FROM OBJECTIVES TO WHAT MODELS LEARN: A LANDAU THEORY OF INVARIANT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Invariant-learning objectives pursue similar goals yet produce qualitatively different regularization paths, leaving unclear when shortcuts can be suppressed without damaging stable modes. Our starting point is simple: in the desired shortcut-suppressed regime, shortcut loading is small, so the objective's low-order expansion governs local stability and residual amplitude. This brings the problem into the domain of Landau phase-transition theory. We establish a mathematical isomorphism between the near-critical normal form of predictive-mode learning and Landau theory, identifying the learning objective as an effective free energy and critical-mode amplitude as an order parameter. The resulting low-order objective signatures predict distinct regularization phenotypes: quadratic R_2 terms shift phase boundaries and enable finite-strength elimination, whereas quartic R_4 terms continuously attenuate acquired modes while leaving nonzero residual loading at finite strengths. Higher-order terms may further shape nonlinear tails. In a canonical bilinear model, we derive exact phase boundaries and equilibrium loadings, identifying conditions for a selective-retention window in which shortcut suppression preserves stable structure. Controlled bilinear and ReLU experiments, including an MNIST construction, support the predicted signature–phenotype relation, while coupled-feature experiments validate its extension to collective modes. The framework connects the mathematical structure of invariant-learning objectives to what models learn as regularization varies.

1 INTRODUCTION

The central promise of invariant learning is simple: exploit heterogeneity across training environments to learn predictive structure that survives distribution shift (Peters et al., 2016; Arjovsky et al., 2019; Sagawa et al., 2020; Krueger et al., 2021). In practice, however, objectives built around this common principle can behave very differently, while their formulas provide little indication of what will happen along the regularization path (Gulrajani & Lopez-Paz, 2021; Zhang et al., 2023). This objective–behavior gap turns objective choice and strength selection into an empirical search problem (Chen et al., 2023). Fundamentally, it leaves unclear whether an objective can distinguish stable from shortcut modes (Geirhos et al., 2020) and suppress the latter without collateral damage.

These limitations call for a structural account of objective behavior: one that makes the regularization phenotype predictable directly from the form of the loss. Hard preselection is an incomplete solution, as spuriousness is mode-wise and is often distributed across interacting collective modes rather than isolated in known, single variables. Conventionally, invariant regularization is viewed as a continuous trade-off, reducing model selection to hunting for a single, fragile performance peak. In reality, suppression is not merely a continuous slide: a change in the local stability of the zero-loading state marks a sharp phase transition. As distinct modes hit this critical boundary at different strengths, the ideal outcome is not a point of compromise, but a robust selective-retention window.

The key to predicting these regularization paths lies in a simple mathematical reality: whether a shortcut is continuously attenuated or eliminated at a finite threshold, its loading in the desired selective-retention state is intrinsically small. The global nonlinear complexity of an invariant penalty is therefore largely irrelevant to this near-critical behavior. What governs the mode is the low-order expansion of the full objective, which determines the stability of the shortcut-free state and the leading behavior of small residual loading. This is precisely the domain of Landau phase-

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.

transition theory, which describes the emergence, attenuation, and loss of macroscopic order through a low-order free-energy expansion. Landau theory has successfully characterized critical behavior across diverse physical systems, from magnets to superconductors and ferroelectrics (Landau, 1937; Landau & Lifshitz, 1980; Ginzburg & Landau, 1950; Hohenberg & Krekhov, 2015).

We establish a mathematical isomorphism between the near-critical normal forms of predictive-mode learning and Landau theory. Through the lens of magnetization, the learning objective acts as an effective free energy, the critical-mode amplitude as the order parameter, regularization strength as a temperature-like control, and environment-dependent support as a mode-dependent external field. A mode that persists without such support resembles spontaneous ferromagnetic order; a shortcut driven mainly by the field resembles a paramagnetic response; and an intrinsically predictive but environment-sensitive mode resembles a soft magnet. Landau theory then predicts mode onset, equilibrium loading, and phase boundaries, as well as how learned modes respond to weak loss perturbations and how quickly they relax toward equilibrium near criticality.

Starting from concrete invariant-learning objectives, we reduce their local equilibria near a phase boundary to a low-order Landau free energy. Its coefficients form a low-order objective signature of the full loss. A quadratic correction R_2 changes the effective mass and moves the phase boundary, much as temperature controls the onset of magnetization; a quartic correction R_4 changes the quartic stiffness and regulates post-onset amplitude; and higher-order terms shape behavior away from criticality. These signatures induce distinct regularization phenotypes, ranging from finite-threshold elimination and gradual attenuation to plateaus, non-monotonic tails, instability, or global collapse. They therefore organize invariant objectives by their regularization phenotypes rather than treating all regularizers as interchangeable penalties. In the canonical bilinear model, this correspondence additionally yields exact phase boundaries and equilibrium loadings.

Contributions. **(i) Ferromagnetic phase-transition perspective.** We show that predictive-mode learning under invariant regularization admits a ferromagnetic Landau description, unifying mode onset, suppression, and selective retention with the language of order parameters, critical points, temperature-like controls, and spin-like collective modes. **(ii) Objective signatures as a predictive methodology.** We derive a Landau reduction in which effective quadratic, quartic, and higher-order terms form a low-order objective signature that predicts phase boundaries, post-onset loadings, and regularization phenotypes. **(iii) Structural analysis of existing objectives.** We apply this signature methodology to widely used invariant-learning objectives and obtain method-specific predictions, including finite-strength elimination, continuous shrinkage, nonlinear tails, instability, and collapse. **(iv) Empirical validation.** Controlled bilinear and ReLU experiments, including an MNIST-based construction, confirm the predicted phase boundaries, loading laws, and method-specific regularization phenotypes; coupled-feature experiments further validate the collective-mode extension.

2 Related Work

Landau theory describes ordering transitions through symmetry-constrained free-energy expansions. Its Ginzburg–Landau and relaxational extensions connect this framework to superconductivity, spatially varying order, and critical dynamics (Hohenberg & Halperin, 1977; Hohenberg & Krekhov, 2015). Related statistical-physics analyses of neural-network learning (Seung et al., 1992; Watkin et al., 1993) have studied online learning in multilayer networks (Saad & Solla, 1995), specialization transitions in committee machines (Biehl et al., 1998; Aubin et al., 2018), order-to-chaos transitions in random networks (Poole et al., 2016), and jamming near interpolation (Geiger et al., 2019).

Exact analyses of deep linear networks reveal distinctive learning dynamics, including plateaus and sequential mode learning (Saxe et al., 2014), while complementary landscape analyses characterize their local-minimum structure and critical-point geometry (Laurent & von Brecht, 2018; Achour et al., 2024).

Invariant learning seeks predictors that remain stable across environments. This principle appears in invariant causal prediction (Peters et al., 2016), IRM and game-theoretic formulations of invariance (Arjovsky et al., 2019; Ahuja et al., 2020), GroupDRO (Sagawa et al., 2020), REx/V-REx (Krueger et al., 2021), Deep CORAL (Sun & Saenko, 2016), and MMD alignment (Gretton et al., 2012). Existing theory has established limitations of finite-environment invariance and practical IRM objectives (Rosenfeld et al., 2021; Kamath et al., 2021), while DomainBed highlights the sen-

sitivity of domain-generalization results to hyperparameter and model-selection choices (Gulrajani & Lopez-Paz, 2021). WILDS further documents substantial performance degradation under naturally occurring distribution shifts (Koh et al., 2021).

3 FORMULATION

We consider supervised learning across training environments $e \in \mathcal{E}_{\text{tr}}$ with joint distributions $P_e(X, Y)$. A stable feature remains predictive across environments, whereas a shortcut is predictive only under particular training conditions and may fail under distribution shift. We study equilibria of the learning objective, namely the stable states reached at the end of optimization. We focus on identifiable low-dimensional predictive modes, which isolate the structural effect of the learning objective from feature-discovery, optimization, and model-selection variability. This controlled setting makes mode onset, suppression, and retention directly measurable and enables quantitative, testable predictions of objective-induced behavior.

As in Landau theory, the form of the order parameter depends on the system. For a general finite network, we denote by q the effective order parameter near a simple continuous transition; it represents the parameter combination that becomes soft at criticality. For closed-form analysis, we use the bilinear parameterization $\beta = w\theta$, in which w provides a concrete realization of q; Theorem B.2 gives the general construction.

Because the effective predictive coefficient $\beta = w\theta$ is unchanged under $(w, \theta) \mapsto (-w, -\theta)$, the objective obeys $\mathcal{F}(-w, -\theta; \lambda) = \mathcal{F}(w, \theta; \lambda)$. This is the $\mathbb{Z}_2$ symmetry of a zero-field Landau free energy. Assume that the θ direction remains noncritical at the reference critical point, $\partial_{\theta\theta}\mathcal{F}(0, 0; \lambda_c) > 0$. The stationary condition $\partial_\theta \mathcal{F} = 0$ then uniquely determines, near the origin,

$$\theta^\star(w; \lambda) = \kappa(\lambda) w + \mathcal{O}(w^3). \tag{1}$$

Thus w is a valid local order-parameter coordinate, with θ locally slaved to it. Define

$$\mathcal{F}_{\text{eff}}(w; \lambda) = \mathcal{F}(w, \theta^\star(w; \lambda); \lambda) = L_0 + \frac{1}{2} r(\lambda) w^2 + \frac{1}{4} u(\lambda) w^4 + \mathcal{O}(w^6). \tag{2}$$

Consequently, $\mathcal{F}_{\text{eff}}(-w; \lambda) = \mathcal{F}_{\text{eff}}(w; \lambda)$, and its expansion contains only even powers. As in Landau theory, the order parameter emerges continuously from zero near a continuous transition, so a low-order Taylor expansion captures the critical behavior; higher-order terms become important mainly away from criticality or at large regularization. Here r is the actual quadratic curvature along w after the auxiliary parameter θ is allowed to readjust to its stationary value whenever w changes:

$$r(\lambda) = \left(\mathcal{F}_{ww} - \frac{\mathcal{F}_{w\theta}^2}{\mathcal{F}_{\theta\theta}} \right) \Bigg|_{(w,\theta)=(0,0)}. \tag{3}$$

The first term is the direct curvature along w at fixed θ; the second accounts for the reduction in curvature produced by the readjustment of θ. Meanwhile, $u(\lambda)$ collects fourth-order contributions after θ readjusts. Corrections of the forms w^2, $w\theta$, and θ^2 in the original objective can all change the effective mass of the same critical mode. We call local operators that change r and u effective R_2 and R_4, respectively. Here λ denotes the control strength of the corresponding method; when the two corrections are tuned independently, we write λ_2 and λ_4. The full reduction is given in Appendix B. We call bilinear controls based on known environmental sensitivity oracle R_2/R_4; their nonlinear counterparts are non-unique.

For closed-form, testable predictions, we specialize to the single-feature squared-loss model with explicit R_2/R_4 corrections. BCE has the same local form; a proof is given in Appendix B.3. The canonical objective is:

$$\mathcal{F}_{\text{can}}(w, \theta; \lambda_2, \lambda_4) = L_0 - aw\theta + \frac{m + \lambda_4 \gamma}{2} w^2\theta^2 + \frac{\mu + \lambda_2 \gamma}{2} w^2 + \frac{\mu}{2}\theta^2. \tag{4}$$

Here a is the task-predictive signal, m is the local task-loss curvature with respect to the final loading β, $\gamma \geq 0$ is the environmental sensitivity, and $\mu > 0$ is the weight-decay coefficient. In this canonical case, λ_2 and λ_4 independently control the effective mass and quartic stiffness. Other algorithms need not take the form of Eq. (4): joint quadratic and quartic corrections in (w, θ) respectively modify r and u after the same reduction. The critical-strength and loading formulas below are exact for this canonical model and asymptotic near the phase-transition critical point for other smooth objectives.

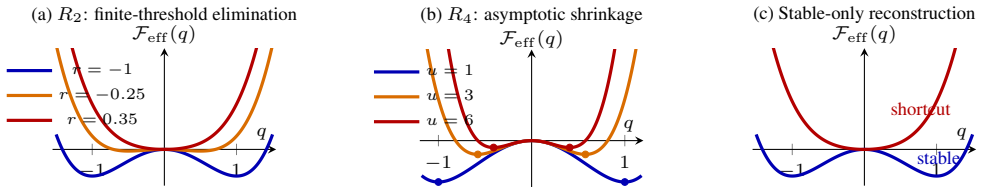


Figure 1: Schematic zero-field free-energy picture of finite-threshold R_2 elimination, asymptotic R_4 shrinkage, and selective retention of the stable mode.

4 A Landau Theory for a Single Feature

4.1 Variational Reduction

Let $A = \mu + \lambda_2\gamma$ and $B = m + \lambda_4\gamma$, and assume $A > 0$ and $B > 0$. For fixed w, Eq. (4) is strictly convex in the auxiliary parameter θ. The stationary condition $\partial_\theta \mathcal{F}_{\text{can}} = 0$ therefore uniquely determines the auxiliary equilibrium and the one-dimensional reduced objective:

$$\theta^\star(w) = \frac{aw}{\mu + Bw^2}, \qquad \mathcal{F}_{\text{eff}}(w) \equiv \mathcal{F}_{\text{can}}(w, \theta^\star(w)) = L_0 + \frac{A}{2}w^2 - \frac{a^2w^2}{2(\mu + Bw^2)}. \tag{5}$$

For this canonical model, strict convexity in θ makes the one-dimensional equilibrium reduction exact rather than merely local: stable equilibria of $\mathcal{F}_{\text{eff}}$ uniquely recover those of the original two-dimensional objective.

4.2 Landau Expansion and Equilibrium Phases

Expanding Eq. (5) around $w = 0$ gives

$$\begin{aligned} \mathcal{F}_{\text{eff}}(w) &= L_0 + \frac{\gamma}{2}(\lambda_2 - \lambda_{2,c})w^2 + \frac{u}{4}w^4 + \mathcal{O}(w^6), \\ \lambda_{2,c} &= \frac{a^2/\mu - \mu}{\gamma}, \qquad u = \frac{2a^2B}{\mu^2} > 0. \end{aligned} \tag{6}$$

The quadratic mass $r(\lambda_2) = \gamma(\lambda_2 - \lambda_{2,c})$ determines the local stability of the origin, while the positive quartic stiffness u stabilizes the nonzero ordered phase. We consider the ordinary case $u(\lambda_{2,c}) > 0$. Figure 1 illustrates the distinct effects of R_2 and R_4 on the resulting wells.

The Landau free energy of a spatially uniform mean-field ferromagnet can be written as (Landau, 1937; Landau & Lifshitz, 1980)

$$F_{\text{mag}}(M) = F_0 + \frac{a_0}{2}(T - T_c)M^2 + \frac{b}{4}M^4 + O(M^6). \tag{7}$$

Equations (6) and (7) share the same local normal form, with the correspondence

$$w \longleftrightarrow M, \qquad \lambda_2 - \lambda_{2,c} \longleftrightarrow T - T_c, \qquad \gamma \longleftrightarrow a_0, \qquad u \longleftrightarrow b. \tag{8}$$

This exact mathematical matching formalizes the strict isomorphism proposed in the introduction. In this canonical control family, λ_2 is a temperature-like control parameter and $|w|$ is the order-parameter amplitude of the learning system: the origin is locally stable when $r > 0$, a nonzero ordered solution appears when $r < 0$, and $r = 0$ gives the local continuous phase boundary.

In the continuous-time limit of gradient descent, the learning parameters obey

$$\frac{dW_i}{dt} = -\eta\frac{\partial \mathcal{F}}{\partial W_i} \qquad \Longrightarrow \qquad \tau_c\frac{dq}{dt} = -r(\lambda)q - u(\lambda_c)q^3 + \mathcal{O}\big(q^5 + |\lambda - \lambda_c|q^3\big), \qquad \tau_c > 0. \tag{9}$$

Near a simple phase boundary, the reduced equation has the standard time-dependent Landau (Model-A) form for the order parameter q (Hohenberg & Halperin, 1977). Its linear term changes sign at $r = 0$, so the early-time growth-rate zero crossing used in Fig. 2 identifies the static phase boundary. Appendix E gives the center-manifold derivation and its scope. Beyond equilibrium loading, the objective's Landau signature predicts how learned modes respond to weak loss perturbations

(Curie–Weiss law; Hohenberg & Krekhov, 2015) and how quickly they relax toward equilibrium during optimization. A conjugate field $-hq$ defines the susceptibility along an equilibrium branch:

$$\chi_q \equiv \left.\frac{\partial q^\star}{\partial h}\right|_{h=0} = \left[\partial_q^2 \mathcal{F}_{\text{eff}}(q^\star)\right]^{-1}, \qquad \tau_{\text{relax}} \sim \tau_c \chi_q. \tag{10}$$

Appendix E.3 tests their $|r|^{-1}$ scaling near criticality. Appendix E.4 shows close agreement between predicted and measured optimization steps near equilibrium for invariant-learning objectives.

4.3 Quantitative Phase Boundary and Loading

Theorem 4.1 (Task-signal–environment-sensitivity phase boundary)**.** *Consider the canonical squared-loss model in Eq. (4), and let $a = \mathbb{E}[XY]$. When $\gamma > 0$, the feature-onset condition and critical regularization strength can be written jointly as*

$$|a|^2 > \mu(\mu + \lambda_2\gamma) \quad \Longleftrightarrow \quad \lambda_2 < \lambda_{2,c}, \qquad \lambda_{2,c} = \frac{a^2/\mu - \mu}{\gamma}. \tag{11}$$

If $\gamma = 0$, quadratic environmental regularization does not move the local phase boundary of this feature, and the condition for a nonzero mode reduces to $|a| > \mu$.

Proof sketch. The exact reduced objective satisfies

$$\frac{d\mathcal{F}_{\text{eff}}}{dw} = w\left[A - \frac{a^2\mu}{(\mu + Bw^2)^2}\right]. \tag{12}$$

The bracketed term is strictly increasing in w^2, from $A - a^2/\mu$ to $A > 0$. Hence a stable nonzero stationary point exists if and only if $A - a^2/\mu < 0$, which is Eq. (11); equality gives the unique continuous phase boundary.

Corollary 4.2 (Selective retention of stable features)**.** *Let $\lambda_{2,c}^{(s)}$ and $\lambda_{2,c}^{(c)}$ denote the shortcut and stable-mode boundaries. If $\lambda_{2,c}^{(s)} < \lambda_{2,c}^{(c)}$, then for every $\lambda_2 \in (\lambda_{2,c}^{(s)}, \lambda_{2,c}^{(c)})$, the shortcut is stabilized at the origin while the stable direction remains in the nonzero acquired phase. Selective retention therefore depends on task signal relative to environmental sensitivity, not on marginal environment association alone.*

Proposition 4.3 (Steady-state feature loading)**.** *Assume $B = m + \lambda_4\gamma > 0$. The effective predictive loading at a stable stationary point of the canonical model is*

$$|\beta^\star| = \frac{\left[|a| - \sqrt{\mu(\mu + \lambda_2\gamma)}\right]_+}{m + \lambda_4\gamma}, \tag{13}$$

where $[z]_+ = \max\{z, 0\}$. Thus, R_2 makes the loading vanish exactly at the finite boundary $\lambda_{2,c}$, whereas R_4 continuously shrinks its nonzero post-onset loading.

Equation (13) captures the distinct roles of the two corrections: increasing λ_2 reduces the numerator and eventually changes the sign of the quadratic mass, while increasing λ_4 only enlarges the denominator. The final loading therefore approaches zero linearly in $\lambda_{2,c} - \lambda_2$ near criticality; the square-root scaling of the order parameter is given in Appendix D.

5 Multi-Component Order Parameters and Collective Modes

A multicomponent order parameter is a natural extension of Landau theory: distinct components can form collective modes through mode coupling. In our setting, environmental responses of different features can reinforce or cancel one another in combination. Let $\boldsymbol{w}, \boldsymbol{\theta} \in \mathbb{R}^p$ and $\boldsymbol{\beta} = \boldsymbol{w} \odot \boldsymbol{\theta}$. Here, Γ_E denotes the environmental sensitivity matrix. The quadratic environmental correction can be written as

$$R_2(\boldsymbol{w}) = \frac{\lambda_2}{2}\sum_i \Gamma_{E,ii} w_i^2 + \lambda_2 \sum_{i<j} \Gamma_{E,ij} w_i w_j. \tag{14}$$

For $\Gamma_E \succeq 0$, diagonal terms give componentwise masses, while off-diagonal terms provide spin-like pairwise coupling; task-supported low-eigenvalue directions form favored collective modes. The corresponding matrix R_4 correction has the same matrix structure but acts on the final loading $\boldsymbol{\beta}$. Below, $C = \text{diag}(\mathbb{E}[X_1Y], \ldots, \mathbb{E}[X_pY])$ denotes the task drive of each component.

Theorem 5.1 (Collective-mode phase boundary)**.** *For the multifeature bilinear extension of Eq. (4), assume $\Gamma_E \succeq 0$. The origin is locally unstable along a collective mode iff*

$$\lambda_{\max}\left[\frac{1}{\mu}C^\top(\mu I + \lambda_2\Gamma_E)^{-1}C\right] > 1. \tag{15}$$

Proposition 5.2 (Predictor-space selectivity of matrix R_4)**.** *Let $\Gamma_E \succeq 0$, $\lambda_2, \lambda_4 > 0$, and $\boldsymbol{\beta} = \boldsymbol{w}\odot\boldsymbol{\theta}$. Define*

$$R_2(\boldsymbol{w}) = \frac{\lambda_2}{2}\boldsymbol{w}^\top\Gamma_E\boldsymbol{w}, \qquad R_4(\boldsymbol{\beta}) = \frac{\lambda_4}{2}\boldsymbol{\beta}^\top\Gamma_E\boldsymbol{\beta}. \tag{16}$$

Then $R_4(\boldsymbol{\beta}) = 0$ iff $\boldsymbol{\beta} \in \ker\Gamma_E$, whereas $R_2(\boldsymbol{w}) = 0$ iff $\boldsymbol{w} \in \ker\Gamma_E$. The latter need not imply $\boldsymbol{\beta} \in \ker\Gamma_E$, since coordinatewise multiplication by $\boldsymbol{\theta}$ need not preserve this nullspace. Thus matrix R_4 directly penalizes environmental sensitivity in predictor space.

