0	30.1 ± 2.9	**97.2** ± 0.1
		GRIT	**78.0** ± 0.2	80.9 ± 0.2	**54.2** ± 0.4	**52.0** ± 0.3	95.3 ± 0.2
54	202	ERM	46.4 ± 2.9	86.5 ± 1.7	16.4 ± 1.8	15.3 ± 1.5	94.6 ± 0.7
		PC	**70.3** ± 1.1	82.9 ± 0.5	34.7 ± 1.3	32.5 ± 2.0	**97.2** ± 0.1
		GRIT	54.9 ± 0.7	**93.5** ± 0.6	**45.6** ± 0.7	**42.8** ± 1.8	95.2 ± 0.2
0	256	ERM	46.4 ± 2.9	86.5 ± 1.7	16.4 ± 1.8	15.3 ± 1.5	94.6 ± 0.7
		PC	**70.5** ± 0.6	82.7 ± 0.5	34.0 ± 1.3	32.1 ± 2.3	**97.1** ± 0.1
		GRIT	53.9 ± 0.7	**93.8** ± 0.6	**45.5** ± 0.3	**44.0** ± 0.7	95.5 ± 0.1

Table 6: Optimizer and pair-budget controls on UrbanCars. Accuracy (%), mean ± sample SD over ten final seeds. WG and TW identify worst-group and training-weighted *validation* selection; the final column is training-weighted *test* accuracy. Controls use no allocation-specific pair banks.

Validation	Method	B shift	O shift	Both shift	Worst group	Train-weighted
WG	ERM	45.4 ± 5.4	76.0 ± 11.7	28.5 ± 12.4	20.4 ± 9.5	83.1 ± 14.9
	ERM, matched	66.0 ± 4.0	82.5 ± 0.8	25.8 ± 4.3	20.6 ± 6.6	96.8 ± 0.5
	Reference	84.0 ± 1.2	83.8 ± 0.7	82.0 ± 0.4	77.4 ± 0.9	88.9 ± 0.9
TW	ERM	46.4 ± 2.9	86.5 ± 1.7	16.4 ± 1.8	15.3 ± 1.5	94.6 ± 0.7
	ERM, matched	69.4 ± 1.1	81.5 ± 0.3	31.3 ± 0.8	27.4 ± 1.9	97.1 ± 0.0
	Reference	93.7 ± 0.3	66.4 ± 0.6	62.5 ± 0.6	61.4 ± 1.1	95.0 ± 0.2

G.4 Uncertainty and interpretation

PC and GRIT means first average the ten final training seeds within each of the three fixed banks, then average the banks equally. ERM and reference rows average ten seeds without bank-specific refits. For a paired contrast, the protocol pairs methods by final seed, averages the bank-specific differences within each seed, and constructs a 95% Student-t interval over those ten differences. This measures training variability conditional on the three banks and fixed test examples. The reported protocol also uses 2,000 within-group test-example bootstrap resamples for selected contrasts, addressing a different source of uncertainty. Neither procedure establishes variation across arbitrary new acquisitions.

Paired intervals clarify these comparisons. Under primary selection, GRIT minus PC on pairs that vary only the object has worst-group difference $1.3\,[-0.5, 3.0]$ percentage points (mean and 95% seed interval). Its interval includes zero, so the higher mean does not establish a separation. Comparing GRIT's 202-object and 142-object collections gives worst-group difference $-3.8\,[-4.8, -2.9]$ despite an O-shift gain of $2.3\,[1.8, 2.9]$. The 202-object collection exceeds matched-optimizer ERM in worst-group accuracy by $47.2\,[42.6, 51.9]$. Under training-weighted validation, that same collection falls below background-only GRIT by 6.5 points, with signed difference $-6.5\,[-8.3, -4.8]$. Contrasts of unrounded values can differ from subtraction of the displayed means.

Tables 4–6 report means and sample standard deviations. For PC and GRIT, the sample standard deviation is across the ten bank-averaged seed scores, with denominator nine; for ERM it is across the ten ordinary seed scores, as for the reference projection. The corresponding standard error is $\mathrm{SD}/\sqrt{10}$. Treating thirty bank-by-seed fits as independent, or recovering row SDs from paired contrast intervals, would give a different or unidentified quantity.