Environmental sensitivity specifies the penalized directions, while the penalty's local dependence on model parameters determines its signature order. The derivations and the Fig. 7 specialization are given in Appendix F. The oracle operators below use the b_F-shift setup for controlled spectral-mechanism checks, but the formulation and criterion apply beyond this particular shift. Crucially, oracle Γ_E is not uniquely identifiable from observational associations in finitely many training environments alone; uniquely recovering it requires additional structural assumptions or causal priors.

6 Regularization-Path Predictions

Table 1: Predicted regularization phenotypes.

Methods	**Objective signature**	**Predicted regularization phenotype**
MMD-mean, CORAL-full, IGA (Koyama & Yamaguchi, 2020), Gaussian MMD	Stable positive R_2 component	Moves the local onset and enables finite-strength mode elimination. Higher-order terms in IGA and Gaussian MMD may modify the path shape.
CORAL-cov, Fishr (Rame et al., 2022), V-REx	Stable positive R_4 component	Continuously shrinks an acquired mode. V-REx may exhibit a small nonlinear tail due to weak R_6/R_8 corrections.
IRM	R_4-leading with negative R_6 and positive R_8 terms	May exhibit non-monotonic paths and sharp drops at large strengths.
GroupDRO	Reweighted R_2/R_4 terms	Suppresses shortcuts at small strengths but may retain a nonzero plateau.
DANN (Ganin et al., 2016)	State-dependent R_2/R_4 terms	May exhibit unstable paths and large seed variability.
Fish (Shi et al., 2022), MLDG (Li et al., 2018)	Indefinite R_2 and R_4 gradient-alignment terms	Shows direction-dependent R_2-like suppression at small strength and becomes unstable at large strength.
ERM, Mixup (Zhang et al., 2018)	No fixed environmental correction	No sign-controlled shortcut suppression.

Table 1 summarizes the leading-order predictions in the low-loading regime. Their local expansions and empirical definitions are detailed in Appendices G and I.2, respectively. Actual selective retention also depends on whether the low-order matrices align with the oracle Γ_E: aligned matrix R_2 or R_4 operators can preserve low-sensitivity stable modes, whereas misaligned operators gradually suppress them as regularization grows. Away from onset, higher-order or sign-indefinite corrections can dominate and produce nonlinear tails, instability, or collapse.

7 Experiments

The preceding analysis yields quantitative predictions for phase boundaries, steady-state loadings, and critical strengths (Eqs. (11) and (13)), together with method-specific regularization phenotypes derived from the local objective signatures (Table 1). We therefore design the experiments to test

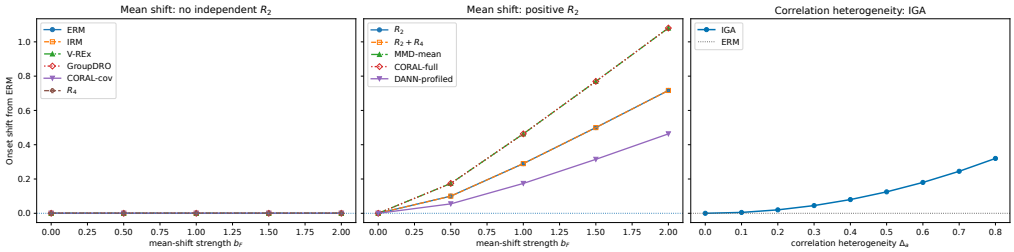
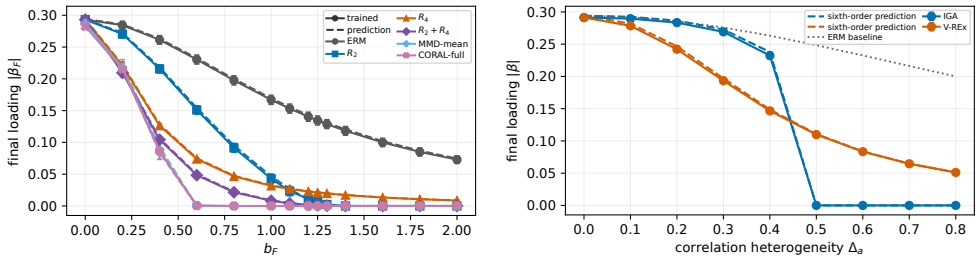


Figure 2: Local taxonomy of onset shifts. The vertical axis shows the shift relative to ERM in the critical signal strength a_s at which the model begins to learn the shortcut. The first two panels use the mean-shift probe b_F; the third uses IGA's matched correlation-heterogeneity probe Δ_a. In all cases, the analytic Landau onset agrees closely with the measurement, with maximum absolute deviation 4.7×10^{-6}. DANN-profiled is the locally profiled discriminator response (Appendix G.5).



(a) Mean-shift probe (b_F)

(b) Matched IGA–V-REx paths

Figure 3: Quantitative validation of shortcut-loading predictions. Dashed curves show reduced Landau predictions; solid curves and error bars show the mean and s.e.m. Panel (a) uses the mean-shift probe b_F; panel (b) uses the matched correlation-heterogeneity probe Δ_a.

these conclusions in sequence: Figs. 2 to 4 evaluate onset shifts, loading laws, and critical strengths; Figs. 5 and 6 examine whether the predicted path distinctions persist along full bilinear and nonlinear regularization paths; and Fig. 7 tests the spectral prediction that stable structure may reside in coupled collective modes (Eq. (15)).

7.1 Quadratic Mass Moves the Onset; Quartic Stiffness Shrinks the Amplitude

Figures 2–4 use scalar single-feature constructions. Figure 2 uses matched environmental probes to measure the onset shift relative to ERM. In this figure, onset is defined by the static condition $r = 0$ in Eq. (11); its numerical measurement is the zero crossing of the early-time local growth rate from a small perturbation, with their local equivalence detailed in Appendix E. Its first two panels scan the mean shift b_F: methods without an independent quadratic correction show almost no onset shift, whereas a positive gate-block mass produces a finite shift. The third panel instead scans the correlation heterogeneity Δ_a for IGA, whose joint quadratic stiffness is non-degenerate only in that construction. These results agree with Eq. (11): R_2 determines whether a mode can grow from the origin, while R_4 mainly controls the amplitude after growth.

Figure 3 quantitatively tests the reduced Landau loading predictions under two matched environmental probes: the mean shift b_F and correlation heterogeneity Δ_a. Without fitting to the trained paths, objective-specific low-order coefficients derived from each learning objective are inserted into the same reduced Landau equilibrium, yielding accurate loading predictions for multiple methods in panel (a); retaining the sextic coefficient extends the same construction to V-REx and IGA in panel (b) (Appendix D).

Figure 4 directly tests the predicted R_2 phase-transition point across explicit R_2, R_2+R_4, MMD-mean, CORAL-full, and matched IGA objectives. Near a simple continuous boundary, the equilibrium loading vanishes linearly with distance to the critical strength (Eq. (13)), motivating a local linear extrapolation from the ordered side. The R_2/R_2+R_4 comparison further confirms the distinct

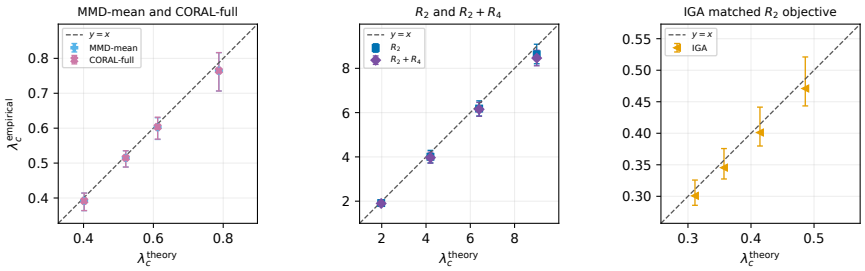
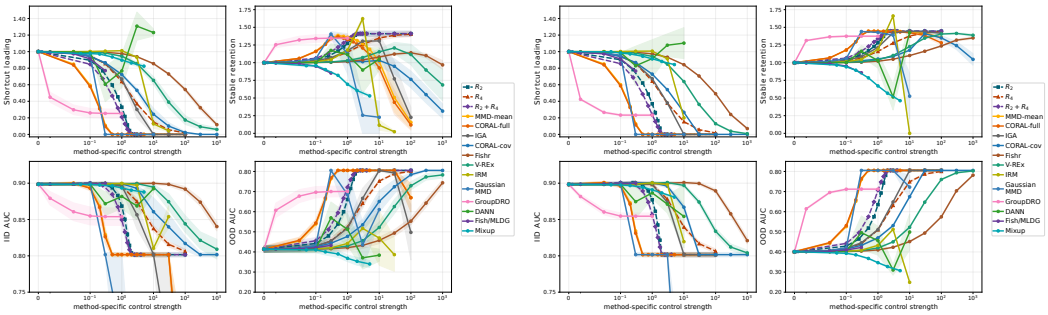


Figure 4: Critical-strength calibration from final shortcut loading. Each marker denotes one setting; vertical bars span the central 95% interval of the seed-resampled critical-strength estimates, and the dashed line is $y = x$. R_2+R_4 shares the R_2 onset, while MMD-mean, CORAL-full, and matched IGA track their corresponding effective-R_2 transition points.



(a) 256 training samples per environment

(b) 8192 training samples per environment

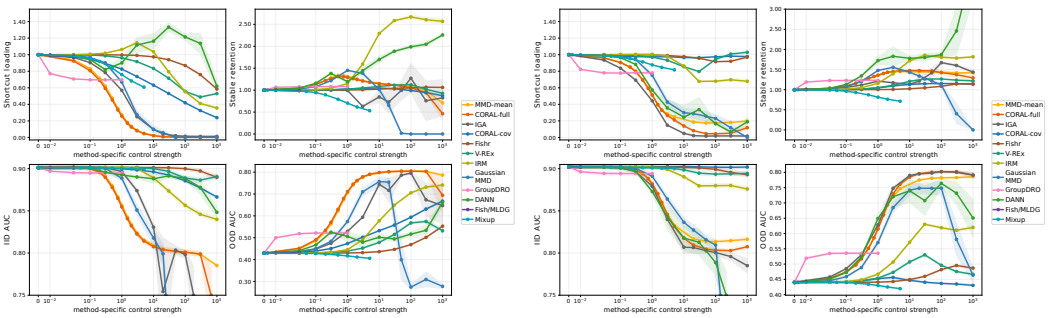
Figure 5: Sample-size dependence of bilinear regularization paths. Curves and shaded bands show means and standard errors over five seeds, with shared IID/OOD test sets across sample sizes. Dashed curves with filled markers denote explicit oracle R_2, R_4, and $R_2 + R_4$ controls. Shortcut loading and stable retention measure mean absolute input derivatives, normalized at $\lambda = 0$.

effects of R_2 and R_4: adding R_4 changes the amplitude and near-critical slope of the loading branch without shifting the R_2-controlled critical point. We estimate the empirical critical strength from a theory-blind coarse sweep followed by a dense rescan; the theoretical value is not used to select the fitting window or perform the fit. Only well-resolved local fits passing pre-specified quality controls are retained; detailed calibration settings and criteria are provided in Appendix I.3.3.

7.2 Regularization Paths

Figures 5–6 use separate stable and shortcut features together with independent noise coordinates. In Fig. 5, stable features suffer earlier collateral damage with 256 samples per environment (left) than with 8192 (right), consistent with larger finite-sample fluctuations in the environmental statistics (Appendix H.1). Oracle R_2 and R_2+R_4 controls preserve the stable mode after shortcut suppression throughout the tested range. OOD AUC improves only when shortcut suppression preserves the stable mode. The observed regularization phenotypes agree with the qualitative predictions derived from the objective signatures in Table 1. MMD-mean and CORAL-full show R_2-like finite-strength suppression, whereas CORAL-cov, Fishr, and V-REx show R_4-leading gradual shrinkage. IGA and Gaussian MMD show modified path shapes, IRM non-monotonic paths, and V-REx a mild nonlinear tail, consistent with their higher-order structure. GroupDRO retains a nonzero shortcut plateau, while DANN shows unstable, seed-sensitive paths.

We transfer the same datasets and objectives to ReLU networks. Figure 6 shows that low-order signatures remain qualitatively predictive: R_2-like objectives retain finite-strength shortcut control, whereas R_4-like objectives primarily yield continuous suppression with residual finite-strength shortcut loading. Across the tested depths and sample sizes, these qualitative signature–phenotype relations persist, while transition scales, large-strength behavior, and higher-order effects vary. Deeper networks and extended paths are reported in Appendix H.2. To study regularization



(a) One-hidden-layer ReLU (b) Two-hidden-layer ReLU

Figure 6: Nonlinear regularization paths using the same data as Fig. 5(b), with 8192 training samples per environment. The bilinear model's qualitative regularization phenotypes persist in both ReLU networks, with different transition scales. Objectives with higher-order terms exhibit network-dependent nonlinear tails.

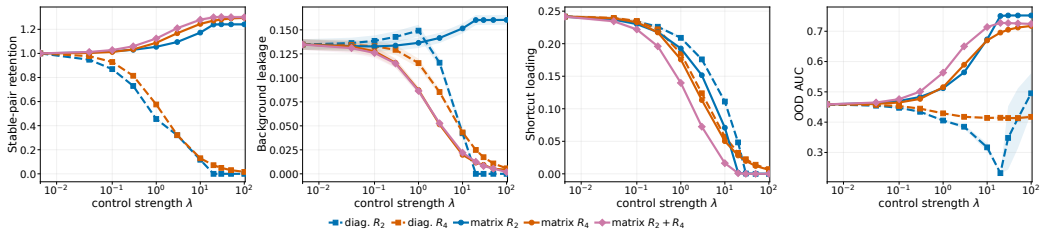


Figure 7: Regularization paths for a compensatory collective mode. The panels report stable-pair retention, task-irrelevant background leakage, shortcut loading, and OOD AUC. Dashed curves use $\mathrm{diag}(\Gamma_E)$, while solid curves use the full oracle Γ_E. The matrix R_4 and matrix R_2+R_4 curves nearly overlap in the background-leakage panel.

paths with more complex, image-based stable features, we extend the experiments to MNIST (Appendix H.3).

7.3 Collective Feature Modes

Figure 7 uses a three-feature construction in which two individually environment-sensitive features form a compensatory collective mode, alongside an independent shortcut. Figure 7 compares coordinate-diagonal R_2, R_4 with coupled matrix $R_2, R_4, R_2 + R_4$. Diagonal regularization destroys the stable combination while suppressing the shortcut. In contrast, coupled matrix R_2, R_4, and their combination agree with the spectral theory: they suppress the shortcut while preserving the collective mode and improving OOD AUC. Matrix R_4 is particularly effective at suppressing background leakage from imperfect pair cancellation. The supplementary full baseline comparison (Fig. 18) indicates that, in this construction, the non-oracle baselines do not consistently exhibit the same joint preservation–suppression behavior across all four diagnostics.

8 Conclusion

We develop a Landau-theoretic framework for predicting how invariant-learning objectives control the onset, suppression, and retention of predictive modes from their low-order structure. The resulting objective signatures organize existing methods by distinct regularization phenotypes and extend naturally from individual features to collective modes. Future work can build on the framework's equilibrium and dynamical descriptions to explain architecture-dependent algorithmic behavior and predict relevant regularization-strength ranges before empirical search. These insights can further guide the design of objectives with desired phenotypes, separating the estimation of environmental sensitivity Γ_E from the construction of penalties with prescribed Landau signatures.

Reproducibility statement

Complete proofs are provided in the appendices. Experimental configurations, objective definitions, random seeds, and critical-strength estimation protocols are documented in Appendix I. Code is included in the supplementary material.

AI use statement

Generative AI tools were used to aid and polish the writing of this paper, including improvements to clarity and readability. All AI-assisted content was reviewed by the authors, who take full responsibility for the final manuscript.

References

El Mehdi Achour, François Malgouyres, and Sébastien Gerchinovitz. The loss landscape of deep linear neural networks: A second-order analysis. *Journal of Machine Learning Research*, 25 (242):1–76, 2024.

Kartik Ahuja, Karthikeyan Shanmugam, Kush R. Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 145–155. PMLR, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

M. Biehl, E. Schlösser, and M. Ahr. Phase transitions in soft-committee machines. *Europhysics Letters*, 44(2):261–267, 1998. doi: 10.1209/epl/i1998-00466-6.

Yimeng Chen, Ruibin Xiong, Zhi-Ming Ma, and Yanyan Lan. When does group invariant learning survive spurious correlations? In *Advances in Neural Information Processing Systems*, volume 35, pp. 7038–7051, 2022.

Yongqiang Chen, Kaiwen Zhou, Yatao Bian, Binghui Xie, Bingzhe Wu, Yonggang Zhang, Han Yang, Kaili Ma, Peilin Zhao, Bo Han, and James Cheng. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. In *International Conference on Learning Representations*, 2023.

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.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019. doi: 10.1103/PhysRevE.100.012115.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Vitaly L. Ginzburg and Lev D. Landau. On the theory of superconductivity. *Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki*, 20:1064–1082, 1950.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Pierre C. Hohenberg and Bertrand I. Halperin. Theory of dynamic critical phenomena. *Reviews of Modern Physics*, 49(3):435–479, 1977. doi: 10.1103/RevModPhys.49.435.

Pierre C. Hohenberg and Alexei P. Krekhov. An introduction to the Ginzburg–Landau theory of phase transitions and nonequilibrium patterns. *Physics Reports*, 572:1–42, 2015. doi: 10.1016/j.physrep.2015.01.001.

Pritish Kamath, Akilesh Tangella, Danica J. Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 4069–4077. PMLR, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021.

Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883*, 2020.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi 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 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021.

Lev D. Landau. On the theory of phase transitions. *Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki*, 7:19–32, 1937.

Lev D. Landau and Evgeny M. Lifshitz. *Statistical Physics, Part 1*, volume 5 of *Course of Theoretical Physics*. Pergamon Press, Oxford, 3rd edition, 1980.

Thomas Laurent and James von Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2902–2907. PMLR, 2018.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11596.

Yotam Norman and Ron Meir. Unsupervised representation learning – an invariant risk minimization perspective. In *International Conference on Learning Representations*, 2026.

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.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Alexandre Rame, 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 of *Proceedings of Machine Learning Research*, pp. 18347–18377. PMLR, 2022.

Elan Rosenfeld, Pradeep K. Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

David Saad and Sara A. Solla. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337–4340, 1995. doi: 10.1103/PhysRevLett.74.4337.

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.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992. doi: 10.1103/PhysRevA.45.6056.

Yuge Shi, Jeffrey Seely, Philip H. S. Torr, N. Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022.

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.

T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499–556, 1993. doi: 10.1103/RevModPhys.65.499.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Yihua Zhang, Pranay Sharma, Parikshit Ram, Mingyi Hong, Kush R. Varshney, and Sijia Liu. What is missing in IRM training and evaluation? Challenges and solutions. In *International Conference on Learning Representations*, 2023.

Appendix

Appendix Contents

A Scope and notation **14**

B Variational reduction **14**

B.1 Joint quadratic corrections and the reduced effective mass 16

B.2 Exact reduction of the canonical model 17

B.3 Local normal-form equivalence of BCE and squared loss 18

C Taylor expansion and Landau coefficients **18**

D Equilibrium branches and loading **19**

E Onset diagnostics, susceptibility, and Landau dynamics **21**

E.1 Linearized onset diagnostic 21

E.2 Nonlinear soft-mode dynamics under gradient flow 21

E.3 Susceptibility and critical slowing down 22

E.4 Optimization steps near equilibrium for invariant-learning objectives 24

F Collective modes and environment-induced mass **25**

G Derivation of objective signatures **27**

G.1 ERM and explicit R_2/R_4 regularizers 27

G.2 MMD-mean, Gaussian MMD, and CORAL 28

G.3 IRMv1, V-REx, and GroupDRO 28

G.4 IGA, Fish/MLDG, and Fishr 29

G.5 Mixup and a DANN surrogate 32

G.6 Summary of local operators 33

H Supplementary results **33**

H.1 Finite-sample operator alignment 33

H.2 Extended nonlinear regularization paths 34

H.3 Regularization paths on MNIST 36

H.4 Collective-mode baseline comparison 39

I Experimental configurations and reproducibility **39**

I.1 Training protocol 39

I.2 Empirical objectives and implementation choices 40

I.3 Figure-specific configurations 41

A SCOPE AND NOTATION

This appendix provides the complete derivations underlying the equilibrium Landau theory in the main text. Two levels of claims should be distinguished. For the canonical bilinear model with squared loss and explicit R_2/R_4 regularization, the reduced objective, critical strengths, and loading formula are exact. For general smooth losses and other algorithms, we compare only the leading local operators that emerge near a common reference point. The sign change of the quadratic mass at the origin, and hence the continuous local transition point, is determined exactly by the local curvature; nonzero equilibrium loading is predicted only asymptotically near criticality.

In environment e with data distribution P_e, define

$$c_e = \mathbb{E}_e[XY], \qquad q_e = \mathbb{E}_e[X^2], \qquad \beta = w\theta. \tag{17}$$

The environmental squared risk can be written exactly as

$$\ell_e(w,\theta) = L_{0,e} - c_e\beta + \frac{q_e}{2}\beta^2. \tag{18}$$

Hereafter, an overbar denotes an average over environments, e.g., $\bar{c} = \overline{c_e}$ and $\bar{q} = \overline{q_e}$, and we define $\Delta c_e = c_e - \bar{c}$ and $\Delta q_e = q_e - \bar{q}$. The reference objective with weight decay on both parameter blocks is

$$\mathcal{F}_0(w,\theta) = \bar{L}_0 - \bar{c}\, w\theta + \frac{\bar{q}}{2}w^2\theta^2 + \frac{\mu_g}{2}w^2 + \frac{\mu_\theta}{2}\theta^2. \tag{19}$$

We allow separate weight-decay coefficients μ_g and μ_θ; the main text sets $\mu_g = \mu_\theta = \mu$.

For any unit direction $v = (v_w, v_\theta)$, we define the coefficients of the objective along the ray $z = \rho v$ by

$$\mathcal{F}(\rho v) - \mathcal{F}(0) = \frac{1}{2}a_2(v)\rho^2 + \frac{1}{4}a_4(v)\rho^4 + O(\rho^6). \tag{20}$$

Equation (19) gives the reference fingerprint

$$\begin{aligned} a_2^{(0)}(v) &= \mu_g v_w^2 + \mu_\theta v_\theta^2 - 2\bar{c}\, v_w v_\theta, \\ a_4^{(0)}(v) &= 2\bar{q}\,(v_w v_\theta)^2. \end{aligned} \tag{21}$$

Accordingly, a "joint R_2 correction" in this paper means an independent correction that changes a_2 or the Hessian at the origin, whereas a "joint R_4 correction" begins at order ρ^4 in the joint variables (w,θ). To determine mode onset, these joint corrections must further be projected onto the critical soft mode to obtain the change in the effective mass of the reduced objective. The sign and magnitude of this change depend on the projection of the joint quadratic operator onto that mode.

B VARIATIONAL REDUCTION

This section states the precise conditions under which noncritical parameter directions can be eliminated to obtain a low-dimensional effective objective. We first give the two-coordinate reduction used by the canonical model and then extend it to the critical soft mode of a finite nonlinear network. The reduction characterizes equilibrium stationary points.

Theorem B.1 (Local elimination of the auxiliary degree of freedom)**.** *Suppose that $\mathcal{F}(w,\theta;\lambda)$ is C^6 in a neighborhood of the critical reference point $(w,\theta,\lambda) = (0,0,\lambda_c)$ and satisfies*

$$\partial_\theta \mathcal{F}(0,0;\lambda_c) = 0, \qquad \partial_{\theta\theta}\mathcal{F}(0,0;\lambda_c) > 0. \tag{22}$$

Then there exists a neighborhood of this point and a unique C^5 function $\theta^\star(w,\lambda)$ such that

$$\partial_\theta \mathcal{F}(w,\theta^\star(w,\lambda);\lambda) = 0. \tag{23}$$

Define

$$\mathcal{F}_{\text{eff}}(w;\lambda) = \mathcal{F}(w,\theta^\star(w,\lambda);\lambda)\,. \tag{24}$$

Then

$$\frac{d\mathcal{F}_{\text{eff}}}{dw} = \partial_w \mathcal{F}(w,\theta^\star(w,\lambda);\lambda)\,. \tag{25}$$

Consequently, stationary points of $\mathcal{F}_{\text{eff}}$ are in one-to-one correspondence with local stationary points of the original objective satisfying $\partial_\theta \mathcal{F} = 0$.

Proof. By Eq. (22), the function $G(w, \theta, \lambda) = \partial_\theta \mathcal{F}(w, \theta; \lambda)$ satisfies, at the reference point, $G = 0$ and $\partial_\theta G = \partial_{\theta\theta}\mathcal{F} \neq 0$. The implicit function theorem therefore immediately yields a unique $\theta^\star(w, \lambda)$. Applying the chain rule to Eq. (24) gives

$$\frac{d\mathcal{F}_{\text{eff}}}{dw} = \partial_w \mathcal{F} + \partial_\theta \mathcal{F} \frac{\partial \theta^\star}{\partial w} = \partial_w \mathcal{F}, \tag{26}$$

where the second term vanishes by Eq. (23). □

Theorem B.2 (Soft-mode reduction and generic phase boundary for a finite nonlinear network)**.** *Let $\mathcal{F}(\Theta; \lambda)$ be C^7 on a finite-dimensional parameter space, and let $\bar{\Theta}(\lambda)$ be a C^6 branch of stationary points near λ_c:*

$$\nabla_\Theta \mathcal{F}\big(\bar{\Theta}(\lambda); \lambda\big) = 0. \tag{27}$$

Suppose that, after restricting to a local slice transverse to any exact parameter symmetries, the critical Hessian

$$H_c = \nabla^2_\Theta \mathcal{F}\big(\bar{\Theta}(\lambda_c); \lambda_c\big) \tag{28}$$

has a one-dimensional kernel $\ker H_c = \operatorname{span}\{v_c\}$, where $\|v_c\| = 1$, and is positive definite on $v_c^\perp$. Then there is a neighborhood of $(q, \lambda) = (0, \lambda_c)$ and a unique C^6 map $\eta^\star(q, \lambda) \in v_c^\perp$, with $\eta^\star(0, \lambda) = 0$, such that

$$P_\perp \nabla_\Theta \mathcal{F}\big(\bar{\Theta}(\lambda) + q v_c + \eta^\star(q, \lambda); \lambda\big) = 0, \tag{29}$$

where $P_\perp$ denotes orthogonal projection onto $v_c^\perp$. Hence all nearby stationary points lie on the scalar soft-mode manifold

$$\Theta^\star(q, \lambda) = \bar{\Theta}(\lambda) + q v_c + \eta^\star(q, \lambda). \tag{30}$$

Defining

$$\mathcal{F}_{\text{eff}}(q; \lambda) = \mathcal{F}(\Theta^\star(q, \lambda); \lambda)\,, \tag{31}$$

we have

$$\frac{d\mathcal{F}_{\text{eff}}}{dq} = \langle v_c, \nabla_\Theta \mathcal{F}(\Theta^\star(q, \lambda); \lambda)\rangle\,. \tag{32}$$

Consequently, stationary points of the full network objective near the critical branch are in one-to-one correspondence with stationary points of $\mathcal{F}_{\text{eff}}$.

If, in addition, there is a local orthogonal involution S satisfying $S^2 = I$, $Sv_c = -v_c$, $S\bar{\Theta}(\lambda) = \bar{\Theta}(\lambda)$, and

$$\mathcal{F}\big(\bar{\Theta}(\lambda) + Sz; \lambda\big) = \mathcal{F}\big(\bar{\Theta}(\lambda) + z; \lambda\big)\,, \tag{33}$$

then $\mathcal{F}_{\text{eff}}(-q; \lambda) = \mathcal{F}_{\text{eff}}(q; \lambda)$ and

$$\mathcal{F}_{\text{eff}}(q; \lambda) = \mathcal{F}_0(\lambda) + \frac{1}{2} r(\lambda) q^2 + \frac{1}{4} u(\lambda) q^4 + \mathcal{O}(q^6), \qquad r(\lambda_c) = 0. \tag{34}$$

If, furthermore,

$$r'(\lambda_c) \neq 0, \qquad u(\lambda_c) > 0, \tag{35}$$

then λ_c is a generic codimension-one continuous phase boundary. The stationary branch $q = 0$ is locally stable on the side where $r(\lambda) > 0$; on the side where $r(\lambda) < 0$, two symmetry-related stable branches emerge:

$$q_\pm(\lambda) = \pm \sqrt{-\frac{r(\lambda)}{u(\lambda_c)}} + \mathcal{O}\Big(|\lambda - \lambda_c|^{3/2}\Big)\,. \tag{36}$$

Proof. Write a nearby parameter vector as $\Theta = \bar{\Theta}(\lambda) + q v_c + \eta$ with $\eta \in v_c^\perp$, and define

$$G(q, \eta, \lambda) = P_\perp \nabla_\Theta \mathcal{F}\big(\bar{\Theta}(\lambda) + q v_c + \eta; \lambda\big)\,. \tag{37}$$

At $(0, 0, \lambda_c)$, the derivative of G with respect to η is the restricted Hessian $H_\perp = P_\perp H_c|_{v_c^\perp}$, which is positive definite and hence invertible by assumption. The implicit function theorem therefore yields the unique map $\eta^\star(q, \lambda)$ satisfying Eq. (29). Because Eq. (27) implies $G(0, 0, \lambda) = 0$, uniqueness gives $\eta^\star(0, \lambda) = 0$.

Along the reduced manifold, the transverse component of the gradient vanishes, so the gradient is parallel to v_c. Since $\partial_q \eta^\star \in v_c^\perp$, the chain rule gives

$$\frac{d\mathcal{F}_{\text{eff}}}{dq} = \langle \nabla_\Theta \mathcal{F}, v_c + \partial_q \eta^\star \rangle = \langle \nabla_\Theta \mathcal{F}, v_c \rangle \,, \tag{38}$$

which proves Eq. (32) and the stationary-point correspondence. At criticality, differentiating the transverse stationarity equation with respect to q gives $H_\perp \partial_q \eta^\star(0, \lambda_c) = -P_\perp H_c v_c = 0$, hence $\partial_q \eta^\star(0, \lambda_c) = 0$. Therefore the reduced quadratic curvature vanishes: $\partial_{qq} \mathcal{F}_{\text{eff}}(0; \lambda_c) = v_c^\top H_c v_c = 0$.

Under Eq. (33), orthogonality of S implies that it preserves $v_c^\perp$. Both $S\eta^\star(q, \lambda)$ and $\eta^\star(-q, \lambda)$ solve the same transverse stationarity equation, so uniqueness gives $\eta^\star(-q, \lambda) = S\eta^\star(q, \lambda)$. Invariance of $\mathcal{F}$ then yields $\mathcal{F}_{\text{eff}}(-q; \lambda) = \mathcal{F}_{\text{eff}}(q; \lambda)$. The even Taylor expansion and $r(\lambda_c) = 0$ give Eq. (34).

Under Eq. (35), the reduced stationarity equation factors as

$$\partial_q \mathcal{F}_{\text{eff}}(q; \lambda) = q \left[r(\lambda) + u(\lambda_c) q^2 + \mathcal{O}\big(q^4 + |\lambda - \lambda_c| q^2\big) \right] . \tag{39}$$

The transverse condition $r'(\lambda_c) \neq 0$ makes r a valid local control coordinate, so the implicit function theorem applied to q^2 yields the two nonzero branches in Eq. (36). Positive definiteness on $v_c^\perp$ keeps every transverse direction stable. Along the reduced direction, $\partial_{qq} \mathcal{F}_{\text{eff}}(0; \lambda) = r(\lambda)$, whereas on the nonzero branches $\partial_{qq} \mathcal{F}_{\text{eff}}(q_\pm; \lambda) = -2r(\lambda) + o(|r(\lambda)|) > 0$ when $r(\lambda) < 0$. This proves the stated local stability and the generic codimension-one continuous phase boundary. □

Here q is a local Landau order-parameter coordinate. When the critical kernel is one-dimensional, the soft direction v_c is unique up to sign, while the normalization of q is a coordinate choice; the two branches $q_\pm$ are symmetry-related states along this same direction. In the canonical bilinear model, the nonzero w projection lets w serve directly as this coordinate.

To connect the full-network result with the scalar formula in the main text, choose a local coordinate s having nonzero projection onto v_c and collect all remaining parameters in ψ. If the Hessian is written as

$$H = \begin{pmatrix} A & B^\top \\ B & C \end{pmatrix}, \qquad C \text{ invertible}, \tag{40}$$

then eliminating the noncritical block gives the reduced mass

$$r = A - B^\top C^{-1} B. \tag{41}$$

At $r = 0$, the critical soft direction is proportional to $(1, -C^{-1}B)^\top$. Thus Eq. (3) is the two-parameter instance of the same Schur-complement reduction, while the vector $-C^{-1}B$ records the coordinated first-order response of all remaining layers.

Remark B.3 (Smooth, piecewise-smooth, and multicomponent cases). The theorem applies directly to finite networks with smooth activations. For a ReLU network, the same argument applies inside an activation region whose interior contains the reference point. Transitions across activation-region boundaries require a nonsmooth analysis. The ReLU experiments assess phenotype-level transfer along finite regularization paths. If $\dim \ker H_c = k > 1$, the same projection yields a vector order parameter $q \in \mathbb{R}^k$ and a multicomponent Landau reduction; only the critical subspace is intrinsic, while the basis and the components of q may be rotated within it.

B.1 Joint quadratic corrections and the reduced effective mass

To show how quadratic corrections on different parameter blocks jointly determine the same onset, consider the following local expansion with simultaneous sign-reversal symmetry:

$$\mathcal{F}(w, \theta; \lambda) = L_0 + \frac{1}{2} A_\lambda w^2 + K_\lambda w\theta + \frac{1}{2} C_\lambda \theta^2 + \mathcal{F}_{4,\lambda}(w, \theta) + O\big(\|(w, \theta)\|^6\big) \,, \tag{42}$$

where $\mathcal{F}_{4,\lambda}$ is a homogeneous quartic polynomial. The quadratic Hessian is

$$H_2(\lambda) = \begin{pmatrix} A_\lambda & K_\lambda \\ K_\lambda & C_\lambda \end{pmatrix} . \tag{43}$$

If $C_\lambda > 0$, the θ direction remains noncritical. Solving $\partial_\theta \mathcal{F} = 0$ and applying the implicit function theorem gives

$$\theta^\star(w;\lambda) = \kappa_\lambda w + O(w^3), \qquad \kappa_\lambda = -\frac{K_\lambda}{C_\lambda}. \tag{44}$$

Whenever the zero eigenvector of the critical Hessian has a nonzero w component, w is a valid local coordinate for the one-dimensional soft mode. If this component vanishes, the roles of the two coordinates can simply be interchanged.

Completing the square in the quadratic part yields

$$\mathcal{F}_2(w,\theta;\lambda) = \frac{1}{2} r_\lambda w^2 + \frac{C_\lambda}{2}\left(\theta - \kappa_\lambda w\right)^2, \qquad r_\lambda = A_\lambda - \frac{K_\lambda^2}{C_\lambda}. \tag{45}$$

Thus r_λ is precisely the Schur complement of the θ block, and, for $C_\lambda > 0$, the origin is locally stable if and only if $r_\lambda > 0$. Substituting Eq. (44) into the quartic term gives

$$\mathcal{F}_{\text{eff}}(w;\lambda) = L_0 + \frac{1}{2} r_\lambda w^2 + \frac{1}{4} u_\lambda w^4 + O(w^6), \qquad u_\lambda = 4\mathcal{F}_{4,\lambda}(1,\kappa_\lambda). \tag{46}$$

Note that $\theta^\star - \kappa_\lambda w = O(w^3)$ contributes only $O(w^6)$ through the completed-square quadratic term and therefore does not alter the quartic coefficient.

Now suppose that the algorithm adds a general quadratic correction at the reference point,

$$\Delta\mathcal{F}_2 = \frac{\lambda}{2}\left(q_{ww} w^2 + 2q_{w\theta} w\theta + q_{\theta\theta}\theta^2\right), \qquad Q = \begin{pmatrix} q_{ww} & q_{w\theta} \\ q_{w\theta} & q_{\theta\theta} \end{pmatrix}. \tag{47}$$

Letting $\kappa_0 = -K_0/C_0$ and differentiating Eq. (45) directly gives

$$\left.\frac{dr_\lambda}{d\lambda}\right|_{\lambda=0} = q_{ww} + 2\kappa_0 q_{w\theta} + \kappa_0^2 q_{\theta\theta} = \begin{pmatrix} 1 & \kappa_0 \end{pmatrix} Q \begin{pmatrix} 1 \\ \kappa_0 \end{pmatrix}. \tag{48}$$

The distinct operators w^2, $w\theta$, and θ^2 all contribute to the effective mass along the critical soft mode. The sign after projection determines whether the quadratic correction suppresses or amplifies that mode. Equation (48) gives the first-order shift at $\lambda = 0$; the finite-λ mass is given by Eq. (45).

B.2 Exact reduction of the canonical model

For the truncated bilinear objective in the main text, define

$$A = \mu + \lambda_2\gamma, \qquad B = m + \lambda_4\gamma, \tag{49}$$

and write

$$\mathcal{F}(w,\theta) = L_0 - aw\theta + \frac{B}{2} w^2\theta^2 + \frac{A}{2} w^2 + \frac{\mu}{2}\theta^2. \tag{50}$$

If $B > 0$, then

$$\partial_{\theta\theta}\mathcal{F} = \mu + Bw^2 > 0, \tag{51}$$

Hence the elimination above is unique for every finite w, with

$$\theta^\star(w) = \frac{aw}{\mu + Bw^2}. \tag{52}$$

Moreover, the canonical objective satisfies the exact completed-square identity

$$\mathcal{F}(w,\theta) = \mathcal{F}_{\text{eff}}(w) + \frac{\mu + Bw^2}{2}\left(\theta - \theta^\star(w)\right)^2. \tag{53}$$

Because the second term is nonnegative, local minima of $\mathcal{F}_{\text{eff}}$ are in one-to-one correspondence with local minima of the original two-dimensional objective.

B.3 Local normal-form equivalence of BCE and squared loss

The exact closed-form results in the main text use squared loss, but the same critical Landau structure does not depend on this choice. Let $Y \in \{0, 1\}$, let the single-feature logit be $z = \beta X = w\theta X$, and consider the binary cross-entropy

$$\ell_{\mathrm{BCE}}(z, Y) = \log(1 + e^z) - Yz. \tag{54}$$

Near zero logit, the softplus expansion is

$$\log(1 + e^z) = \log 2 + \frac{z}{2} + \frac{z^2}{8} - \frac{z^4}{192} + O(z^6). \tag{55}$$

Assume that the required feature moments are finite, so that expectations may be taken term by term. Define

$$a_e^{\mathrm{BCE}} = \mathbb{E}_e\left[\left(Y - \frac{1}{2}\right) X\right], \qquad q_e = \mathbb{E}_e[X^2], \qquad h_e = \mathbb{E}_e[X^4], \tag{56}$$

Then the BCE risk in environment e satisfies

$$L_e^{\mathrm{BCE}}(\beta) = \log 2 - a_e^{\mathrm{BCE}}\beta + \frac{q_e}{8}\beta^2 - \frac{h_e}{192}\beta^4 + O(\beta^6). \tag{57}$$

After substituting $\beta = w\theta$, the joint-parameter expansion through fourth order is

$$L_e^{\mathrm{BCE}}(w, \theta) = L_{0,e} - a_e^{\mathrm{BCE}} w\theta + \frac{m_e^{\mathrm{BCE}}}{2} w^2\theta^2 + O\big(\|(w, \theta)\|^8\big), \qquad m_e^{\mathrm{BCE}} = \frac{q_e}{4}. \tag{58}$$

This has the same quadratic $w\theta$ coupling and quartic $w^2\theta^2$ stabilizing term as the canonical squared-loss objective, after replacing the task signal and curvature by a_e^{BCE} and m_e^{BCE}. If the feature is centered, $\mathbb{E}_e[X] = 0$, then $a_e^{\mathrm{BCE}} = \mathbb{E}_e[YX]$, identical to the task signal under squared loss. For the encoding $Y \in \{-1, +1\}$, using $\log(1 + e^{-Yz})$ similarly gives $a_e^{\mathrm{BCE}} = \frac{1}{2}\mathbb{E}_e[YX]$ and $m_e^{\mathrm{BCE}} = q_e/4$; constant rescalings of the labels or logits merely reparameterize these coefficients.

Therefore, BCE and squared loss have the same Landau normal form near the zero-logit critical point and induce the same effective R_2/R_4 order classification and phase-boundary mechanism. The equivalence holds near zero logit: the higher-order terms in Eq. (57) generally do not vanish, so the exact reduced objective and closed-form loading formula derived for squared loss are only asymptotic predictions near criticality for BCE. If an OOD algorithm constructs its penalty directly from risks or gradients, its numerical coefficients must still be re-expanded under the BCE objective.

C Taylor expansion and Landau coefficients

Because

$$\mathcal{F}(-w, -\theta) = \mathcal{F}(w, \theta), \qquad \theta^\star(-w) = -\theta^\star(w), \tag{59}$$

the reduced objective satisfies $\mathcal{F}_{\mathrm{eff}}(-w) = \mathcal{F}_{\mathrm{eff}}(w)$. This symmetry excludes odd powers from the reduced Taylor expansion. The two branches $\pm w$ correspond to the same $\beta = w\theta$, so the physically relevant order-parameter magnitude is $|w|$.

Substituting Eq. (52) into Eq. (50) gives the exact reduced objective

$$\mathcal{F}_{\mathrm{eff}}(w) = L_0 + \frac{A}{2}w^2 - \frac{a^2 w^2}{2(\mu + Bw^2)}. \tag{60}$$

Using

$$\frac{1}{\mu + Bw^2} = \frac{1}{\mu} - \frac{B}{\mu^2}w^2 + \frac{B^2}{\mu^3}w^4 + O(w^6), \tag{61}$$

we obtain

$$\mathcal{F}_{\mathrm{eff}}(w) = L_0 + \frac{1}{2}rw^2 + \frac{1}{4}uw^4 + O(w^6), \tag{62}$$

where

$$r = A - \frac{a^2}{\mu} = \gamma(\lambda_2 - \lambda_{2,c}), \qquad \lambda_{2,c} = \frac{a^2/\mu - \mu}{\gamma}, \qquad u = \frac{2a^2 B}{\mu^2}. \tag{63}$$

Provided that

$$B = m + \lambda_4 \gamma > 0, \tag{64}$$

we have $u > 0$, so the quartic term stabilizes the reduced free energy near the critical point. Higher-order task terms, if present, modify only coefficients of order $O(w^6)$ and above and do not change the leading division of roles between the quadratic mass and quartic stiffness.

When the two layers have different baseline curvatures,

$$\mathcal{F} = L_0 - aw\theta + \frac{B}{2} w^2\theta^2 + \frac{A_w}{2} w^2 + \frac{\mu_\theta}{2}\theta^2, \tag{65}$$

the same calculation gives

$$r = A_w - \frac{a^2}{\mu_\theta}, \qquad u = \frac{2a^2 B}{\mu_\theta^2}. \tag{66}$$

Thus the shared curvature μ in the main text is only a notational simplification.