The coverage interpretation is qualitative in this feature space. Reference background and object covariances require 151 and 166 directions, respectively, to explain 90% of their energy, and their leading subspaces overlap strongly. These empirical covariances are not ground-truth unstable subspaces. Consequently, two semantic factors do not mean two orthogonal feature directions, and these results do not measure the theorem's clean weakest eigenvalue or establish its pair-noise assumptions. CueMNIST supplies the complementary known-geometry experiment. The claim supported here is that combining the two supplied relations improves worst-group means under the specified selection procedure.

G.5 Construction fidelity and component overlap

The construction follows the UrbanCars generator's component pools, split rules, image placement, and JPEG output: a 256×256 canvas with a centered car and the co-object at the right edge. The upstream generator revision is `c91b0970`. CLIP preprocessing retains the full canvas. Sixteen car images required fallback masks: fourteen used reduced MaskFormer confidence thresholds and two used GrabCut. The remaining masks and pool order were unchanged. A reported comparison of 200 decoded images per split found 597 of 600 pixel-identical to the original generator with the same masks; the remaining three differed by at most three intensity levels on a few pixels.

Component IDs are split-disjoint, but the construction audit records six pairs of byte-identical files under different names in different upstream splits: three Places backgrounds and three Stanford Cars photographs. It reports that at most six of the 2,000 validation and test images share one component with a training image, with no shared full composite. The audit also records 908 COCO photos supplying distinct object instances to different splits. These inherited overlaps were retained to preserve the upstream construction. Thus component-ID separation should not be interpreted as complete separation of source photographs or bytes.

H CueMNIST and controlled pair-information sweeps

H.1 Independent sweeps of weakest signal, pair noise, and pair count

These experiments vary the smallest clean pair-covariance eigenvalue, pair-noise level, and number of pairs separately. CueMNIST, MNIST (LeCun et al., 1998) with two independently controlled color cues, supplies the running example; a synthetic task provides a second controlled setting.

Both use $d = 96$ fixed features and two orthonormal cue directions, collected in $\boldsymbol{U} = [\boldsymbol{u}_B, \boldsymbol{u}_O]$. The known unstable subspace is $\mathcal{U} = \mathrm{span}(\boldsymbol{U})$, with $r_* = 2$, and $\mathcal{S} = \mathcal{U}^\perp$. Task examples, target shifts, feature geometry, and labeled sample size remain fixed across all sweep conditions within a replicate.

CueMNIST task and fixed features. The binary label is one for an MNIST digit less than five and zero otherwise. Each replicate uses 3,000 training identities and 1,000 disjoint validation identities from the official training pool, and 1,000 identities from the official test pool, with equal numbers of the two labels in every split. A 40×48 float RGB canvas contains the unchanged 28×28 grayscale digit, a colored frame B, a separate colored header O, and a strip of 30 color patches. With $t = 2y - 1$, the cue values are

$$b = 2.5q_B + 0.3\eta_B, \qquad o = 2.5q_O + 0.3\eta_O, \qquad \eta_B, \eta_O \overset{\text{iid}}{\sim} \mathcal{N}(0, 1),$$

where the signs q_B, q_O independently equal t with source probability 0.95. The strip coordinates are independent $\mathcal{N}(0, 0.2^2)$ variables. Each scalar v is rendered as color $(a(v), 1 - a(v), 1/2)$, where $a(v) = (1 + \tanh(v/4))/2$. Regional extraction recovers v from the mean red channel using $4 \operatorname{arctanh}(2a - 1)$, avoiding clipping the Gaussian cue values.