D EQUILIBRIUM BRANCHES AND LOADING

Theorem D.1 (Exact phase boundary of the canonical model)**.** *Let $A > 0$ and $B > 0$. The derivative of the canonical reduced objective is*

$$\mathcal{F}'_{\text{eff}}(w) = w\left[A - \frac{a^2\mu}{(\mu + Bw^2)^2}\right]. \tag{67}$$

The origin is the unique stable equilibrium when $a^2 \le A\mu$; when $a^2 > A\mu$, a pair of stable nonzero equilibria exists. Therefore, the exact phase boundary is

$$a^2 = A\mu \quad \Longleftrightarrow \quad \lambda_2 = \frac{a^2/\mu - \mu}{\gamma}. \tag{68}$$

Proof. Let

$$g(s) = A - \frac{a^2\mu}{(\mu + Bs)^2}, \qquad s = w^2 \ge 0.$$

Because $B > 0$, $g(s)$ is strictly increasing and converges to $A > 0$. If $g(0) = A - a^2/\mu \ge 0$, then $\mathcal{F}'_{\text{eff}}(w)$ has the same sign as w, and the origin is the unique minimum. If $g(0) < 0$, then g has a unique zero on the positive half-line, producing a symmetric pair of nonzero minima while the origin becomes a local maximum. The boundary is therefore $g(0) = 0$. □

Proposition D.2 (Equilibrium branches and the mean-field critical exponent)**.** *Suppose that $u(\lambda_c) > 0$ and*

$$r(\lambda) = c(\lambda - \lambda_c) + o(\lambda - \lambda_c), \qquad c > 0, \tag{69}$$

Then the stable equilibrium branches at zero external field satisfy

$$w^\star = 0 \quad (r > 0), \qquad |w^\star| = \sqrt{-\frac{r}{u}} + O(|r|^{3/2}) \quad (r < 0). \tag{70}$$

Consequently,

$$|w^\star| \sim \sqrt{\frac{c}{u(\lambda_c)}}\,(\lambda_c - \lambda)^{1/2}, \tag{71}$$

i.e., the order parameter has the mean-field critical exponent $1/2$.

Proof. From $\partial_w \mathcal{F}_{\text{eff}} = rw + uw^3 + O(w^5) = 0$, the nonzero branch satisfies $w^2 = -r/u + O(r^2)$. When $u(\lambda_c) > 0$, this branch is a local minimum on the $r < 0$ side. Substituting Eq. (69) then gives the result. □

Proposition D.3 (Second-order nonanalyticity of the minimum free energy)**.** *Under the conditions of Proposition D.2,*

$$\mathcal{F}_{\min}(\lambda) - \mathcal{F}_{\min}(\lambda_c) = \begin{cases} 0, & r \geq 0, \\ -\dfrac{r^2}{4u(\lambda_c)} + O(|r|^3), & r < 0. \end{cases} \tag{72}$$

Thus the equilibrium order parameter emerges continuously, while the first derivative of the minimum free energy with respect to the control parameter remains continuous and the second derivative jumps at the critical point. This is the standard mean-field second-order Landau transition.

Proof. For $r \geq 0$, the local minimum is $w^\star = 0$. For $r < 0$, substituting $w_\star^2 = -r/u + O(r^2)$ into $\mathcal{F}_{\text{eff}} - \mathcal{F}_0 = rw^2/2 + uw^4/4 + O(w^6)$ yields Eq. (72). □

Theorem D.4 (Exact loading and the distinct roles of quadratic and quartic terms)**.** *If $B > 0$, the stable equilibrium of Eq. (50) satisfies*

$$|\beta^\star| = \frac{\left[|a| - \sqrt{A\mu}\right]_+}{B}. \tag{73}$$

Thus λ_2 drives the loading exactly to zero at the finite strength $\lambda_{2,c} = (a^2/\mu - \mu)/\gamma$ whereas a pure λ_4 term only compresses a nonzero loading continuously by increasing B.

Proof. A nonzero stationary point satisfies

$$Aw = \theta(a - B\beta), \qquad \mu\theta = w(a - B\beta). \tag{74}$$

Multiplying the two equations and canceling the nonzero factor $w\theta$ gives

$$(a - B\beta)^2 = A\mu. \tag{75}$$

Selecting the branch with the same sign as the task drive and that connects continuously to zero as $a^2 \downarrow A\mu$ gives Eq. (73). □

Sixth-order loading extension. For the higher-order predictions in Fig. 3(b), we apply the same soft-mode reduction but retain the sextic coefficient. Writing the sixth-order truncation as

$$\mathcal{F}_{\text{eff}}^{(6)}(q) = \mathcal{F}_0 + \frac{1}{2}rq^2 + \frac{1}{4}uq^4 + \frac{1}{6}v_6q^6, \tag{76}$$

and setting $s = q^2$, every nonzero stationary branch satisfies

$$r + us + v_6s^2 = 0, \qquad s_\star = \frac{-u + \sqrt{u^2 - 4v_6r}}{2v_6}, \tag{77}$$

where the displayed root is the branch continuous with $s_\star = -r/u$ as $v_6 \to 0$ (and the stable positive root is selected whenever more than one real branch exists). For a scalar bilinear soft mode, the slaved loading has the even expansion

$$\beta(q) = b_2q^2 + b_4q^4 + O(q^6), \qquad |\beta^\star| = \left|b_2s_\star + b_4s_\star^2\right| + O(s_\star^3). \tag{78}$$

The limit $v_6 \to 0$ recovers the quartic branch, while the canonical explicit R_2/R_4 model further admits the exact loading formula in Eq. (73). Thus the quartic and sixth-order predictions use the same reduced equilibrium law, differing only in the retained low-order coefficients. For V-REx and IGA these coefficients are obtained from the objective expansions in Eqs. (117) and (129); they are not fitted to the trained loading paths.

Proposition D.5 (Selective-retention window)**.** *Suppose that the stable and shortcut directions have critical values $\lambda_{2,c}^{(c)}$ and $\lambda_{2,c}^{(s)}$, respectively, with $\lambda_{2,c}^{(s)} < \lambda_{2,c}^{(c)}$. Then any*

$$\lambda_{2,c}^{(s)} < \lambda_2 < \lambda_{2,c}^{(c)} \tag{79}$$

places the shortcut direction in the disordered phase while the stable direction remains ordered.

E ONSET DIAGNOSTICS, SUSCEPTIBILITY, AND LANDAU DYNAMICS

We first explain why the early-time growth-rate zero crossing used in the experiments measures the same static phase boundary, and then give a local dynamical interpretation under deterministic continuous-time gradient flow.

E.1 LINEARIZED ONSET DIAGNOSTIC

Let $z = (w, \theta)^\top$. The continuous-time gradient flow of the canonical objective linearized near the origin is

$$\dot{z} = -H_0 z, \qquad H_0 = \begin{pmatrix} A & -a \\ -a & \mu \end{pmatrix}. \tag{80}$$

The fastest local growth rate is $-\lambda_{\min}(H_0)$, so its zero crossing occurs when

$$\lambda_{\min}(H_0) = 0 \quad \Longleftrightarrow \quad \det H_0 = A\mu - a^2 = 0 \quad \Longleftrightarrow \quad r(\lambda_2) = 0. \tag{81}$$

Thus the measured dynamical zero crossing is a local numerical diagnostic of the static Landau boundary.

E.2 NONLINEAR SOFT-MODE DYNAMICS UNDER GRADIENT FLOW

Consider deterministic continuous-time gradient flow

$$\dot{\Theta} = -\nabla_\Theta \mathcal{F}(\Theta; \lambda). \tag{82}$$

At a simple soft-mode critical point satisfying the spectral assumptions of Theorem B.2, the linearized flow has one neutral direction and all transverse directions decay. Both reductions use the order-parameter coordinate q along the critical direction v_c. The equilibrium map $\eta^\star$ enforces transverse stationarity; the dynamical map η_{cm} enforces invariance under gradient flow. The center-manifold theorem gives a locally invariant one-dimensional manifold, which we parameterize as

$$\Theta_{\mathrm{cm}}(q, \lambda) = \bar{\Theta}(\lambda) + q v_c + \eta_{\mathrm{cm}}(q, \lambda), \qquad \eta_{\mathrm{cm}} \perp v_c. \tag{83}$$

Define the objective restricted to this manifold by $\mathcal{F}_{\mathrm{cm}}(q; \lambda) = \mathcal{F}(\Theta_{\mathrm{cm}}(q, \lambda); \lambda)$.

Proposition E.1 (Time-dependent Landau normal form)**.** *Under the same local $\mathbb{Z}_2$ symmetry and transverse-crossing assumptions as in Theorem B.2, with $u(\lambda_c) > 0$, the center-manifold dynamics satisfy*

$$\tau(q, \lambda)\dot{q} = -\frac{\partial \mathcal{F}_{\mathrm{cm}}}{\partial q}, \qquad \tau(q, \lambda) = \|\partial_q \Theta_{\mathrm{cm}}(q, \lambda)\|^2 > 0. \tag{84}$$

Near criticality,

$$\mathcal{F}_{\mathrm{cm}}(q; \lambda) = \mathcal{F}_0(\lambda) + \frac{1}{2} r(\lambda) q^2 + \frac{1}{4} u(\lambda) q^4 + \mathcal{O}(q^6), \tag{85}$$

so, after absorbing the positive local mobility into the time scale,

$$\tau_c \dot{q} = -r(\lambda) q - u(\lambda_c) q^3 + \mathcal{O}\big(q^5 + |\lambda - \lambda_c| q^3\big), \qquad \tau_c > 0. \tag{86}$$

Proof sketch. Because the center manifold is invariant under Eq. (82), $\partial_q \Theta_{\mathrm{cm}}\, \dot{q} = -\nabla_\Theta \mathcal{F}$ along it. Taking the inner product with $\partial_q \Theta_{\mathrm{cm}}$ gives $\|\partial_q \Theta_{\mathrm{cm}}\|^2 \dot{q} = -\partial_q \mathcal{F}_{\mathrm{cm}}$, which proves Eq. (84). The symmetry makes $\mathcal{F}_{\mathrm{cm}}$ even in q. With $r(\lambda_c) = 0$ and the assumed $u(\lambda_c) > 0$, differentiating the Landau expansion yields Eq. (86).

For comparison, the spatially uniform Landau free energy of a nonconserved scalar order parameter is

$$F_{\mathrm{mag}}(M) = F_0 + \frac{1}{2} a_0 (T - T_c) M^2 + \frac{1}{4} b M^4 + \mathcal{O}(M^6), \tag{87}$$

and its standard relaxational time-dependent Landau dynamics are (Hohenberg & Krekhov, 2015)

$$\tau_M \dot{M} = -\frac{\partial F_{\mathrm{mag}}}{\partial M} = -a_0 (T - T_c) M - b M^3 + \mathcal{O}(M^5). \tag{88}$$

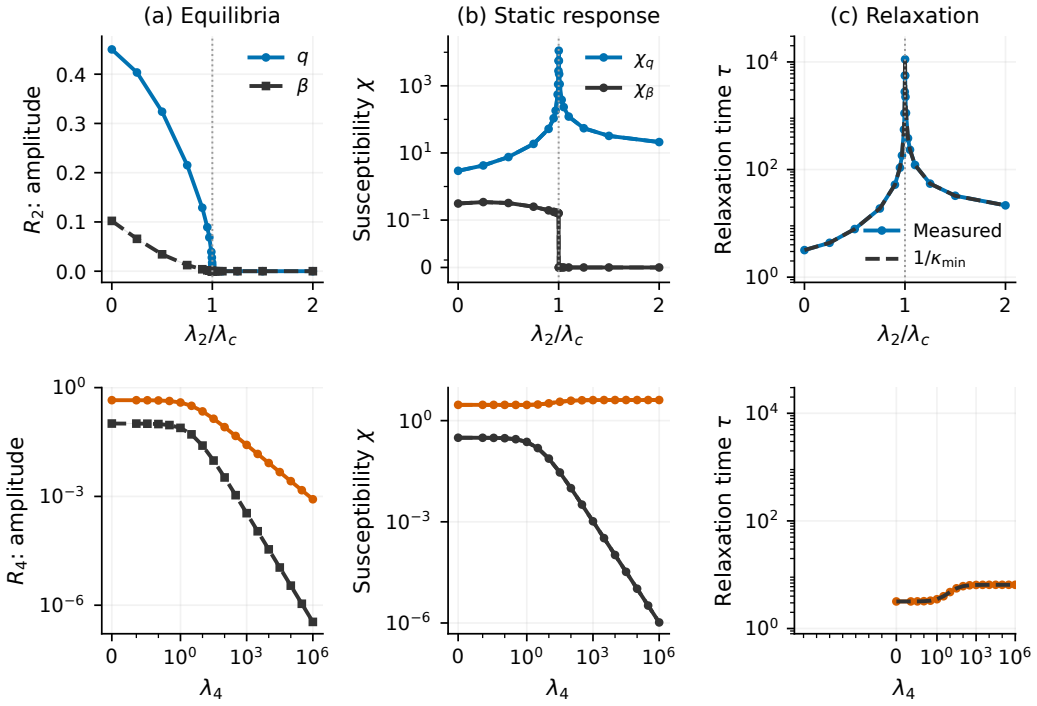


Figure 8: Equilibrium amplitudes, static responses, and relaxation times in the smooth network under R_2 (top) and R_4 (bottom). In the response and relaxation panels, markers are numerical measurements and dashed curves are Hessian predictions. Time is in gradient-flow units; R_2 strength is normalized by each seed's critical value.

Under the correspondence $q \leftrightarrow M$, $r(\lambda) \leftrightarrow a_0(T - T_c)$, and $u \leftrightarrow b$, Eqs. (86) and (88) have the same local normal form. This spatially uniform relaxational dynamics is often referred to as Model-A or time-dependent Ginzburg–Landau dynamics. In particular, in the linear regime,

$$\frac{d}{dt} \log |q| = -\frac{r(\lambda)}{\tau_c} + o(1), \tag{89}$$

so the growth-rate zero crossing, the Hessian sign change, and the Landau mass condition $r = 0$ identify the same local phase boundary.

The same normal form predicts critical slowing down. On the disordered side $r > 0$, linearization around $q^\star = 0$ gives

$$\tau_{\text{relax}} = \frac{\tau_c}{r} [1 + o(1)]. \tag{90}$$

On the ordered side $r < 0$, where $(q^\star)^2 = -r/u + O(r^2)$, the local curvature is $\partial_{qq} \mathcal{F}_{\text{cm}}(q^\star) = -2r + o(|r|)$ and hence

$$\tau_{\text{relax}} = \frac{\tau_c}{2|r|} [1 + o(1)]. \tag{91}$$

Thus the relaxation time diverges as $|r|^{-1}$ when the phase boundary is approached from either side, giving the standard mean-field critical slowing down of relaxational Landau dynamics (Hohenberg & Krekhov, 2015).

E.3 Susceptibility and critical slowing down

We test whether the same critical curvature controls equilibrium response and relaxation in the canonical bilinear model and a smooth finite network. These experiments compare explicit R_2 and R_4 controls under deterministic gradient flow.

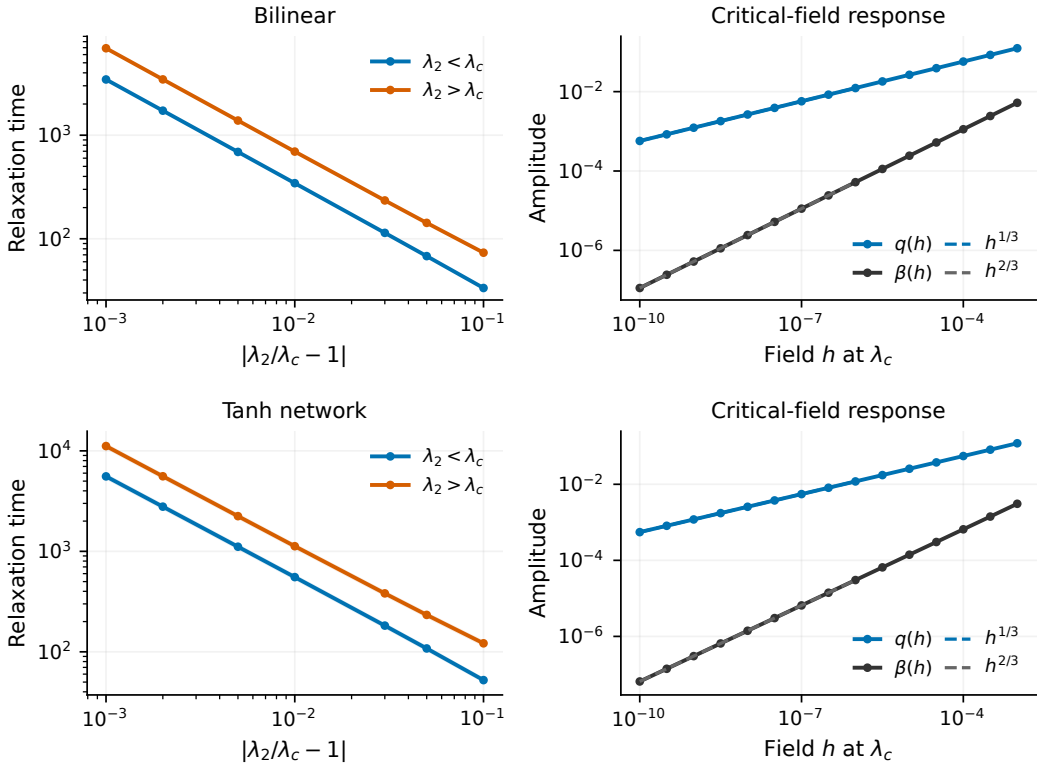


Figure 9: Critical scaling in the bilinear and smooth-network controls. Left: relaxation time versus distance from the R_2 boundary. Right: amplitudes under $-hq$ at criticality; dashed guides show $h^{1/3}$ and $h^{2/3}$ scaling, anchored at the smallest sampled field.

Response predictions. A field conjugate to the reduced coordinate gives

$$\mathcal{F}_{\text{eff}}(q;h) = \mathcal{F}_0 + \frac{1}{2}rq^2 + \frac{1}{4}uq^4 - hq + O(q^6). \tag{92}$$

For $u > 0$, differentiating the equilibrium condition on a fixed stable branch yields

$$\chi_q = \left.\frac{\partial q^\star}{\partial h}\right|_{h=0} = \begin{cases} r^{-1}[1+o(1)], & r > 0, \\ (2|r|)^{-1}[1+o(1)], & r < 0. \end{cases} \tag{93}$$

Together with the relaxation law above and a transverse crossing of r, this predicts $\chi_q, \tau_{\text{relax}} \propto |\lambda_2 - \lambda_c|^{-1}$. At $r = 0$, the nonlinear response satisfies $q \sim (h/u)^{1/3}$. Increasing quartic stiffness at fixed negative mass can reduce the equilibrium amplitude while its curvature remains finite, distinguishing continuous attenuation from critical softening.

For a differentiable probe $p(\Theta)$ of the full network, consider $\mathcal{F}_h = \mathcal{F} - hp$. At a nondegenerate stable equilibrium with Hessian H, differentiating the full stationarity equations gives the exact local response

$$\chi_p = \left.\frac{d\,p(\Theta_h^\star)}{dh}\right|_{h=0} = \nabla p^\top H^{-1} \nabla p. \tag{94}$$

Models and environments. The bilinear control uses Eq. (4) with $(a, m, \gamma, \mu) = (0.5, 2.25, 1, 0.2)$, giving $\lambda_c = 1.05$. The smooth network has eight inputs and one $\tanh$ unit, $f_\Theta(X) = \theta \tanh(w^\top X)$, with parameters $\Theta = (w, \theta)$. For fixed orthonormal directions $b, c \in \mathbb{R}^8$, we draw

$$\widetilde{X}_i = (0.5b + 0.06c)Y_i + \xi_i, \qquad X_{ei} = \widetilde{X}_i - \overline{\widetilde{X}} + d_e b, \quad d_e \in \frac{1}{\sqrt{5}}\{-3, -1, 1, 3\}, \tag{95}$$

where $Y_i \in \{-1, 1\}$ is balanced and $\xi_i \sim \mathcal{N}(0, I_8)$. Each environment contains the same 1,024 base examples, differing only in its offset. The task objective is mean squared error with factor $1/2$ and ridge $\mu\|\Theta\|^2/2$, where $\mu = 0.2$. For the nonlinear experiment, we construct two mean-matching controls whose leading contributions near the zero reference are quadratic (R_2) and quartic (R_4), respectively. Writing $\bar{\phi}_e$ for the environment mean of $\tanh(w^\top X)$ and $g_e = (\bar{\phi}_e - \bar{\phi})/2$, we set

$$\Omega_2 = \frac{\lambda_2}{2}\|g(\Theta)\|^2, \qquad \Omega_4 = \frac{\lambda_4}{2}\|\theta g(\Theta)\|^2. \tag{96}$$

At the zero reference, $g = 0$; with $J_E = \partial g/\partial\Theta$, the environmental quadratic operator is $G_E = J_E^\top J_E$ for R_2, whereas R_4 has zero Hessian there. The critical reference has a one-dimensional Hessian kernel and positive transverse curvature in every run.

Both scans use $q = v_c^\top \Theta$, with v_c the unit critical eigenvector at the R_2 boundary, held fixed along both scans. We separately measure predictive loading: $\beta = w\theta$ in the bilinear model and

$$\beta(\Theta) = \frac{\widehat{\mathbb{E}}[(b^\top X) f_\Theta(X)]}{\widehat{\mathbb{E}}[(b^\top X)^2]} \tag{97}$$

in the smooth network, with empirical averages over the pooled training environments. Thus q is a parameter-space critical coordinate, while β measures the predictor's linear projection onto the designated input direction. Figures show unnormalized amplitudes.

Observable-specific responses. We denote the response to $\mathcal{F} - h_q q$ by χ_q (order-parameter susceptibility), and that to $\mathcal{F} - h_\beta \beta$ by χ_β (predictive-loading response). Each differentiates its own observable with respect to its own field at zero field, as in Eq. (94). In the canonical model, $-h_\beta\beta$ replaces the task drive a by $a + h_\beta$, so $\chi_\beta = \partial\beta^\star/\partial a$ with the other coefficients fixed. Near criticality, $\beta \propto q^2$, so its conjugate perturbation changes the quadratic coefficient and its response can remain finite as χ_q diverges.

Measurements. We solve the full stationarity equations and verify positive Hessian curvature at each equilibrium. The Hessian zero crossing determines λ_c. For each probe, central differences between equilibria of $\mathcal{F} \mp hp$ measure susceptibility; repeating the measurement at $h/2$ checks the linear-response regime. On the acquired side, both perturbations follow the same local branch. Relaxation uses unit-mobility gradient flow $\dot{\Theta} = -\nabla_\Theta \mathcal{F}$. For ordinary gradient descent with a small fixed step size η, $t \simeq \eta k$ after k steps, so $\tau_{\text{steps}} \simeq \tau/\eta$. At each equilibrium $\Theta^\star$, let $(\kappa_{\min}, v_{\min})$ be the smallest eigenpair of $H = \nabla^2_\Theta \mathcal{F}(\Theta^\star)$, with $\|v_{\min}\| = 1$ and $\kappa_{\min} > 0$. We integrate from $\Theta_\pm(0) = \Theta^\star \pm \epsilon v_{\min}$ and fit the projected displacement:

$$d(t) = \frac{1}{2} v_{\min}^\top [\Theta_+(t) - \Theta_-(t)] \simeq d(0) e^{-t/\tau}, \qquad \tau_{\text{pred}} = \kappa_{\min}^{-1}.$$

The plotted τ is the fitted time for the local parameter-displacement amplitude to decrease by a factor e, in the above gradient-flow units. The current direction $v_{\min}$ determines the slowest local relaxation, while q uses the fixed critical direction v_c; the relation in Eq. (10) holds asymptotically near criticality. We use five independent data and initialization seeds for the smooth network, and five initialization repeats of the fixed bilinear population objective. Figures report means and standard errors.

Critical response and relaxation. The R_2 paths show the predicted inverse-distance scaling of χ_q and the relaxation time (Figs. 8 and 9). At criticality, the finite-field response follows $q \propto h^{1/3}$ and $\beta \propto h^{2/3}$, consistent with $\beta \propto q^2$. In these controls, R_4 attenuation retains finite relaxation times even at small loading.

E.4 Optimization steps near equilibrium for invariant-learning objectives

We test local relaxation under the eight full objectives in Fig. 11. The bilinear predictor is $f_\Theta(X) = w\theta X$, with representation wX. Four equally weighted environments have balanced $Y \in \{-1, 1\}$ and

$$X_e = (0.5 + 0.2\zeta_e)Y + 0.5\zeta_e + \epsilon, \qquad \zeta_e \in \frac{1}{\sqrt{5}}\{-3, -1, 1, 3\}, \quad \epsilon \sim \mathcal{N}(0, 1). \tag{98}$$

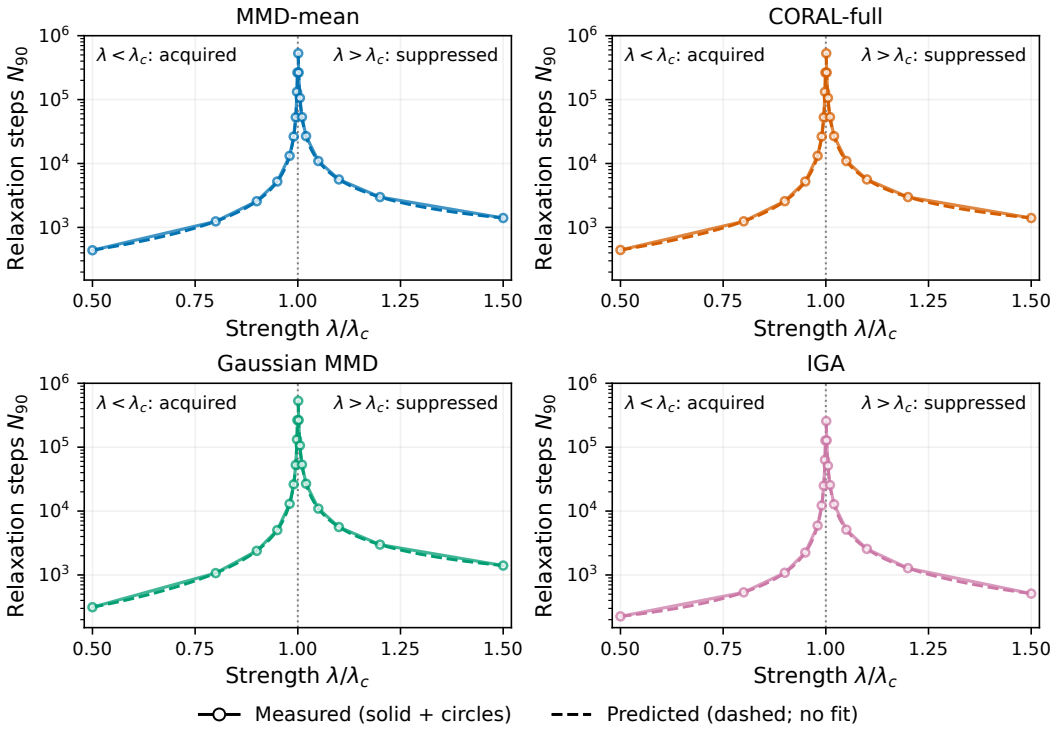


Figure 10: Critical slowing down near $\lambda_c = 2.1$ for MMD-mean, CORAL-full, and Gaussian MMD, and $\lambda_c = 3.75$ for IGA. Solid curves with circles show measured N_{90}; dashed curves show Hessian predictions. Dotted lines mark the critical boundary, approached to relative distance 10^{-3} on either side.

We use population squared loss with factor $1/2$, weight decay $0.2\|(w, \theta)\|^2/2$, and the environment-variance normalization of Appendix G. Gaussian MMD uses a unit-bandwidth RBF kernel; Fishr uses per-example classifier-gradient variances.

Measured steps and predictions. At each stable equilibrium, we apply gradient descent with $\eta = 0.03$ to paired perturbations along the minimum-curvature Hessian eigenvector, as in Appendix E.3. Writing their projected half-difference as d_k, let N_{90} be the first iteration with $|d_k/d_0| \le 0.1$. For Hessian eigenvalue $\kappa_{\min}$, linearization predicts

$$\frac{d_k}{d_0} \simeq (1 - \eta\kappa_{\min})^k, \qquad N_{90}^{\text{pred}} = \frac{\log 0.1}{\log(1 - \eta\kappa_{\min})} \simeq \frac{\log 10}{\eta\kappa_{\min}}. \tag{99}$$

Perturbation amplitudes are $10^{-3}\|\Theta^\star\|$ on the acquired branch and $10^{-3}\min(0.1, \sqrt{\kappa_{\min}})$ at the origin; all sampled points satisfy $0 < \eta\kappa_{\max} < 1$.

Figures 10 and 11 show inverse-distance slowing near the R_2 boundaries, while the R_4-leading paths retain finite relaxation times over this scan. In this setting, IRM's negative R_6 term softens the nonzero equilibrium, producing a finite peak in relaxation steps while its minimum Hessian curvature remains positive. Hessian predictions agree with measured N_{90} within 0.6%; Fig. 12 shows the underlying CORAL-full trajectories.

F Collective Modes and Environment-Induced Mass

Let $C = \mathrm{diag}(c_1, \ldots, c_p)$ and $M = \mu I + \lambda_2 \Gamma_E$, with $\Gamma_E \succeq 0$. The multi-feature quadratic Hessian is

$$H_0^{(p)} = \begin{pmatrix} M & -C \\ -C^\top & \mu I \end{pmatrix}. \tag{100}$$

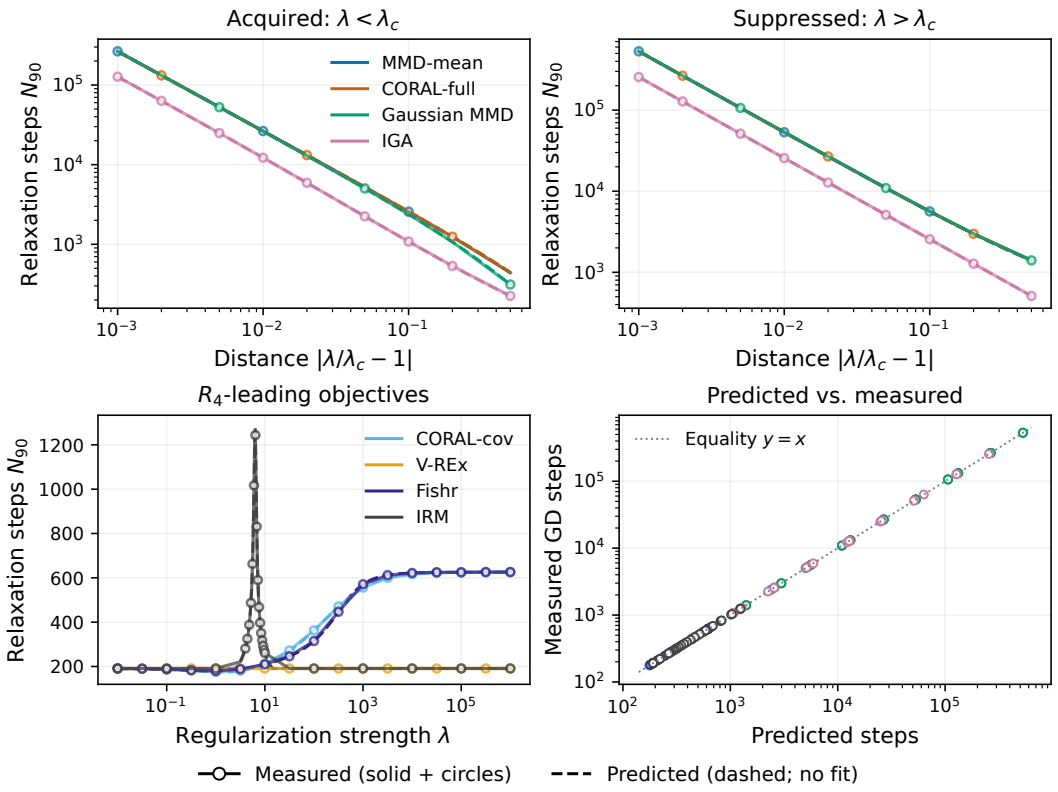


Figure 11: Top: distance-to-boundary scaling of N_{90}. Bottom left: R_4-leading paths at this reference. Bottom right: predicted versus measured steps, with dotted diagonal $y = x$. Line conventions follow Fig. 10.

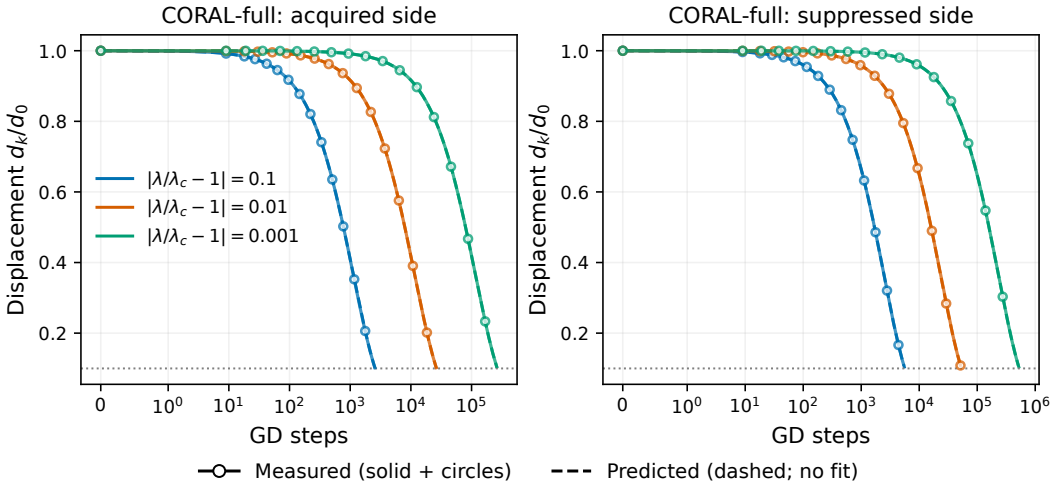


Figure 12: CORAL-full relaxation on both sides of the boundary. Colors denote distance to criticality. Solid curves with circles show d_k/d_0; dashed curves show $(1 - \eta\kappa_{\min})^k$. The dotted line marks the 0.1 threshold defining N_{90}.

Theorem F.1 (Collective-mode onset)**.** *Assume $M \succ 0$. A collective direction grows away from the origin if and only if*

$$\lambda_{\max}\left[\frac{1}{\mu}C^\top M^{-1}C\right] > 1. \tag{101}$$

At criticality, the largest eigenvalue equals 1*. The corresponding gating direction can be written as $w_c \propto M^{-1}Cu_c$, where u_c is the leading eigenvector in Eq.* (101)*.*

Proof. Because $M \succ 0$, taking the Schur complement of the upper-left block in Eq. (100) gives

$$H_0^{(p)} \succeq 0 \quad \Longleftrightarrow \quad S \equiv \mu I - C^\top M^{-1}C \succeq 0. \tag{102}$$

Multiplying S on both sides by $\mu^{-1/2}I$ shows that S has a negative eigenvalue if and only if the matrix in Eq. (101) has an eigenvalue larger than 1. At the boundary, the quadratic stationarity equation $Mw = C\theta$ gives $w_c \propto M^{-1}Cu_c$. □

Proof of Proposition 5.2. For $\Gamma_E \succeq 0$, $x^\top\Gamma_E x = \left\|\Gamma_E^{1/2}x\right\|^2$ vanishes exactly when $x \in \ker\Gamma_E$. Applying this identity to $\boldsymbol{w}$ and $\boldsymbol{\beta}$, with positive λ_2, λ_4, proves both equivalences. To see why zero gate penalty need not imply zero predictor penalty, take $\Gamma_E = dd^\top$ with $d = (1,-1)^\top$, $\boldsymbol{w} = (1,1)^\top$, and $\boldsymbol{\theta} = (1,2)^\top$. Then $\boldsymbol{\beta} = (1,2)^\top$, so $R_2(\boldsymbol{w}) = 0$ but $R_4(\boldsymbol{\beta}) = \lambda_4/2 > 0$. □

Compensatory example (Fig. 7). Let $\delta_c, \delta_s > 0$ denote the environmental mean-shift amplitudes of the pair and shortcut, and set $d_{\text{pair}} = (\delta_c, -\delta_c, 0)^\top$ and $d_{\text{short}} = (0, 0, \delta_s)^\top$. The oracle sensitivity matrix is

$$\Gamma_E = d_{\text{pair}}d_{\text{pair}}^\top + d_{\text{short}}d_{\text{short}}^\top, \tag{103}$$

so direct expansion gives

$$\boldsymbol{w}^\top\Gamma_E\boldsymbol{w} = \delta_c^2(w_1 - w_2)^2 + \delta_s^2 w_3^2, \qquad \boldsymbol{\beta}^\top\Gamma_E\boldsymbol{\beta} = \delta_c^2(\beta_1 - \beta_2)^2 + \delta_s^2\beta_3^2. \tag{104}$$

Hence the compensated direction $(1,1,0)^\top/\sqrt{2}$ lies in $\ker(\Gamma_E)$, whereas the leakage direction $(1,-1,0)^\top/\sqrt{2}$ and shortcut direction e_3 have positive sensitivity. Matrix R_4 directly penalizes predictor-space leakage $\beta_1 - \beta_2$ and shortcut loading β_3, leaving the compensated direction unpenalized; matrix R_2 applies the same geometry to the gates $\boldsymbol{w}$.

G Derivation of objective signatures

This section derives, term by term, the fingerprints of all methods appearing in the Figure 5 legend: ERM, GroupDRO, IRMv1, V-REx, IGA, Fish/MLDG, Fishr, Mixup, MMD-mean, Gaussian MMD, CORAL-cov, CORAL-full, DANN, and the explicit R_2, R_4, and R_2+R_4 objectives. All identities refer to the bilinear population probe in Eq. (18). Their application to deep networks and alternating optimization describes local operators under the stated reference assumptions.

We use Var_{P_e} and Cov_{P_e} for within-environment moments, and Var_e and Cov_e for variation across environments. Let $m_e = \mathbb{E}_e[X]$ and $s_e = \mathrm{Cov}_{P_e}(X)$, and define

$$\begin{aligned} M_2 &= \overline{(m_e - \bar{m})^2}, \quad S_2 = \overline{(s_e - \bar{s})^2}, \quad V_c = \overline{(c_e - \bar{c})^2}, \\ C_{cq} &= \overline{(c_e - \bar{c})(q_e - \bar{q})}, \quad V_q = \overline{(q_e - \bar{q})^2}. \end{aligned} \tag{105}$$

If an implementation uses pairwise environment penalties, it differs from Eq. (105) only by a constant depending on the number of environments; we absorb that constant into the corresponding λ. The quantities Δa_2 and Δa_4 below are reported under the convention of Eq. (20).

G.1 ERM and explicit R_2/R_4 regularizers

The environment-averaged ERM objective is exactly Eq. (19):

$$\mathcal{F}_{\text{ERM}}^{(2)} = -\bar{c}\,w\theta, \qquad \mathcal{F}_{\text{ERM}}^{(4)} = \frac{\bar{q}}{2}w^2\theta^2. \tag{106}$$

Thus it defines the mixed task curvature and quartic data stabilization at zero regularization, without adding an independent R_2 or R_4 operator.

The explicit quadratic and quartic environment regularizers are

$$\Omega_{R_2} = \frac{\lambda_2\gamma}{2}w^2, \qquad \Omega_{R_4} = \frac{\lambda_4\gamma}{2}(w\theta)^2. \tag{107}$$

Substituting $w = \rho v_w$ and $\theta = \rho v_\theta$ gives

$$\Delta a_2^{R_2} = \lambda_2\gamma v_w^2, \qquad \Delta a_4^{R_4} = 2\lambda_4\gamma(v_w v_\theta)^2. \tag{108}$$

For $R_2 + R_4$, the two corrections add directly. This also shows that even when both terms appear in the learning objective, only the former changes the Hessian at the origin.

G.2 MMD-mean, Gaussian MMD, and CORAL

Here MMD-mean denotes MMD with a linear kernel, equivalently squared mean matching in representation space. For mean matching, if the representation is $h = wX$, the difference in environment means is $\mathbb{E}_e[h] - \overline{\mathbb{E}[h]} = w(m_e - \bar{m})$. Hence the local MMD-mean term is

$$\Omega_{\text{MMD-mean}} = \lambda_\mu M_2 w^2, \qquad \Delta a_2 = 2\lambda_\mu M_2 v_w^2. \tag{109}$$

This is a genuine gate-block R_2 mass. By contrast, the centered covariance satisfies

$$\mathrm{Cov}_{P_e}(h) = w^2 s_e, \qquad \Omega_{\text{CORAL-cov}} = \lambda_\Sigma S_2 w^4, \qquad \Delta a_4 = 4\lambda_\Sigma S_2 v_w^4. \tag{110}$$

Thus covariance-only CORAL begins at quartic order. Full CORAL is the sum of the two: its R_2 component comes from mean matching and its R_4 component from covariance matching.

For Gaussian MMD between two environment distributions P, Q, let $h = tX$ and consider the single-bandwidth kernel

$$k_\sigma(tX, tX') = \exp\left[-\frac{t^2(X - X')^2}{2\sigma^2}\right]. \tag{111}$$

Define

$$C_{2j}(P, Q) = \mathbb{E}_{P,P}(X - X')^{2j} + \mathbb{E}_{Q,Q}(X - X')^{2j} - 2\mathbb{E}_{P,Q}(X - X')^{2j}. \tag{112}$$

Expanding the exponential term by term gives the exact even-power series

$$\mathrm{MMD}_\sigma^2(tP, tQ) = \sum_{j\geq 1} \frac{(-1)^j t^{2j}}{j!(2\sigma^2)^j} C_{2j}(P, Q) = A_2 t^2 + A_4 t^4 + O(t^6). \tag{113}$$

Averaging over environment pairs and kernel bandwidths preserves this form, with A_2 and A_4 averaged accordingly. Therefore,

$$\Omega_{\text{MMD-Gaussian}} = \lambda_G(A_2 w^2 + A_4 w^4 + O(w^6)), \quad \Delta a_2 = 2\lambda_G A_2 v_w^2, \quad \Delta a_4 = 4\lambda_G A_4 v_w^4. \tag{114}$$

When environment means differ, typically $A_2 > 0$, so the leading term is R_2. If the mean term vanishes by symmetry, the leading term may instead be R_4 or higher order. This is why Gaussian MMD cannot be assigned a data-independent fixed Landau order.

G.3 IRMv1, V-REx, and GroupDRO

IRMv1. Let the classifier scaling be α. By Eq. (18), at $\alpha = 1$,

$$\partial_\alpha \ell_e(\alpha\beta)\big|_{\alpha=1} = -c_e\beta + q_e\beta^2. \tag{115}$$

Therefore, under the environment-average convention, the IRMv1 penalty satisfies exactly

$$\begin{aligned} \Omega_{\text{IRM}} &= \lambda_{\text{IRM}}\overline{\left(-c_e\beta + q_e\beta^2\right)^2} \\ &= \lambda_{\text{IRM}}\left[\overline{c_e^2}\beta^2 - 2\overline{c_e q_e}\beta^3 + \overline{q_e^2}\beta^4\right], \\ \Delta a_4^{\text{IRM}} &= 4\lambda_{\text{IRM}}\overline{c_e^2}(v_w v_\theta)^2. \end{aligned} \tag{116}$$

Because $\beta = w\theta = O(\|(w, \theta)\|^2)$, the three terms are of fourth, sixth, and eighth order, respectively. IRMv1 has no independent R_2 term in (w, θ); using a sum rather than an average merely rescales λ_{IRM}.

V-REx. Treat the environmental variation of $L_{0,e}$ as the constant term. Direct expansion of Eq. (18) gives

$$\begin{aligned}\mathrm{Var}_e(\ell_e) = \; & \mathrm{Var}_e(L_{0,e}) - 2\,\mathrm{Cov}_e(L_{0,e}, c_e)\beta \\ & + \big[\mathrm{Var}_e(c_e) + \mathrm{Cov}_e(L_{0,e}, q_e)\big]\beta^2 \\ & - \mathrm{Cov}_e(c_e, q_e)\beta^3 + \frac{1}{4}\,\mathrm{Var}_e(q_e)\beta^4.\end{aligned} \tag{117}$$

Hence the genuine R_2 fingerprint of $\lambda_{\text{V-REx}}\,\mathrm{Var}_e(\ell_e)$ is

$$\Delta a_2^{\text{V-REx}} = -4\lambda_{\text{V-REx}}\,\mathrm{Cov}_e(L_{0,e}, c_e) v_w v_\theta. \tag{118}$$

Under the equal-null-risk reference used in Figure 5, this term and $\mathrm{Cov}_e(L_{0,e}, q_e)$ both vanish, so

$$\Delta a_2^{\text{V-REx}} = 0, \qquad \Delta a_4^{\text{V-REx}} = 4\lambda_{\text{V-REx}} V_c (v_w v_\theta)^2. \tag{119}$$

V-REx is therefore R_4-leading under this equal-null-risk reference.