The extractor adaptively average-pools the digit to 8×8 values, subtracts the training-coordinate means, and applies one scalar normalization so the training digit vectors have root-mean-square norm five. Concatenating the 64 digit values, two decoded cues, and 30 decoded strip values gives 96 coordinates. A random orthogonal embedding per replicate hides the cue axes. It is obtained by QR factorization of a standard-Gaussian 96×96 matrix, with signs chosen so the diagonal of R is positive. Normalization uses training images only, and the feature map remains fixed during fitting. Batch construction implements these regional operations analytically; a test verifies equality with extraction from rendered float canvases. This semi-synthetic task uses the engineered feature map described above. Gaussian pair noise, defined below, is added after feature extraction and need not correspond to a valid pixel-space perturbation.

Validation and test contain all four cue-alignment combinations for each identity, producing 4,000 views per split. The four views share the digit, strip, and within-cue Gaussian variations. Eight groups combine the binary label with the two cue-alignment indicators. Both-cue-reversal accuracy averages the two label groups in which both cues oppose the label. Single-cue shifts and source-weighted metrics are also retained. For each label, the four joint source-group probabilities are $\frac{1}{2}(0.9025, 0.0475, 0.0475, 0.0025)$ for both aligned, only B misaligned, only O misaligned, and both misaligned, respectively. Ten replicates redraw the split, embedding, nuisance values, and pair noise. Their uncertainty describes construction variability on the fixed MNIST dataset; the four views of an identity are not treated as independent replicates.

Synthetic task. Each of twenty replicates draws mutually orthogonal $\boldsymbol{S} \in \mathbb{R}^{96\times 4}$, $\boldsymbol{u}_B$, and $\boldsymbol{u}_O$, and an independent orthogonal matrix $\boldsymbol{V}$. These directions and rotations use the same Gaussian QR construction. For a uniform binary label y and $t = 2y - 1$, examples follow

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{S}\boldsymbol{s} + \boldsymbol{u}_B b + \boldsymbol{u}_O o + \boldsymbol{V}\,\mathrm{diag}(1, 2^{-1/2}, \ldots, 96^{-1/2})\boldsymbol{\eta}, \\ \boldsymbol{s} &= 1.28t\,\mathbf{1}_4/2 + \boldsymbol{\eta}_S, \qquad b = 2.5q_B + \eta_B, \qquad o = 2q_O + \eta_O. \end{aligned}$$

All displayed Gaussian variables are independent standard Gaussian variables of the required dimensions. The cue signs have the same independent 0.95 source alignment as in CueMNIST. There are 3,000 labeled training examples, 160 validation examples per group, and 400 test examples per group. The anisotropic Gaussian feature noise is part of the source generator; it is distinct from the isotropic acquisition noise added to pair differences. The same group definitions and source weights are used for both tasks.

Pairs with an exact prescribed spectrum. Let g denote the smallest clean eigenvalue and fix the largest eigenvalue at one. For even k, the clean differences alternate between

$$\boldsymbol{u}_j = \begin{cases} \sqrt{2}\,\boldsymbol{u}_B, & j \text{ odd}, \\ \sqrt{2g}\,\boldsymbol{u}_O, & j \text{ even}. \end{cases} \qquad \boldsymbol{\Sigma}_k = \frac{1}{k}\sum_{j=1}^{k} \boldsymbol{u}_j \boldsymbol{u}_j^\top = \boldsymbol{u}_B \boldsymbol{u}_B^\top + g\boldsymbol{u}_O \boldsymbol{u}_O^\top.$$

Consequently, the clean empirical eigenvalues on $\mathcal{U}$ are exactly $(1, g)$ for every tested budget. The runner constructs these differences analytically and assigns them to recorded training anchors, cycling through the 3,000 training examples when necessary. The ideal partner of anchor $\boldsymbol{z}_{a_j}$ is $\boldsymbol{z}_{a_j} - \boldsymbol{u}_j$. It changes one cue while preserving the task content and all other coordinates. The CueMNIST renderer and extractor realize these clean scalar cue interventions exactly; noisy image endpoints are not materialized. No partner is added to the labeled training sample.