GroupDRO. GroupDRO reweights environment-specific task risks toward high-loss environments. With equal weight decay μ on both parameter blocks, its hard objective is

$$\mathcal{F}_{\text{GDRO}} = \max_{p \in \Delta} \sum_e p_e \ell_e + \frac{\mu}{2}(w^2 + \theta^2). \tag{120}$$

During training, $p_{e,t+1} \propto p_{e,t} \exp[\eta \ell_e(w_t, \theta_t)]$, normalized over environments, with update step size η. Holding the current weights p fixed during a predictor update gives

$$\mathcal{F}_p = L_{0,p} - c_p w\theta + \frac{q_p}{2} w^2\theta^2 + \frac{\mu}{2}(w^2 + \theta^2), \tag{121}$$

where $c_p = \sum_e p_e c_e$ and $q_p = \sum_e p_e q_e$. Relative to uniform ERM, $\Delta c_p = c_p - \bar{c}$ and $\Delta q_p = q_p - \bar{q}$ renormalize the R_2 task drive and R_4 stiffness, with joint-variable coefficients

$$\Delta a_2^{\text{GroupDRO}} = -2\Delta c_p v_w v_\theta, \qquad \Delta a_4^{\text{GroupDRO}} = 2\Delta q_p (v_w v_\theta)^2. \tag{122}$$

At a stable equilibrium with limiting weights $p^\star$ and $q_{p^\star} > 0$, Eq. (73) gives

$$|\beta^\star_{p^\star}| = \frac{[|c_{p^\star}| - \mu]_+}{q_{p^\star}}. \tag{123}$$

Reweighting can reduce shortcut loading while a residual task drive $|c_{p^\star}| > \mu$ sustains a nonzero plateau.

G.4 IGA, FISH/MLDG, AND FISHR

Let $z = (w, \theta)^\top$, and write each environment risk as

$$\ell_e = L_{0,e} + \frac{1}{2} z^\top H_e z + r_e^{(4)}(z), \qquad H_e = \begin{pmatrix} 0 & -c_e \\ -c_e & 0 \end{pmatrix}, \qquad r_e^{(4)}(z) = \frac{q_e}{2} w^2\theta^2. \tag{124}$$

IGA. IGA penalizes dispersion among environment gradients. Let $\Delta H_e = H_e - \bar{H}$ and $\Delta r_e^{(4)} = r_e^{(4)} - \bar{r}^{(4)}$. Then

$$\begin{aligned}\Omega_{\text{IGA}} &= \lambda_{\text{IGA}} \overline{\big\|\nabla \ell_e - \overline{\nabla \ell_e}\big\|^2} \\ &= \lambda_{\text{IGA}} z^\top \overline{\Delta H_e^\top \Delta H_e} z + 2\lambda_{\text{IGA}} \overline{\Big\langle \Delta H_e z, \nabla \Delta r_e^{(4)}(z) \Big\rangle} + O(\|z\|^6).\end{aligned} \tag{125}$$

Under the Hessian convention $\frac{1}{2} z^\top K z$,

$$K_{\text{IGA}} = 2\lambda_{\text{IGA}} \overline{\Delta H_e^\top \Delta H_e} \succeq 0. \tag{126}$$

Thus IGA contributes a positive-semidefinite matrix R_2 over the full parameter space and can rotate the unstable direction. In the scalar probe, Eq. (124) gives

$$\begin{aligned}\nabla_w \ell_e - \overline{\nabla_w \ell_e} &= -\Delta c_e\, \theta + \Delta q_e\, w\theta^2, \\ \nabla_\theta \ell_e - \overline{\nabla_\theta \ell_e} &= -\Delta c_e\, w + \Delta q_e\, w^2\theta.\end{aligned} \tag{127}$$

$$\Omega_{\mathrm{IGA}} = \lambda_{\mathrm{IGA}} V_c(w^2+\theta^2) + O(\|(w,\theta)\|^4), \quad \Delta a_2^{\mathrm{IGA}} = 2\lambda_{\mathrm{IGA}} V_c(v_w^2+v_\theta^2). \tag{128}$$

For the canonical squared-loss probe, this statement can be made explicit: substituting Eq. (127) and collecting powers gives the exact identity

$$\Omega_{\mathrm{IGA}} = \lambda_{\mathrm{IGA}} \left[V_c(w^2+\theta^2) - 2C_{cq} w\theta(w^2+\theta^2) + V_q w^2\theta^2(w^2+\theta^2) \right]. \tag{129}$$

The second term is the generally signed joint R_4 correction, and the last is a nonnegative sixth-order correction. Thus IGA is R_2-leading with quartic and sixth-order corrections. These higher-order terms preserve the strict local onset and can deform the finite-amplitude path as λ_{IGA} grows.

Fish/MLDG. The Fish/MLDG implementation in the figure uses cross-environment gradient inner products,

$$\Omega_{\mathrm{Fish/MLDG}} = -\lambda_{\mathrm{Fish}} \overline{\langle \nabla \ell_e, \nabla \ell_f \rangle}_{e \neq f}. \tag{130}$$

Substituting Eq. (124) gives

$$\begin{aligned} \Omega^{(2)}_{\mathrm{Fish/MLDG}} &= -\lambda_{\mathrm{Fish}} \overline{z^\top H_e^\top H_f z}_{e\neq f}, \\ \Omega^{(4)}_{\mathrm{Fish/MLDG}} &= -\lambda_{\mathrm{Fish}} \overline{\left\langle H_e z, \nabla r_f^{(4)}(z) \right\rangle + \left\langle H_f z, \nabla r_e^{(4)}(z) \right\rangle}_{e\neq f}. \end{aligned} \tag{131}$$

Because a scalar quadratic form depends only on the symmetric part of its matrix, define

$$\begin{aligned} G_{\mathrm{Fish}} &:= -2\overline{\mathrm{sym}(H_e^\top H_f)}_{e\neq f}, \qquad \mathrm{sym}(A) = \frac{1}{2}(A+A^\top), \\ \Phi^{(4)}_{\mathrm{Fish}}(z) &:= -\overline{\left\langle H_e z, \nabla r_f^{(4)}(z) \right\rangle + \left\langle H_f z, \nabla r_e^{(4)}(z) \right\rangle}_{e\neq f}. \end{aligned} \tag{132}$$

The low-order Fish/MLDG correction can therefore be written as

$$\Omega_{\mathrm{Fish/MLDG}}(z) = \frac{\lambda_{\mathrm{Fish}}}{2} z^\top G_{\mathrm{Fish}} z + \lambda_{\mathrm{Fish}} \Phi^{(4)}_{\mathrm{Fish}}(z) + O(\|z\|^6). \tag{133}$$

The quadratic and quartic corrections have signs and directions determined by the geometry of cross-environment gradients; both stabilizing and destabilizing contributions are possible.

In the scalar probe, $H_e^\top H_f = c_e c_f I$, so

$$G_{\mathrm{Fish}} = -2\overline{c_e c_f}_{e\neq f} I, \qquad \Delta a_2^{\mathrm{Fish/MLDG}} = -2\lambda_{\mathrm{Fish}} \overline{c_e c_f}_{e\neq f} (v_w^2+v_\theta^2). \tag{134}$$

The correction is then isotropic in the two-dimensional probe, but its sign is still determined by products of task gradients across environments. In a high-dimensional parameter space, projections along different feature directions can have different signs, producing positive-mass suppression in some directions and negative-mass amplification in others.

The canonical scalar probe also makes the first omitted term explicit. Define the ordered-pair averages

$$A_{cc} = \overline{c_e c_f}_{e\neq f}, \qquad A_{cq} = \overline{c_e q_f + q_e c_f}_{e\neq f}, \qquad A_{qq} = \overline{q_e q_f}_{e\neq f}. \tag{135}$$

The complete squared-loss correction is

$$\Omega_{\mathrm{Fish/MLDG}} = -\lambda_{\mathrm{Fish}} A_{cc}(w^2+\theta^2) + \lambda_{\mathrm{Fish}} A_{cq} w\theta(w^2+\theta^2) - \lambda_{\mathrm{Fish}} A_{qq} w^2\theta^2(w^2+\theta^2). \tag{136}$$

Hence Fish/MLDG has an explicit sixth-order correction in addition to its sign-indefinite quadratic and quartic terms. Since $q_e = \mathbb{E}_e[X^2] \geq 0$, $A_{qq} \geq 0$ and this sixth-order contribution is nonpositive for positive λ_{Fish}.

Theorem G.1 (Finite-strength local instability of Fish/MLDG)**.** *Translate any locally stable reference point to the origin, and let the Hessian without the Fish/MLDG correction be $H_{\mathrm{ref}} \succ 0$. If G_{Fish} has a negative-curvature direction, define*

$$\lambda_{\mathrm{inst}} = \min_{\substack{\|v\|=1 \\ v^\top G_{\mathrm{Fish}} v < 0}} \frac{v^\top H_{\mathrm{ref}} v}{-v^\top G_{\mathrm{Fish}} v}. \tag{137}$$

Then $0 < \lambda_{\text{inst}} < \infty$*, and*

$$H_{\text{ref}} + \lambda_{\text{Fish}} G_{\text{Fish}} \begin{cases} \succ 0, & 0 \le \lambda_{\text{Fish}} < \lambda_{\text{inst}}, \\ \textit{has a zero eigenvalue}, & \lambda_{\text{Fish}} = \lambda_{\text{inst}}, \\ \textit{has a negative eigenvalue}, & \lambda_{\text{Fish}} > \lambda_{\text{inst}}. \end{cases} \tag{138}$$

In particular, when $H_{\text{ref}} = \mu I$*,*

$$\lambda_{\text{inst}} = \frac{\mu}{|\lambda_{\min}(G_{\text{Fish}})|}. \tag{139}$$

Proof. Along any unit direction v, the total quadratic curvature is

$$\kappa_2(v; \lambda_{\text{Fish}}) = v^\top H_{\text{ref}} v + \lambda_{\text{Fish}} v^\top G_{\text{Fish}} v. \tag{140}$$

If $v^\top G_{\text{Fish}} v \ge 0$, increasing λ_{Fish} cannot destabilize that direction. If $v^\top G_{\text{Fish}} v < 0$, κ_2 crosses zero exactly at $v^\top H_{\text{ref}} v / [-v^\top G_{\text{Fish}} v]$. Taking the minimum over all negative-curvature directions gives Eq. (137). Because $H_{\text{ref}} \succ 0$ and the set of negative-curvature directions is nonempty, the threshold is strictly positive and finite. The Rayleigh–Ritz criterion then gives Eq. (138). When $H_{\text{ref}} = \mu I$, the first unstable direction is the minimum-eigenvalue eigenvector of G_{Fish}, yielding Eq. (139). □

This theorem also explains how "R_2-like behavior at small regularization, followed by sudden instability" can coexist. If the shortcut direction e_s satisfies $e_s^\top G_{\text{Fish}} e_s > 0$, then its effective mass

$$m_s(\lambda_{\text{Fish}}) = e_s^\top H_{\text{ref}} e_s + \lambda_{\text{Fish}} e_s^\top G_{\text{Fish}} e_s \tag{141}$$

increases with regularization and therefore behaves as a positive R_2 term under this projection. However, if another direction satisfies $v^\top G_{\text{Fish}} v < 0$, the full system still loses local stability at λ_{inst}.

Proposition G.2 (Quartic bifurcation after instability)**.** *Along a unit direction* v*, write the total local free energy as*

$$\mathcal{F}_\lambda(\rho v) - \mathcal{F}_\lambda(0) = \frac{1}{2}\kappa_2(\lambda)\rho^2 + \frac{1}{4}\kappa_4(\lambda)\rho^4 + O(\rho^6). \tag{142}$$

If $\kappa_2(\lambda) < 0$*, then:*

1. *If* $\kappa_4(\lambda) > 0$*, the quartic truncation produces a stable nonzero branch* $\rho_\star^2 = -\kappa_2/\kappa_4$*, corresponding to mode amplification, switching, or a new metastable state;*
2. *If* $\kappa_4(\lambda) \le 0$*, the quartic truncation has no stable nonzero well, and the radial gradient flow escapes outward.*

Proof. Neglecting $O(\rho^6)$, $\partial_\rho \mathcal{F}_\lambda = \rho[\kappa_2 + \kappa_4 \rho^2]$. When $\kappa_2 < 0 < \kappa_4$, in addition to the unstable origin there is a stationary point at $\rho_\star^2 = -\kappa_2/\kappa_4$, whose second derivative is $-2\kappa_2 > 0$. When $\kappa_4 \le 0$, for every $\rho > 0$, $\kappa_2 + \kappa_4 \rho^2 < 0$, and hence $-\partial_\rho \mathcal{F}_\lambda > 0$; the gradient flow of the quartic truncation therefore moves continuously toward larger amplitude. □

Remark G.3 (Theoretical instability versus numerical collapse). Theorem G.1 proves that the local Hessian loses positive definiteness at a finite regularization strength. Positive higher-order terms may restabilize the system farther away, in which case one observes a large mode rearrangement or a nonmonotone path. If the quartic and accessible higher-order terms fail to provide a stable well, optimization rapidly escapes along a negative-curvature direction and appears as parameter explosion or numerical collapse under finite-precision training. In the main experiments, the Fish/MLDG path is truncated at the first observed numerical instability.

Fishr. Standard Fishr matches the coordinate-wise variances of per-example *classifier* gradients. Let

$$g_{e,n}^{\text{cls}} = \nabla_{\theta_{\text{cls}}} \ell_{e,n}, \qquad v_e = \text{Var}_{n \in e}\big(g_{e,n}^{\text{cls}}\big), \tag{143}$$

where the variance is taken coordinate-wise. We study the instantaneous static objective

$$\Omega_{\text{Fishr}} = \frac{\lambda_{\text{Fishr}}}{d_{\text{cls}}} \sum_{j=1}^{d_{\text{cls}}} \text{Var}_e[(v_e)_j]. \tag{144}$$

For a controlled comparison of objective geometries, no EMA, penalty annealing, or warm-up is used. In the scalar bilinear probe, take θ to be the classifier parameter. Then

$$\nabla_\theta \ell_{e,n} = [(\beta X_n - Y_n)X_n]w, \qquad \nu_e(\beta) = \mathrm{Var}_{n\in e}[(\beta X_n - Y_n)X_n]. \tag{145}$$

For squared loss, define the environment-wise coefficients

$$\nu_{0,e} = \mathrm{Var}_{P_e}(YX), \qquad \nu_{1,e} = -2\,\mathrm{Cov}_{P_e}(YX, X^2), \qquad \nu_{2,e} = \mathrm{Var}_{P_e}(X^2), \tag{146}$$

so that $\nu_e(\beta) = \nu_{0,e} + \nu_{1,e}\beta + \nu_{2,e}\beta^2$ exactly. Let $V_{ij}^{\mathrm{Fishr}} = \mathrm{Cov}_e(\nu_{i,e}, \nu_{j,e})$. Hence $v_e = \nu_e(\beta)w^2$ and the exact static scalar penalty is

$$\begin{aligned}\Omega_{\mathrm{Fishr}} = \lambda_{\mathrm{Fishr}}\big[&V_{00}^{\mathrm{Fishr}}w^4 + 2V_{01}^{\mathrm{Fishr}}w^5\theta + (V_{11}^{\mathrm{Fishr}} + 2V_{02}^{\mathrm{Fishr}})w^6\theta^2 \\ &+ 2V_{12}^{\mathrm{Fishr}}w^7\theta^3 + V_{22}^{\mathrm{Fishr}}w^8\theta^4\big].\end{aligned} \tag{147}$$

Thus standard Fishr has no independent R_2 contribution and is locally R_4-leading. Its successive terms have joint orders $4, 6, 8, 10, 12$; only the leading $V_{00}^{\mathrm{Fishr}} \geq 0$ and final $V_{22}^{\mathrm{Fishr}} \geq 0$ coefficients have a fixed sign. Fishr and covariance-only CORAL are both quartic geometry-matching objectives, but Fishr matches classifier-gradient variances whereas CORAL matches representation covariances.

G.5 Mixup and a DANN surrogate

Mixup. For an environment pair (e, f) and mixing coefficient α, define

$$\widetilde{X} = \alpha X_e + (1-\alpha)X_f, \qquad \widetilde{Y} = \alpha Y_e + (1-\alpha)Y_f. \tag{148}$$

Let $\widetilde{c} = \mathbb{E}[\widetilde{X}\widetilde{Y}]$ and $\widetilde{q} = \mathbb{E}[\widetilde{X}^2]$, where the expectation also averages over the sampled (e, f, α). The mixed risk remains exactly

$$\mathcal{F}_{\mathrm{Mixup}} = \widetilde{L}_0 - \widetilde{c}\,w\theta + \frac{\widetilde{q}}{2}w^2\theta^2 + \text{weight decay}. \tag{149}$$

Mixup renormalizes the data moments of the task loss:

$$\Delta a_2^{\mathrm{Mixup}} = -2(\widetilde{c} - \bar{c})v_w v_\theta, \qquad \Delta a_4^{\mathrm{Mixup}} = 2(\widetilde{q} - \bar{q})(v_w v_\theta)^2. \tag{150}$$

DANN. To derive a comparable local expression, first consider a ridge-regularized linear environment head B that predicts the centered environment encoding s:

$$\begin{aligned}L_D(g, B) &= \frac{1}{2}\mathbb{E}\,\|s - BD_gX\|^2 + \frac{\mu_B}{2}\,\|B\|_F^2\,, \\ D_g &= \mathrm{diag}(g), \qquad C_s = \mathbb{E}[Xs^\top], \qquad \Sigma = \mathbb{E}[XX^\top].\end{aligned} \tag{151}$$

Minimizing over B yields the profiled loss

$$L_D^\star(g) = L_D^\star(0) - \frac{1}{2}\,\mathrm{Tr}\big[C_s^\top D_g(D_g\Sigma D_g + \mu_B I)^{-1}D_g C_s\big]\,. \tag{152}$$

The feature-side DANN term is $-\lambda_D L_D^\star(g)$. Expanding the inverse matrix at $g = 0$ gives

$$\begin{aligned}\Omega_{\mathrm{DANN}}^{(2)} &= \frac{\lambda_D}{2\mu_B}\,\mathrm{Tr}(C_s^\top D_g^2 C_s), \\ \Omega_{\mathrm{DANN}}^{(4)} &= -\frac{\lambda_D}{2\mu_B^2}\,\mathrm{Tr}(C_s^\top D_g^2 \Sigma D_g^2 C_s).\end{aligned} \tag{153}$$

The first term is a positive-semidefinite gate mass along environment-predictive directions, whereas the second is a saturating negative quartic correction. For a single feature and scalar environment encoding, $C_s = c_D := \mathbb{E}[Xs]$ and $\Sigma = \bar{q}$ give the exact geometric series

$$\Omega_{\mathrm{DANN}}^{\mathrm{loc}} = \frac{\lambda_D c_D^2}{2\mu_B}w^2 - \frac{\lambda_D c_D^2 \bar{q}}{2\mu_B^2}w^4 + \frac{\lambda_D c_D^2 \bar{q}^2}{2\mu_B^3}w^6 + O(w^8), \tag{154}$$

and hence

$$\Delta a_2^{\mathrm{DANN}} = \lambda_D \frac{c_D^2}{\mu_B}v_w^2, \qquad \Delta a_4^{\mathrm{DANN}} = -2\lambda_D \frac{c_D^2 \bar{q}}{\mu_B^2}v_w^4. \tag{155}$$

The expansion above applies to the ridge-profiled linear MSE surrogate. The experiments use a linear environment classifier trained by alternating cross-entropy updates.

G.6 SUMMARY OF LOCAL OPERATORS

Table 2: Effective local expansions at the controlled bilinear reference state. Here $z = (w, \theta)$, $\beta = w\theta$, and v denotes a measurement direction. Terms denoted $O(\|z\|^k)$ are the remaining higher-order terms; every "quartic" statement refers to order in the joint variables (w, θ).

Method in figure	Effective correction relative to ERM	Local order, sign, and conditions
ERM	None: $-\bar{c}\,\beta + \frac{\bar{q}}{2}\beta^2$	Reference mixed task curvature and quartic data term; see Eq. (106).
GroupDRO	$-\Delta c_p \beta + \frac{1}{2}\Delta q_p \beta^2$	Reweighted R_2/R_4 task terms at fixed weights; residual drive can sustain a nonzero plateau.
IRMv1	$\lambda_{\mathrm{IRM}}[\overline{c_e^2}\beta^2 - 2\overline{c_e q_e}\beta^3 + \overline{q_e^2}\beta^4]$	R_4-leading, with explicit sixth- and eighth-order terms; no R_2 contribution.
V-REx	$-2\,\mathrm{Cov}(L_{0,e}, c_e)\beta + [\mathrm{Var}(c_e) + \mathrm{Cov}(L_{0,e}, q_e)]\beta^2 - \mathrm{Cov}(c_e, q_e)\beta^3 + \frac{1}{4}\mathrm{Var}(q_e)\beta^4$	R_2 under unequal null risks; R_4-leading under equal null risks, with sixth- and eighth-order terms.
IGA	$\lambda z^\top \overline{\Delta H_e^\top \Delta H_e} z + 2\lambda\langle \Delta H_e z, \nabla \Delta r_e^{(4)}(z)\rangle + O(\|z\|^6)$	PSD matrix-R_2 plus generally signed R_4 and nonnegative R_6 gradient couplings.
Fish/MLDG	$\frac{\lambda}{2} z^\top G_{\mathrm{Fish}} z + \lambda \Phi^{(4)}_{\mathrm{Fish}}(z) + O(\|z\|^6)$	Quadratic and quartic corrections are sign-indefinite; a negative direction of G_{Fish} causes finite-strength local instability; see Theorem G.1.
Fishr	$\lambda_{\mathrm{Fishr}}\,\mathrm{Var}_e[\nu_e(0)]w^4 + O(\|z\|^6)$	Classifier-gradient variance matching; R_4-leading with state-dependent finite-amplitude higher-order corrections.
Mixup	$-(\widetilde{c} - \bar{c})\beta + \frac{1}{2}(\widetilde{q} - \bar{q})\beta^2$	Renormalizes the task curvature and quartic data term; no fixed additive operator.
MMD-mean	$\lambda_\mu M_2 w^2$	Gate-block R_2; see Eq. (109).
Gaussian MMD	$\lambda_G(A_2 w^2 + A_4 w^4 + \cdots)$	Determined by moment differences; may begin at R_2, R_4, or higher order.
CORAL-cov	$\lambda_\Sigma S_2 w^4$	Quartic gate-covariance geometry.
CORAL-full	$\lambda_\mu M_2 w^2 + \lambda_\Sigma S_2 w^4$	Contains both the MMD-mean R_2 and the CORAL-cov R_4.
DANN	$+\,\mathrm{Tr}(C_s^\top D_g^2 C_s) - O(g^4)$	The profiled linear MSE surrogate yields a PSD gate R_2 and a saturating negative R_4; experiments use an alternating linear cross-entropy classifier.
R_2	$\frac{1}{2}\lambda_2 \gamma w^2$	Explicit gate R_2.
R_4	$\frac{1}{2}\lambda_4 \gamma (w\theta)^2$	Explicit joint R_4.
$R_2 + R_4$	$\frac{1}{2}\lambda_2 \gamma w^2 + \frac{1}{2}\lambda_4 \gamma (w\theta)^2$	R_2 determines the exact onset; R_4 determines the post-onset amplitude.

H SUPPLEMENTARY RESULTS

H.1 FINITE-SAMPLE OPERATOR ALIGNMENT

The MMD-mean and CORAL-full objectives in Figure 5 are evaluated from empirical environment moments. Let $\widehat{m}_e = n^{-1}\sum_{i\in e} X_i$ and define

$$\widehat{\Gamma}_\mu = \frac{2}{\binom{E}{2}} \sum_{e<f} \mathrm{diag}\big((\widehat{m}_e - \widehat{m}_f)^{\odot 2}\big)\,.$$

For the gate vector g, the MMD-mean penalty is $\Omega_{\mathrm{MMD}} = \frac{\lambda}{2} g^\top \widehat{\Gamma}_\mu g$. CORAL-full has this identical leading R_2 operator and adds a covariance term that begins at R_4 order. Although the population stable-coordinate mean is invariant by construction, its empirical counterpart fluctuates across finite training environments. With $\gamma_c = e_c^\top \widehat{\Gamma}_\mu e_c$ and $\gamma_s = e_s^\top \widehat{\Gamma}_\mu e_s$, Table 3 audits this projection before training on the five Figure 5 datasets at each sample size.

The stable projection is small but nonzero. At $\lambda = 100$, its mean contribution is $\lambda\bar{\gamma}_c \simeq 2.86$ for $n = 256$, compared with the gate weight-decay coefficient $\mu = 0.2$. For $n = 8192$, the corresponding contribution is only 0.0527, consistent with the weaker collateral suppression in Fig. 5(b).

Table 3: Leading empirical R_2 signatures in the finite-sample Figure 5 objective. Values are mean $\pm$ s.e.m. over five pre-training datasets at each sample size n per environment. γ_c and γ_s are the stable- and shortcut-mode gate masses.

Objective	n	$\bar{\gamma}_c$	$\bar{\gamma}_s$	$\overline{\gamma_c/\gamma_s}$
oracle R_2	both	0	1	0
MMD-mean	256	0.0286 ± 0.0078	5.092 ± 0.212	$(5.73 \pm 1.57) \times 10^{-3}$
CORAL-full	256	0.0286 ± 0.0078	5.092 ± 0.212	$(5.73 \pm 1.57) \times 10^{-3}$
MMD-mean	8192	$(5.27 \pm 0.65) \times 10^{-4}$	5.342 ± 0.021	$(9.88 \pm 1.24) \times 10^{-5}$
CORAL-full	8192	$(5.27 \pm 0.65) \times 10^{-4}$	5.342 ± 0.021	$(9.88 \pm 1.24) \times 10^{-5}$

H.2 Extended nonlinear regularization paths

Figures 13 and 14 extend the same runs shown in Fig. 6 to larger control strengths, and Fig. 15 adds a three-hidden-layer network. All three architectures use exactly the same training and IID/OOD test sets as Fig. 5(b), with 8192 training samples per environment. The three-hidden-layer network has widths $(32, 32, 32)$ and uses the two-hidden-layer optimization settings; loading definitions are unchanged. Curves and shaded bands report means and standard errors over five seeds. Each method uses the same displayed endpoint across all four metrics.

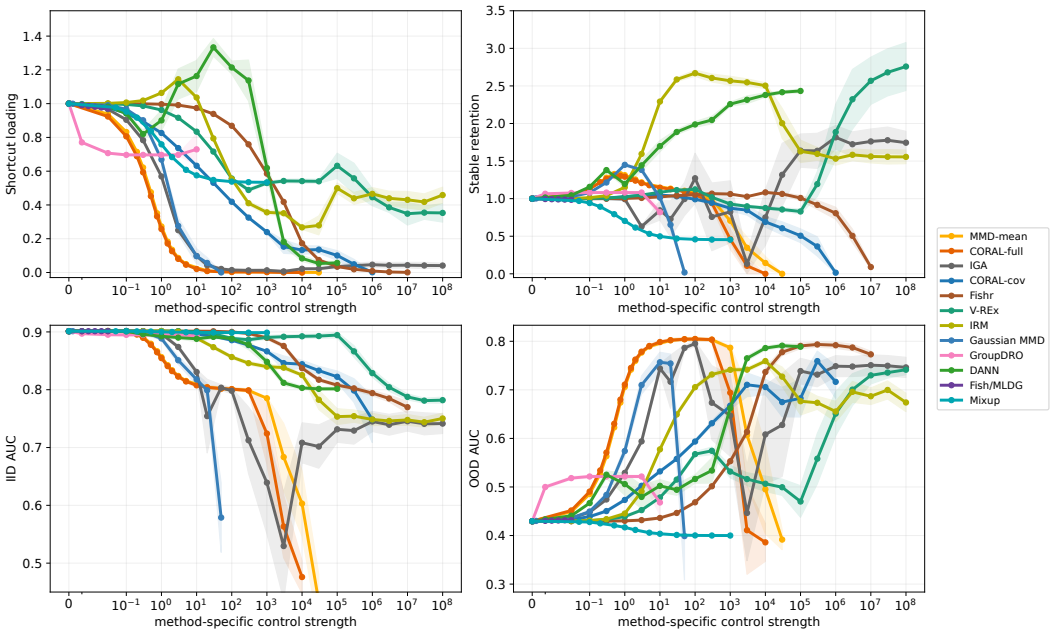


Figure 13: Extended regularization paths for the one-hidden-layer ReLU network with 8192 training samples per environment.

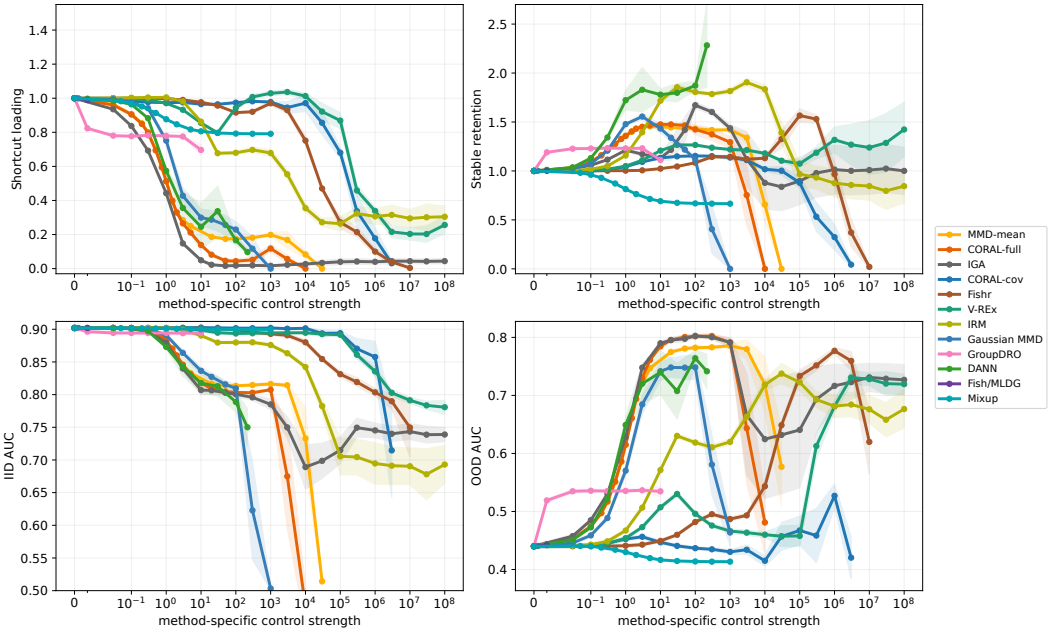


Figure 14: Extended regularization paths for the two-hidden-layer ReLU network with 8192 training samples per environment.

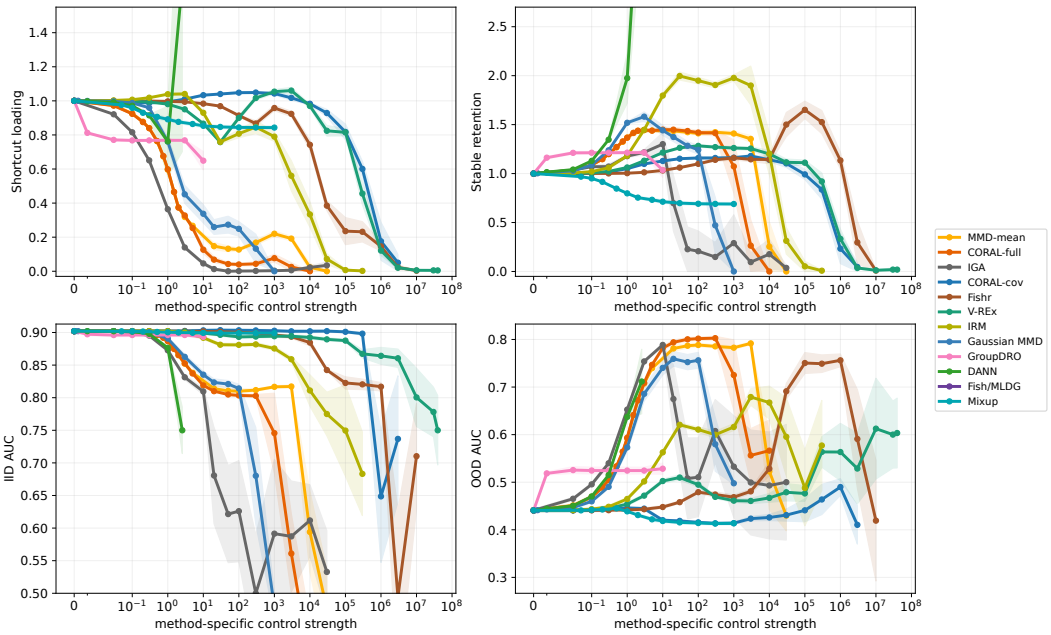


Figure 15: Extended regularization paths for the three-hidden-layer ReLU network with 8192 training samples per environment, using the same data as Figs. 13 and 14.

H.3 Regularization paths on MNIST

Prior invariant-learning studies have used MNIST constructions with label-correlated corner patches (Chen et al., 2022) or environment-specific white squares (Norman & Meir, 2026). We adapt these localized cues to a binary bright marker with controlled environment-dependent prevalence and label association.

We examine whether the qualitative signature–phenotype relation in Figure 6 remains informative when stable features are learned from MNIST images. We evaluate thirteen established methods under two controlled environment allocations.

Task and shortcut construction. Following the binary digit task of Arjovsky et al. (2019), the clean label is $Y_c = 2\mathbf{1}\{d < 5\} - 1$ for digit $d \in \{0, \dots, 9\}$; the task label $Y \in \{-1, +1\}$ flips Y_c with rate 0.25, enforced by rounded counts within each digit. The stable signal is therefore the grayscale digit shape. The shortcut is a bright corner marker: when $B = 1$, initially zero-valued pixels within an upper-left 8×8 patch are set to unit brightness, whereas $B = 0$ leaves the background unchanged. This intervention never overwrites nonzero digit pixels. With $S = 2B - 1$ and training environment index $\zeta_e \in \{-3, -1, 1, 3\}/\sqrt{5}$, the label-conditional marker probabilities are chosen to realize the target moments

$$\mathbb{E}[S \mid e] = \delta\zeta_e, \qquad c_e := \mathbb{E}[SY \mid e] = a_s + \Delta_a\zeta_e, \qquad (a_s, \delta, \Delta_a) = (0.70, 0.10, 0.03). \quad (156)$$

Equivalently, $P(B = 1 \mid Y = y, e) = (1 + \delta\zeta_e + yc_e)/2$ for balanced Y. Conditional marker counts are rounded within the two task-label classes. This separates the marker's marginal mean shift from its label-correlation heterogeneity: $\mathrm{Var}_e(\mathbb{E}[S \mid e]) = 0.01$, whereas $\mathrm{Var}_e(c_e) = 0.0009$. The nominal marker-only AUC ranges from 0.830 to 0.870 across training environments, averaging 0.85. OOD evaluation uses held-out MNIST test images in two environments with $\zeta_e = \pm 2$, $\mathbb{E}[S \mid e] = 0.10\zeta_e$, and $\mathbb{E}[SY \mid e] = -0.70$, reversing the shortcut association (marker-only AUC 0.15). IID AUC is measured on a stratified 20% holdout from each training environment; OOD AUC is averaged over the two test environments. Both use the noisy task label.

Two environment allocations. In the *replicated-digit* setting, every environment contains the same digit images and the same task-label flip mask. Each has $43{,}370$ training and $10{,}840$ IID-holdout examples, with respectively $4{,}337$ and $1{,}084$ examples per digit. There are $1{,}084$ flipped training labels and 271 flipped IID labels per digit in every environment. A shared holdout assignment keeps the training and IID image sets disjoint across environments. Only the marker assignment differs.

In the *matched-disjoint* setting, the environments receive four distinct images from each same-digit matched quartet. Matching uses six principal components of 7×7 average-pooled image descriptors and a Morton ordering; within-quartet assignment approximately balances the coarse image means across environments. Each quartet shares its label-flip status. The resulting $54{,}200$ source images are unique across environments. Each environment has $10{,}840$ training and $2{,}710$ IID examples: respectively $1{,}084$ and 271 per digit, with 271 and 68 flipped labels per digit. Digit counts, task-label counts, and the positions of flipped examples within the matched sequence agree exactly; grayscale images and their higher-order statistics may still differ. Before the IID split, the maximum pairwise mean-image RMSE is 1.54×10^{-3} on the $[0, 1]$ scale. Both settings use synchronized minibatch indices and no environment-specific augmentation. The comparison jointly varies residual stable-feature mismatch and per-environment sample size.

Replicated-digit results. In the replicated-digit run, Fig. 16 is consistent with a broad selective-retention window for several objectives when stable-image and label variation across environments is removed. At $\lambda = 10^6$, MMD-mean yields the across-seed mean triplet $(Q_s, \text{Clean-digit AUC}, \text{OOD AUC}) = (0.0098 \pm 0.0011, 0.981 \pm 0.002, 0.734 \pm 0.002)$, compared with ERM's $(1, 0.977 \pm 0.001, 0.215 \pm 0.001)$, where uncertainties denote standard errors over five seeds. CORAL-full at 10^6 gives $(0.010 \pm 0.002, 0.980 \pm 0.003, 0.731 \pm 0.004)$. V-REx at 3×10^4 and Fishr at 3×10^7 give $(0.006 \pm 0.002, 0.974 \pm 0.001, 0.735 \pm 0.002)$ and $(0.008 \pm 0.002, 0.978 \pm 0.001, 0.735 \pm 0.003)$, respectively. Across these four method–strength pairs, every seed has $Q_s < 0.016$ and Clean-digit AUC above 0.967. These sampled path points approach the noisy-label digit-only performance level at method-specific transition scales. At the

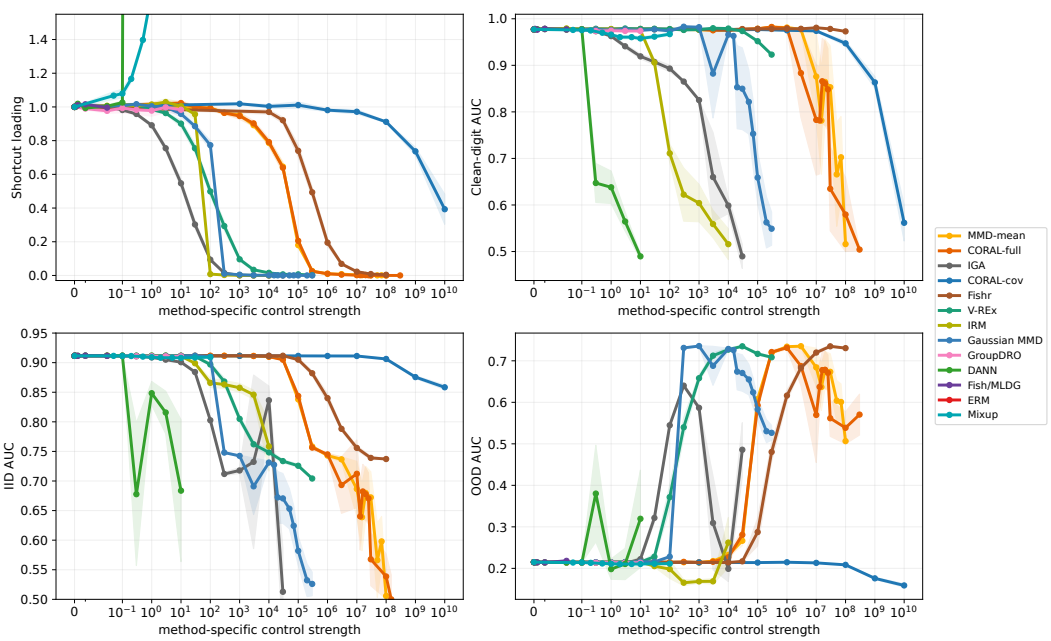


Figure 16: MNIST regularization paths with identical digit images and label-flip masks across four training environments. Curves and shaded bands show means and standard errors over five training seeds. Several objectives exhibit a selective-retention window in which marker response decreases while Clean-digit AUC remains near its ERM value.

larger sampled strengths, the mean-matching and Gaussian MMD paths lose stable digit information or collapse; IGA and IRM show earlier damage in this run, while covariance-only CORAL retains substantial shortcut loading over much of its scan.

Matched-disjoint results. The matched-disjoint run in Fig. 17 shows greater seed variation and more apparent collateral damage despite exact matching of digit and flip counts. At $\lambda = 10^3$, V-REx gives the across-seed mean triplet $(Q_s, \text{Clean-digit AUC}, \text{OOD AUC}) = (0.088 \pm 0.008, 0.961 \pm 0.003, 0.545 \pm 0.025)$, compared with ERM's $(1, 0.977 \pm 0.001, 0.217 \pm 0.002)$. Across the five V-REx runs, Q_s ranges from 0.055 to 0.101 and Clean-digit AUC from 0.953 to 0.966, whereas OOD AUC ranges from 0.505 to 0.637. Fishr at 10^6 gives $(0.150 \pm 0.014, 0.967 \pm 0.001, 0.552 \pm 0.017)$; at 10^7, the mean marker response decreases further, giving $(0.013 \pm 0.002, 0.914 \pm 0.011, 0.543 \pm 0.028)$, with clear digit-feature loss. MMD-mean at 4×10^4 and CORAL-full at 5×10^4 have still larger seed variation, including runs with low Clean-digit AUC. On the sampled grid, shortcut attenuation and stable-feature loss overlap more closely than in the replicated setting. ERM, Mixup, and GroupDRO retain substantial marker response over their sampled ranges in this run. The sampled covariance-only, adversarial, and gradient-alignment paths also do not display a uniformly high-retention suppression window. In particular, a high OOD AUC accompanied by large marker response and near-chance Clean-digit AUC is not counted as stable-feature recovery.

Model, objectives, and scan protocol. The predictor uses the width and spatial-subsampling convention of the two-hidden-layer ColoredMNIST MLP in Arjovsky et al. (2019): the rendered luminance image is subsampled from 28×28 to 14×14, passed through two width-390 ReLU layers, and mapped to a single logit. Every path point starts afresh and runs 2,000 float32 Adam updates with learning rate $\simeq 2.82 \times 10^{-3}$, an explicit parameter-squared penalty with coefficient $\simeq 8.60 \times 10^{-4}$, gradient-norm clipping at 10, and 128 matched samples per environment per update. The task loss is binary cross-entropy (BCE). There is no penalty annealing, learning-rate schedule, or early stopping. The dataset construction is fixed at seed 0; model initialization and minibatch sampling use training seeds $\{0, 1, 2, 3, 4\}$. Curves report the across-training-seed mean and standard error.

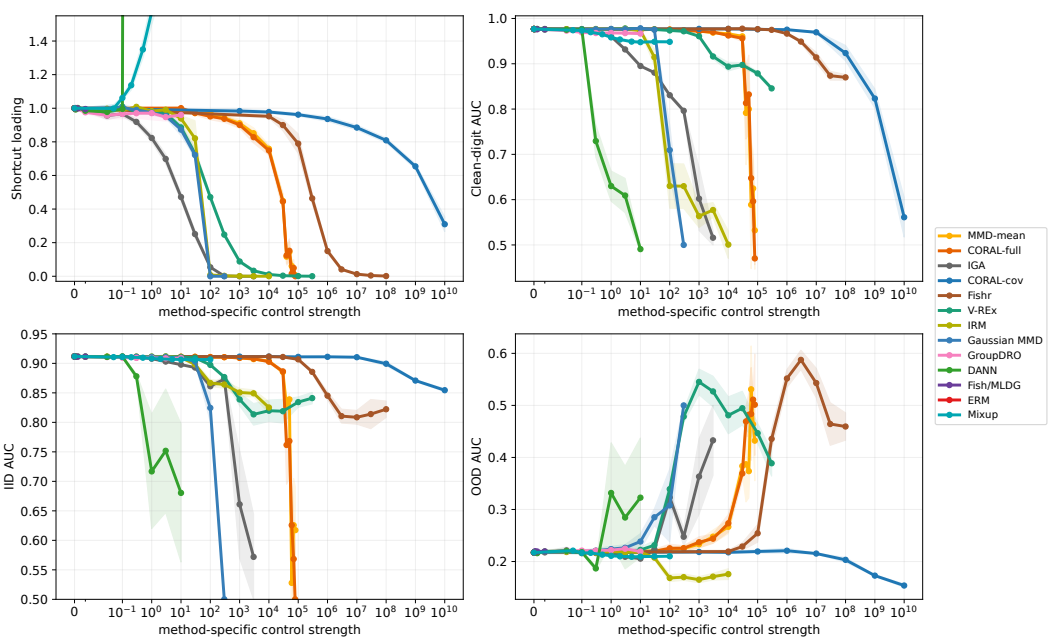


Figure 17: MNIST regularization paths with matched but non-overlapping digit images across four environments. Digit and label counts, label-flip positions, and minibatch indices are aligned. Curves and shaded bands show means and standard errors over five training seeds. Shortcut attenuation and digit-feature loss overlap more closely than in the replicated setting.