Observed differences satisfy $\boldsymbol{\delta}_j = \boldsymbol{u}_j + \sigma_\xi \boldsymbol{\epsilon}_j$, with independent $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{96})$ independent of the task data. A master array supplies the same Gaussian draws across conditions and nested prefixes across budgets. Anchor indices and clean differences likewise have nested prefixes when g is fixed. Here k counts independent noisy measurements, including repeated measurements at reused training anchors; it does not count distinct digit identities. Conditioning on the designed clean signals leaves the k acquisition noises independent, as required by the pair-noise model. Neither clean nor observed differences are centered, and noisy differences are not renormalized. The pair covariance is $\widetilde{\boldsymbol{\Sigma}}_k = k^{-1} \sum_j \boldsymbol{\delta}_j \boldsymbol{\delta}_j^\top$, equivalently the covariance obtained by assigning half of each pair's weight to each orientation.[2]

Three primary sweeps. The primary sweeps share the reference $(g, \sigma_\xi, k) = (1, 0.1, 256)$. The prespecified grids are

$$
\begin{aligned}
g &\in \{1/64, 1/32, 1/16, 1/8, 1/4, 1/2, 3/4, 7/8, 1\}, \\
\sigma_\xi &\in \{0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3, 0.45, 0.6, 0.8, 1\}, \\
k &\in \{64, 96, 128, 192, 256, 384, 512, 1024, 2048, 4096\}.
\end{aligned}
$$

Each sweep holds the other two quantities at their reference values. Changing g holds the other eigenvalue fixed, so clean trace is $1 + g$; this sweep does not fix total clean energy. Changing k preserves both clean eigenvalues exactly. The expected noisy trace is $1 + g + 96\sigma_\xi^2$. A $g = 0$ control at the reference noise and budget omits the second clean direction and lies outside the spanning assumption. A $\sigma_\xi = 0$ control at the reference spectrum and budget isolates exact geometry. Both controls are reported separately and excluded from logarithmic fits. Conditions shared across axes reuse the same fitted models.

Secondary geometry settings. A follow-up specified after inspecting the primary results adds two nearby settings around each reference geometry curve. The eigenvalue sweep uses $\sigma_\xi \in \{0.05, 0.1, 0.2\}$ at $k = 256$; the noise sweep uses $k \in \{128, 256, 512\}$ at $g = 1$; and the pair-count sweep uses $\sigma_\xi \in \{0.05, 0.1, 0.2\}$ at $g = 1$. All original x values, replicates, clean signals, and Gaussian noise streams are retained, including nested noise prefixes across budgets. The runner reconstructs each known subspace and master noise array from the deterministic construction. It requires matching hashes for the original subspaces, clean and noisy differences, noise, and anchor indices, and checks default subspace errors within 10^{-12} of the frozen measurements. This follow-up computes subspace error only and reuses the primary predictor results. The three top-row curves show these geometry settings; bottom-row accuracies and the appendix's fitted-instability mirrors retain the primary defaults. The secondary settings were fixed before their computation.

Predictor fitting and validation. For $f_i = \boldsymbol{w}^\top \boldsymbol{z}_i + c$, all methods minimize binary logistic empirical risk plus $0.1\|\boldsymbol{w}\|_2^2/2$, with an unpenalized intercept. GRIT additionally imposes $P_{k,2}\boldsymbol{w} = 0$, where $P_{k,2}$ removes the leading two pair-covariance directions. The shared SVD implementation excludes singular values below 10^{-4} of the largest and caps the requested rank at numerical rank. We fix $r = r_* = 2$ to isolate estimation; this experiment does not select rank or assume that the correct rank is known in general applications. ERM and projection onto the true $\mathcal{U}^\perp$ are fitted once per replicate as controls.

Quadratic prediction consistency (CR) adds

$$
\frac{\alpha}{2\,\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_k)} \boldsymbol{w}^\top \widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{w}, \qquad \alpha \in \{0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}.
$$

[2] Symmetrization leaves the matrix unchanged and does not increase the number of independent measurements, which remains k.

The factor $1/2$ matches the sum of squared differences of two symmetric logits $(-f/2, f/2)$. For comparison with the score-penalty coefficient in the theory, the effective coefficient is $\lambda = \alpha/[2\,\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_k)]$. We record both coefficients because changing the pair spectrum or noise changes the trace. Each condition's CR strength maximizes worst-group validation accuracy, with ties resolved toward smaller α. The balanced validation groups expose cue misalignment to this selector.

Fitting uses deterministic float64 Newton steps, zero initialization, Armijo backtracking, at most 80 iterations, and a required gradient infinity norm no greater than 10^{-7} in the optimization coordinates (retained features and intercept). Backtracking starts at step size one and halves it until the objective decrease is at least 10^{-4} times the step size times the gradient–Newton-step inner product. A step size below 10^{-12} or failure to meet the gradient tolerance terminates the run as unconverged. Every candidate, selection, coefficient hash, and fitting-input hash is written to a frozen record before that replicate's test views are generated. The selected CR model, fixed-rank GRIT, ERM, and true-subspace control are then evaluated without further selection. The master seed is 20260927. Source files, package versions, dataset checksums, content identities, coefficients, and evaluation-input hashes are saved with the resolved configuration.

Geometry, instability, and uncertainty. For the effective coefficient in the original feature coordinates, define

$$e_{\mathrm{sub}} = \|(\boldsymbol{I} - P_{k,2})\boldsymbol{U}\|_{\mathrm{op}}, \qquad s_{\mathrm{rel}} = \frac{\|\boldsymbol{U}^\top \boldsymbol{w}\|_2}{\|\boldsymbol{w}\|_2}.$$

The first quantity is the largest missed unstable-direction component. The second measures the fitted predictor's unstable sensitivity relative to its coefficient norm. Subspace error is a GRIT diagnostic; CR does not estimate a removal subspace. For a nonzero hard-projected predictor, $s_{\mathrm{rel}} \le e_{\mathrm{sub}}$ holds deterministically; the pipeline checks this inequality. Normalization is undefined for a zero coefficient, and the report rejects coefficients with norm at most 10^{-12}. Raw sensitivity $\|\boldsymbol{U}^\top \boldsymbol{w}\|_2$, coefficient norm, logistic loss, and shifted accuracy are retained alongside the normalized quantities. The risk decomposition uses raw sensitivity, so the normalized plot must be interpreted together with the coefficient norm.

Curves display replicate means and sample standard deviations. Endpoint contrasts use matched replicate differences with 5,000 percentile-bootstrap resamples and base seed 260925. The pair-count reference curve is proportional to $k^{-1/2}$ and anchored to mean subspace error at the prespecified $k = 256$; its vertical normalization is descriptive. Log–log slopes are fit separately per replicate using $k \ge 512$, giving the four points $512, 1024, 2048, 4096$. Replicate resampling gives uncertainty for the mean slope. Any normalized-sensitivity slope describes the fitted predictors and is not asserted to equal the subspace-error slope.

Connection to the finite-pair bound. Let $h = d + \log(1/\eta)$. The GRIT theorem's pair-dependent factor is

$$T(g, \sigma_\xi, k) = \left(\frac{\sigma_\xi}{\sqrt{g}} + \frac{\sigma_\xi^2}{g}\right)\sqrt{\frac{h}{k}} + \frac{\sigma_\xi^2}{g}\frac{h}{k}.$$

The largest clean eigenvalue does not appear in this factor. The three sweeps separately control its weakest-signal, noise, and finite-pair inputs. In the larger-k regime, controlled noise-to-signal ratio yields a $k^{-1/2}$ subspace-estimation scale; the full expression retains its $1/k$ contribution. Weak-signal and high-noise settings can instead saturate at subspace error one. The displayed shape reference does not estimate unknown theorem constants or supply a calibrated bound.

The risk theorem also contains source prediction risk, bridge covariate shift, and predictor-dependent factors; its extra-rank term vanishes at the fixed correct rank studied here. Finite-label ridge-logistic fitting differs from the theorem's unregularized population optimization. Accordingly, the sweeps examine how pair information affects estimated geometry and fitted instability, alongside target performance. The complete target-risk bound additionally requires the source-risk, bridge, and predictor-dependent quantities for each fitted model. The CR comparison uses the same supplied pair information and the stated validation rule.