The objectives follow Appendix I.2, with BCE risks on minibatches and the following normalization details. Mean- and covariance-matching terms divide squared discrepancies by 390 and 390^2, respectively, following the elementwise-mean convention of DomainBed (Gulrajani & Lopez-Paz, 2021); CORAL-full sums the mean and covariance terms. Gaussian MMD averages five RBF bandwidths $\{0.1, 0.5, 1, 2, 10\}$ using the biased V-statistic. Fishr matches instantaneous diagonal variances of per-example classifier gradients, including the bias, without EMA. Fish/MLDG uses the same full-parameter gradient-inner-product surrogate as in Figure 6. DANN alternates one linear-domain-head Adam update (learning rate 0.01, weight decay 0.001) with each predictor update. The horizontal axis is the group-weight update rate for GroupDRO, the Beta concentration for Mixup, and the penalty/adversarial coefficient for the remaining non-ERM methods. Control strengths therefore have method-specific scales.

Within each seed, all methods share one zero-strength baseline. MMD-mean, CORAL-full, and Gaussian MMD curves end at the first point with mean Clean-digit AUC ≤ 0.55; this rule also applies to IGA and IRMv1 in the matched-disjoint setting. Fish/MLDG curves end at the first observed loading instability.

Measurements. For a fixed test digit image x, let $f_\lambda(x, B)$ be the logit with marker state B. Define the marker-marginalized score and counterfactual response by

$$\bar{f}_\lambda(x) = \tfrac{1}{2}\big[f_\lambda(x,0) + f_\lambda(x,1)\big], \qquad Q_s(\lambda) = \frac{\mathbb{E}_x|f_\lambda(x,1) - f_\lambda(x,0)|}{\mathbb{E}_x|f_0(x,1) - f_0(x,0)|}. \tag{157}$$

The panels show shortcut loading Q_s, *Clean-digit AUC* of $\bar{f}_\lambda$ against the unflipped digit label, and factual IID/OOD AUC against the noisy task label. Clean-digit AUC is the task-based stable-retention diagnostic. Selective retention combines low marker response with preserved Clean-digit AUC.

Interpretation. We compare the MNIST paths qualitatively with the local signatures in Appendix G. These signatures depend on the reference state; in particular, the R_4-leading V-REx expression in Eq. (119) assumes equal null risks across environments. The matched-disjoint V-REx

tail illustrates that small but nonzero shortcut loading can still alter OOD rankings when stable-score differences also shrink, even while Clean-digit AUC remains relatively high.

H.4 COLLECTIVE-MODE BASELINE COMPARISON

Figure 18 reports the full baseline comparison for the collective-mode experiment.

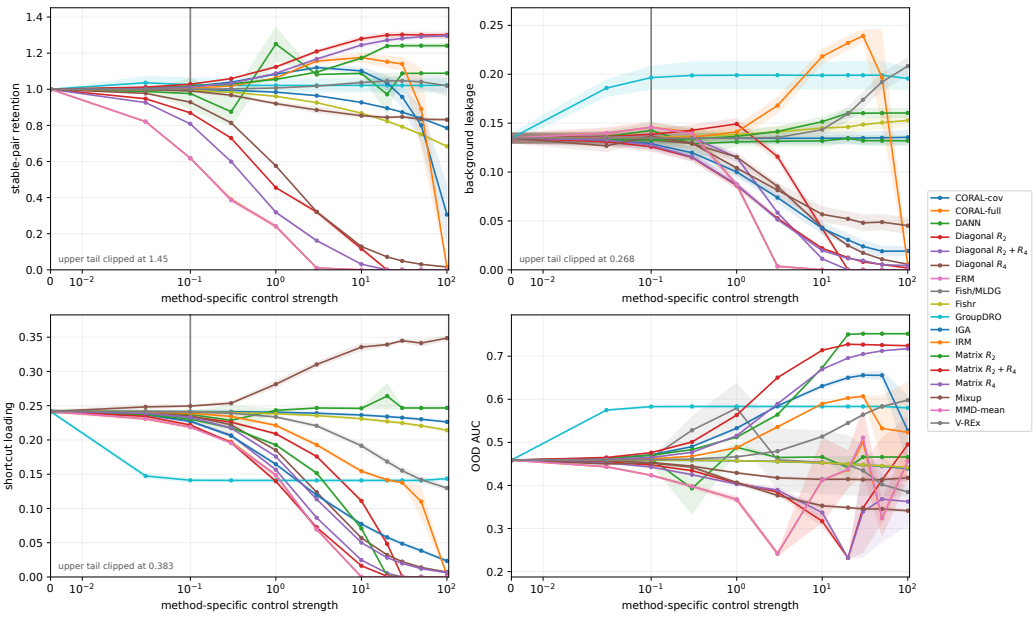


Figure 18: Supplementary full algorithm comparison for the collective-mode experiment (Fig. 7 in the main text). The four panels report retention of the compensating pair, background leakage, shortcut loading, and OOD AUC, respectively. In this construction, the broader set of sampled paths does not uniformly combine preservation of the compensating features, removal of leakage, and improved OOD performance.

I EXPERIMENTAL CONFIGURATIONS AND REPRODUCIBILITY

Unless stated otherwise, all stochastic training curves report the mean and standard error over five random seeds. Figure 4 uses the pre-specified, theory-blind two-stage empirical calibration protocol detailed below. Settings for the response, objective-specific relaxation, and MNIST experiments are given in Appendices E.3, E.4, and H.3, respectively.

I.1 TRAINING PROTOCOL

Table 4 summarizes the common training protocol for the main-text synthetic experiments. Except for the deterministic population audit in Figure 2, every synthetic update computes the environment risk on all samples from each of the four training environments and then averages across environments; no mini-batch sampling is used. Synthetic stochastic curves use seeds $\{0, 1, 2, 3, 4\}$ and report mean $\pm$ standard error. No penalty annealing, learning-rate schedule, optimizer reset, or early stopping is used.

Table 4: Main-text synthetic training protocol. Sample counts are per environment; the listed experiments except Figure 2 use Adam with its default momentum parameters.

Figure	Model	Samples and batching	Optimization settings
Figure 2	Single-shortcut population objective	No finite-sample batches	Gradient descent for 50 steps with step size 0.03; initialized from a perturbation of radius 10^{-3}.
Figure 3	Bilinear predictor	256 training samples in (a), 1024 in (b); full-batch over four environments	Adam for 2000 steps with learning rate 0.03 and ℓ_2 coefficient $\mu = 0.2$.
Figure 4	Single-feature bilinear predictor	512 training samples; full-batch over four environments; no test set is used	Adam for 2500 steps with learning rate 0.03, gradient clipping at 20, and $\mu \in \{0.15, 0.2\}$.
Figures 5–6	Bilinear predictor; or plain ReLU MLP with one hidden layer (32) / two hidden layers (32,32)	256/8192 training samples in Figure 5(a)/(b); 8192 in Figure 6; 2048 test samples; full-batch over four environments	3000 steps with gradient clipping at 10; learning rate 0.01 and $\mu = 0.2$ for the bilinear and one-hidden-layer models, and learning rate 0.005 and $\mu = 0.05$ for the two-hidden-layer model.
Figure 7	Three-feature bilinear collective-mode predictor	1024 training samples and 4096 samples for each of IID/OOD testing; full-batch over four environments	Adam for 3000 steps; learning rate 0.03 for both predictor and DANN domain head, gradient clipping at 10, $\mu = 0.02$ (domain-head $\mu_B = 10^{-3}$).

I.2 Empirical objectives and implementation choices

This section specifies the objectives optimized in the finite-sample synthetic experiments. Let $R_e = \frac{1}{2n_e}\sum_i (y_{ei} - f_\phi(x_{ei}))^2$ denote the full-batch MSE in environment e, let $\bar{R} = E^{-1}\sum_e R_e$, let h_{ei} be the representation used for alignment or domain discrimination, and let $g_e = \nabla_\phi R_e$. Except for the specialized updates of Mixup, GroupDRO, and DANN, all methods minimize $\bar{R} + \mathcal{R}_{\ell_2} + \lambda\Omega$. Every finite-sample synthetic experiment uses float64 full-batch training, five random seeds, and the default momentum parameters of PyTorch Adam. For the bilinear model, $h = g \odot x$; for the plain ReLU MLP, h is the final hidden layer.

Table 5: Empirical objectives and implementation choices for the finite-sample regularization-path experiments. Here C_e is the sample covariance of representations in environment e. Every pairwise average is normalized as $\binom{E}{2}^{-1}\sum_{e<f}$.

Method	Objective or additional term used in training	Implementation choice
ERM; explicit R_2 / R_4	ERM uses $\bar{R} + \mathcal{R}_{\ell_2}$. For the bilinear parameterization $\beta = g \odot \theta$, explicit R_2 adds $\frac{\lambda}{2} g^\top \Gamma g$, explicit R_4 adds $\frac{\lambda}{2}\beta^\top\Gamma\beta$, and the joint objective uses their sum.	Scalar experiments set $\Gamma = \gamma$; the matrix version in Figure 7 uses the known oracle Γ_E, while the diagonal baseline uses $\mathrm{diag}(\Gamma_E)$.
IRMv1	$\lambda E^{-1}\sum_e [\partial_\alpha R_e(\alpha f_\phi)\|_{\alpha=1}]^2$.	Uses the autograd derivative at the shared scalar scaling $\alpha = 1$ and averages across environments.
V-REx	$\lambda E^{-1}\sum_e (R_e - \bar{R})^2$.	Uses the population-variance convention over the finite set of environments.
GroupDRO	Optimizes $\sum_e p_e R_e + \mathcal{R}_{\ell_2}$, with $p_e \leftarrow p_e \exp(\eta R_e)/\sum_f p_f \exp(\eta R_f)$.	Initializes p_e uniformly; the scanned control is the group-weight update step size η.
IGA	$\lambda E^{-1}\sum_e \|g_e - \bar{g}\|_2^2$.	Computes environment-risk gradients with respect to all trainable predictor parameters.

Method	Objective or additional term used in training	Implementation choice
Fish/MLDG	$-\lambda\binom{E}{2}^{-1}\sum_{e<f} g_e^\top g_f$.	Uses a cross-environment gradient-alignment surrogate with full-parameter gradient inner products.
Fishr	$\lambda d_{\text{cls}}^{-1}\sum_{j=1}^{d_{\text{cls}}} \text{Var}_e[(v_e)_j]$.	Matches the instantaneous coordinate-wise variances of per-example classifier gradients; the penalty is averaged over classifier coordinates. No EMA or penalty annealing is used.
MMD-mean; Gaussian MMD	The former adds $\lambda\|\bar{h}_e - \bar{h}_f\|_2^2$; the latter adds $\lambda\text{MMD}^2(h_e, h_f)$.	Gaussian MMD averages RBF kernels with bandwidths $\{0.1, 0.5, 1, 2, 10\}$, uses a biased V-statistic including diagonal entries, and uses at most the first 128 samples from each environment.
CORAL-full; CORAL-cov	CORAL-full adds $\lambda\|\bar{h}_e - \bar{h}_f\|_2^2 + \|\mathsf{C}_e - \mathsf{C}_f\|_F^2$; CORAL-cov retains only the covariance term.	CORAL-full uses the DomainBed mean-and-covariance implementation with one shared λ; CORAL-cov is the covariance-only component corresponding to the original Deep CORAL loss. Covariances use the sample denominator $n_e - 1$.
Mixup	Replaces ERM with the mean MSE on $\alpha x_e + (1-\alpha)x_{e+1}$ and $\alpha y_e + (1-\alpha)y_{e+1}$.	Uses $\alpha \sim \text{Beta}(\lambda, \lambda)$ and cyclically pairs adjacent environments; $\lambda = 0$ is defined as the ERM baseline.
DANN	The predictor minimizes $\bar{R} + \mathcal{R}_{\ell_2} - \lambda\ \text{CE}(D_\psi(h), e)$.	D_ψ is a linear environment-classification head; before each predictor update, the domain head receives one full-batch Adam update on cross-entropy.

Table 5 describes the finite-sample training objectives, and Appendix G gives their local R_2/R_4 expansions in the controlled bilinear population setting. The DANN derivation uses a ridge-profiled linear MSE surrogate; the DANN curves in Figures 5–6 use the alternating cross-entropy domain classifier listed above.

I.3 Figure-specific configurations

I.3.1 Local onset taxonomy (Figure 2)

This figure computes the Hessian at the origin directly for a single-shortcut population objective and runs gradient descent for 50 steps from a perturbation of radius 10^{-3} using step size 0.03; the early-growth rate is fitted from the first 20 steps, and its zero crossing in a_s defines the reported dynamic onset. GroupDRO starts from uniform environment weights and updates them with step size $\eta = 1$ before each predictor step. Equal null risks make the initial weight changes quadratic in (w, θ), preserving the ERM onset. The four training environments are $\zeta_e = (-3, -1, 1, 3)/\sqrt{5}$. The first two panels scan $b_F = \delta \in \{0, 0.5, 1, 1.5, 2\}$ over $a_s \in [0, 2]$ with step size 0.02, with shared parameters $\mu_w = \mu_\theta = 0.2$, $\sigma = 1$, and $\Delta_a = 0.5$. The IGA panel instead fixes $b_F = 0$ and scans $\Delta_a \in \{0, 0.1, \ldots, 0.8\}$ using the same onset diagnostic.

I.3.2 Scalar loading validation (Figure 3)

Panel (a) uses $X_F = a_F Y + b_F \zeta_e + \sigma_F \epsilon$, where $Y \in \{-1, +1\}$ is balanced and $\zeta_e = (-3, -1, 1, 3)/\sqrt{5}$. It sets $a_F = 0.6$, $\sigma_F = 1$, and scans $b_F \in \{0, 0.2, \ldots, 1.0, 1.1, 1.2, 1.25, 1.3, 1.4, 1.6, 1.8, 2.0\}$; each environment has 256 training samples, and the bilinear model uses Adam (learning rate 0.03) for 2000 steps with weight decay $\mu = 0.2$. Panel (b) uses the matched probe $X = (0.6 + \Delta_a \zeta_e)Y + \epsilon$, fixes $b_F = 0$, and scans $\Delta_a \in \{0, 0.1, \ldots, 0.8\}$ with $\lambda_{\text{IGA}} = 1$ and

$\lambda_{\text{V-REx}} = 5$. It uses 1024 training samples per environment and otherwise the same optimizer and number of steps.

I.3.3 THEORY-BLIND CRITICAL-STRENGTH CALIBRATION (FIGURE 4)

The calibration uses the same four-environment scalar construction $X_F = a_F Y + b_F \zeta_e + \epsilon$ with $\zeta_e = (-3, -1, 1, 3)/\sqrt{5}$ for explicit R_2, R_2+R_4, MMD-mean, and CORAL-full. We enumerate $a_F \in \{0.5, 0.6\}$, $\mu \in \{0.15, 0.2\}$, and $b_F \in \{0.5, 0.7, 0.9\}$, yielding 12 settings per method. Matched IGA instead uses $X_F = (0.5 + \Delta_a \zeta_e)Y + \epsilon$ with $(\mu, \Delta_a) = (0.15, 0.6), (0.15, 0.65), (0.15, 0.7), (0.15, 0.75)$, so the calibration contains 52 settings in total. For every setting, we first train a fixed coarse grid $\{0, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1, 1.25, 1.5, 2, 2.5, 3, 4, 5, 6, 8, 10, 12\}$, independently of $\lambda_c^{\text{theory}}$. Let $(\bar{b}_{\text{floor}}, s_{\text{floor}})$ be the mean and s.e.m. of the final loading at the three largest coarse strengths. At strength λ, let $(\bar{b}_\lambda, s_\lambda)$ denote the corresponding mean and s.e.m. of final $|\beta|$. We call this point *resolved ordered* iff

$$\bar{b}_\lambda - 1.96 s_\lambda > \bar{b}_{\text{floor}} + 1.96 s_{\text{floor}}.$$

Thus $\bar{x} \pm 1.96\,\text{s.e.m.}$ is used as the usual approximate two-sided 95% normal confidence interval for a mean. The first persistent non-resolved tail and the preceding resolved point form the empirical bracket $[\lambda_{\text{lo}}, \lambda_{\text{hi}}]$. Settings without such a bracket are discarded.

We next use a uniform 9-point refinement as an initial empirical screen within each bracket (21 points for IGA). A setting is retained only when this scan contains at least five *resolved ordered* strengths; its five largest such strengths, i.e., those closest to the transition from the ordered side, define a common local fitting window. To quantify finite-seed variation, we enumerate all $5^5 = 3125$ ordered five-seed bootstrap draws with replacement. For each draw, we average the loading at the five fixed strengths and fit its zero by ordinary least squares. Figure 4 reports the mean and the 2.5th–97.5th percentile bootstrap interval of these 3125 critical-strength estimates. The bracket and fitting points are selected from empirical loading data. The aggregate weighted local fit is used for quality control: we reject a nonnegative slope, $R^2 < 0.95$, or an extrapolated root more than half a bracket width outside the empirical bracket. The one-sided confidence criterion restricts the fit to resolved ordered points.

Of the 52 settings, 24 were excluded for having fewer than five *resolved ordered* strengths: four each for R_2 and R_2+R_4, and eight each for MMD-mean and CORAL-full. The remaining 28 passed all quality checks. Figure 4 displays four settings per method (20 in total), selected using fixed theoretical critical-strength targets.

I.3.4 REGULARIZATION PATHS AND NONLINEAR TRANSFER (FIGURES 5–6)

The bilinear and ReLU experiments share a four-environment stable–shortcut construction: $X_c = a_c Y + \sigma_c \epsilon_c$, $X_s = (a_s + \Delta_a \zeta_e) Y + \delta \zeta_e + \sigma_s \epsilon_s$, together with four independent noise features. We use $(a_c, a_s, \delta, \Delta_a, \sigma_c, \sigma_s) = (0.6, 0.8, 1, 0.3, 1, 0.5)$. Figure 5 compares 256 and 8192 training samples per environment with shared IID/OOD test sets within each seed. Figure 6 and Appendix H.2 use the same 8192-sample training sets and IID/OOD test sets as Figure 5(b), for all five seeds and all methods. Each test environment contains 2048 samples. OOD testing uses $\zeta_e = \pm 2$ and reverses the shortcut label coefficient, $a_s \mapsto -a_s$. The bilinear model is trained with Adam for 3000 steps at learning rate 0.01. Nonlinear validation uses plain ReLU MLPs on the raw inputs: either one hidden layer of width 32 or two hidden layers of widths $(32, 32)$, trained for 3000 steps at learning rates 0.01 and 0.005, respectively. Functional loading is the mean absolute Jacobian of the prediction with respect to each input coordinate on OOD samples, normalized by the corresponding zero-regularization value for the same method.

I.3.5 COLLECTIVE-MODE EXPERIMENT (FIGURE 7)

The compensating-pair construction uses two features with opposite environmental backgrounds to test whether the stable signal is a collective direction. Figure 7 and its supplementary diagnostics use four composite training environments. Both $\zeta_{\text{pair}} = (-3, -1, 1, 3)/\sqrt{5}$ and $\zeta_{\text{short}} = (1, -3, 3, -1)/\sqrt{5}$ have zero mean and unit second moment, and they are mutually orthogonal.

With independent noise $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$, the training and IID-test features are

$$X_1 = (a_c + \Delta_a \zeta_{\text{pair}})Y + \delta_c \zeta_{\text{pair}} + \epsilon_1, \qquad X_2 = -\delta_c \zeta_{\text{pair}} + \epsilon_2,$$
$$X_3 = (a_s + \Delta_{a,s} \zeta_{\text{short}})Y + \delta_s \zeta_{\text{short}} + \epsilon_3.$$

Thus, when $\Delta_a = 0$, $X_1 + X_2$ cancels the pair background, whereas $X_1 - X_2$ is the background-leakage direction. The OOD environments keep both background axes fixed and only reverse $a_s \mapsto -a_s$. The corresponding oracle mass is $\Gamma_E = d_{\text{pair}} d_{\text{pair}}^\top + d_{\text{short}} d_{\text{short}}^\top$, where $d_{\text{pair}} = (\delta_c, -\delta_c, 0)^\top$ and $d_{\text{short}} = (0, 0, \delta_s)^\top$. Hence $\boldsymbol{\beta}^\top \Gamma_E \boldsymbol{\beta} = \delta_c^2(\beta_1 - \beta_2)^2 + \delta_s^2 \beta_3^2$. The diagonal baseline uses $\text{diag}(\Gamma_E)$, whereas the matrix R_2 and R_4 objectives use the full oracle Γ_E to isolate the spectral mechanism.

The construction uses $a_c = 0.6$, $\delta_c = 1$, $a_s = 0.75$, $\delta_s = 1$, $\sigma = 1$, and $\Delta_a = \Delta_{a,s} = 0$. Each training environment contains 1024 samples, and each IID and OOD test environment contains 4096 samples. For each $\lambda \in \{0, 0.03, 0.1, 0.3, 1, 3, 10, 20, 30, 50, 100\}$, training runs for 3000 steps over five seeds. The stable-mode loading is $(\beta_1 + \beta_2)/\sqrt{2}$, normalized by its absolute value for the same method and seed at $\lambda = 0$; the shortcut loading is $|\beta_3|$.