H.2 Results and interpretation

Figure 2 pairs GRIT's subspace recovery with the target accuracy of GRIT and validation-selected CR; Figure 3 gives the Gaussian counterpart. Figures 4 and 5 compare subspace error with normalized fitted instability for the same GRIT predictors. The numerical contrasts below use the primary settings.

Weakest-direction strength. At fixed $k = 256$, $\sigma_\xi = 0.1$, and largest clean eigenvalue one, increasing g from $1/64$ to one reduces CueMNIST subspace error from 0.535 to 0.064. Normalized fitted instability falls from 0.469 to 0.029. Target accuracy increases from 3.43% to 75.86%, a paired gain of 72.43 percentage points (95% bootstrap interval: 70.11–74.57). This isolates weakest-direction strength: the two directions already span $\mathcal{U}$ at every positive g, and the largest eigenvalue stays fixed. The separate nonspanning control has subspace error 0.997 and target accuracy 0.02%. The noiseless full-span control reaches the true-subspace reference accuracy of 80.31%, with subspace error below 10^{-14}.

Noise and pair count. With clean spectrum $(1, 1)$ and $k = 256$, raising noise SD from 0.025 to one increases CueMNIST subspace error from 0.016 to 0.780 and normalized instability from 0.002 to 0.673. At the upper noise levels, prediction can fail despite the clean collection spanning the full unstable subspace. Conversely, at noise SD 0.1 and the same clean spectrum, increasing k from 64 to 4,096 reduces subspace error from 0.129 to 0.016 and raises target accuracy from 60.45% to 79.97%. The subspace-error log–log slope over the prespecified $k \geq 512$ range is -0.500 (95% interval: $[-0.519, -0.482]$). Normalized instability has a steeper empirical slope, -1.004 ($[-1.209, -0.839]$). The geometric bound is an upper bound for this quantity, so it need not describe the rate of every fitted predictor.

Synthetic replication. The Gaussian task gives the same qualitative dependencies (Figure 3). Its weakest-eigenvalue sweep reduces subspace error from 0.568 to 0.064 and improves target accuracy from 55.49% to 88.43%. Increasing pair count from 64 to 4,096 reduces subspace error from 0.128 to 0.016, with a large-k slope of -0.497 (95% interval: $[-0.511, -0.482]$). The synthetic and CueMNIST geometric curves are similar because both have the same prescribed signal eigenvalues, dimension and isotropic pair-noise model. Their fitted sensitivities and target accuracies also depend on their different labeled-data distributions.

GRIT and consistency regularization. At the shared reference setting, GRIT and CR attain 75.86% and 71.58% on CueMNIST, and 88.43% and 86.77% on the synthetic task. On CueMNIST, CR exceeds GRIT at $k = 64$ (63.83% versus 60.45%) and at noise SD 0.3 (39.14% versus 35.94%). At $k = 4096$, GRIT attains 79.97% versus 71.40% for CR on CueMNIST, and 89.61% versus 87.18% on the synthetic task. The figures retain all tested noise, signal-strength, and budget values. The ERM and true-subspace controls use the same labeled data and fitting procedure, respectively retaining all features or removing the known cue subspace.

With few pairs, CR's soft penalty can yield lower fitted instability than hard removal using an imperfect subspace: at $k = 64$, CueMNIST normalized instability is 0.056 for CR and 0.091 for GRIT. As k increases, GRIT estimates the unstable subspace more accurately. In contrast, the mean noisy pair covariance is $\boldsymbol{\Sigma}_k + \sigma_\xi^2 \boldsymbol{I}$, so CR's penalty also acts on stable directions even as sampling fluctuations decrease. The large-budget fitted diagnostics are consistent with this tradeoff: at $k = 4096$, normalized instability is 0.0022 for GRIT versus 0.0503 for CR on CueMNIST, and 0.0021 versus 0.0386 on the synthetic task. These diagnostics describe the association between geometry, fitted sensitivity, and accuracy in the tested settings.

The CR strength reaches the largest tested value, $\alpha = 1000$, in nine of ten CueMNIST replicates and four of twenty synthetic replicates at $k = 4096$ (and all ten CueMNIST replicates at $k = 64$). The comparison therefore concerns the stated finite grid and validation rule.

Reproducing the comparisons. The construction, grids, objective, selector, and estimators in Appendix H.1 specify the experiment directly. For each replicate, construct the fixed labeled features and cue subspace, then form the clean differences and one master Gaussian-noise array. For each sweep condition, take the required prefix, add the specified noise, and compute the leading two

singular directions. Fit GRIT and every CR candidate on the same labeled training examples, select CR strength by worst-group validation accuracy, and freeze the choices before evaluating the test groups. Fit ERM and the true-subspace control once per replicate. Aggregate test metrics across the ten CueMNIST or twenty synthetic replicates; use matched replicate differences for contrasts and replicate resampling for their intervals. The primary fits used CPU float64 and took under a minute after environment and data setup.

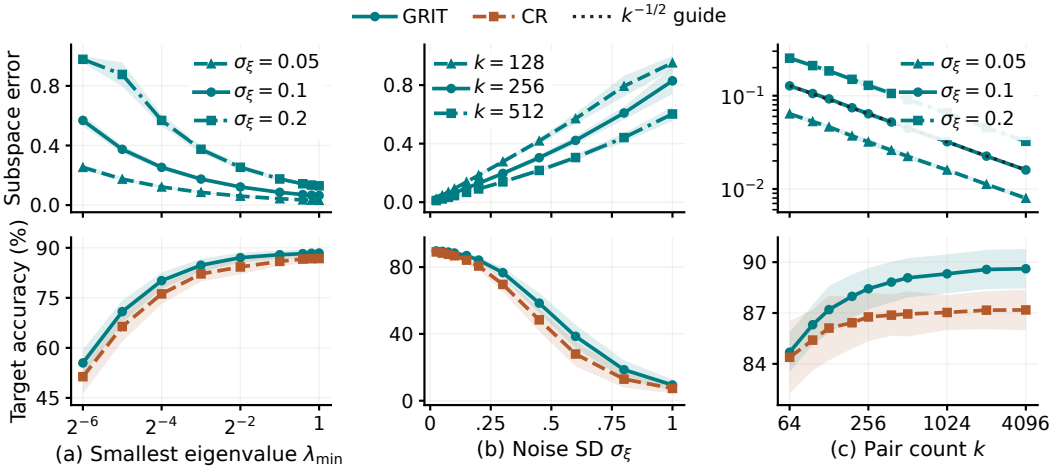

Figure 3: **Stronger signal in the weakest direction and less noise improve recovery and target accuracy; more pairs yield near-$k^{-1/2}$ subspace error and a GRIT advantage over CR at larger budgets.** On the Gaussian synthetic task, the top row shows GRIT's subspace error; the bottom compares target accuracy. Defaults are $k = 256$, $\sigma_\xi = 0.1$, and clean spectrum $(1, 1)$. Top-panel legends vary a second setting; bottom panels use the defaults. Bands show sample SD across 20 replicates; the dotted line is a shape guide. Y-axis ranges differ by panel.

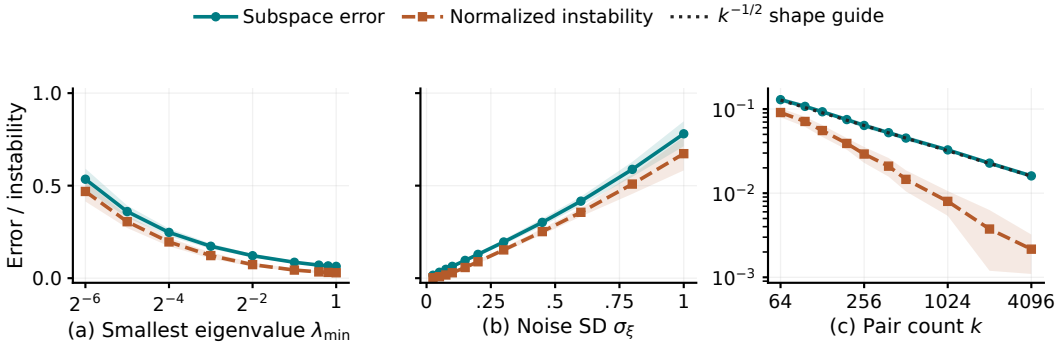

Figure 4: **GRIT's fitted instability follows the recovery trends and decreases faster than subspace error as pair count grows.** Mirror of the three sweeps on CueMNIST, adding normalized predictor sensitivity to the subspace-error curves. Defaults are $k = 256$, $\sigma_\xi = 0.1$, and clean spectrum $(1, 1)$; bands show sample SD. The dotted $k^{-1/2}$ line describes subspace recovery, not a predictor-rate guarantee.

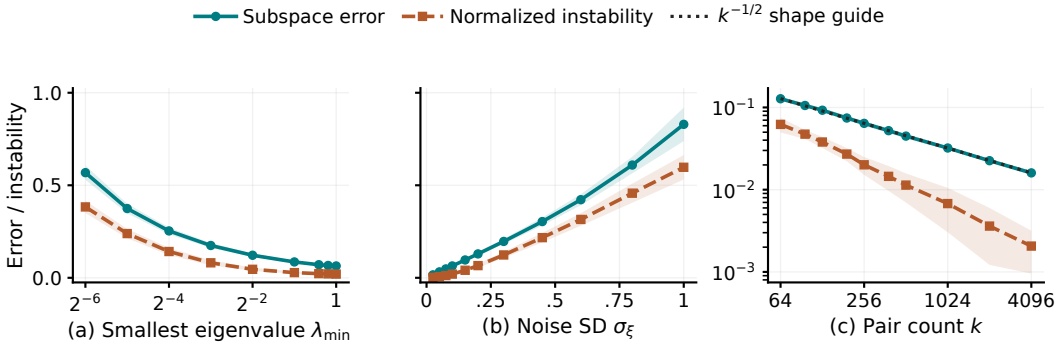

Figure 5: **GRIT's fitted instability follows the recovery trends and decreases faster than subspace error as pair count grows.** Mirror of the three sweeps on the Gaussian synthetic task, adding normalized predictor sensitivity to the subspace-error curves. Defaults are $k = 256$, $\sigma_\xi = 0.1$, and clean spectrum $(1, 1)$; bands show sample SD. The dotted $k^{-1/2}$ line describes subspace recovery, not a predictor-rate guarantee.

I Further Discussion and Future Directions

I.1 Extending the predictor and the invariance geometry

Two distinct linearity assumptions shape the present analysis. A1 places the stable information in a linear subspace of the fixed, potentially nonlinear representation; our risk guarantees additionally use a linear probe on that representation. Replacing the probe with a nonlinear head is straightforward empirically: the head can be fitted on the same projected features, and remains exactly invariant to translations in the removed subspace. We expect extensions of the risk analysis to be feasible with suitable control of the head's sensitivity to projection error. Relaxing the probe therefore leaves the underlying geometric model intact.

The deeper challenge is to relax A1 itself, allowing stable information to depend on a nonlinear map of the features. A single subspace and its orthogonal complement may then be insufficient to describe the variation that prediction should ignore. Nonlinear PCA offers a useful analogy: kernel PCA and neural approaches introduce choices about feature geometry and learned mappings beyond a linear eigenspace (Schölkopf et al., 1998; Kramer, 1991). Learning such geometry from invariant pairs would also require identifying which variation can be removed while preserving predictive information. Developing this extension is a promising direction for building on GRIT's foundation of estimating invariance before enforcing it.

I.2 Pair acquisition, validity, and empirical scope

Progress in either direction also depends on the information supplied by the pairs. Acquiring pairs and checking that they preserve task information remain application-specific tasks. Pooling compatible sources can improve coverage, but their relative counts and scales affect which variations dominate the covariance. Moreover, preserving task content does not by itself establish our geometric or noise assumptions; biased or dependent acquisition may require a different analysis. The current bounds explain dependencies rather than certify individual experimental runs, and the controlled noise study does not establish why GRIT improves performance on pretrained features. Broader studies of pair acquisition and validity can test whether invariant pairs can serve as a reusable interface between application-specific knowledge and robust learning.