THE ROLE OF DATA REUSE IN SHORTCUT LEARNING WITH LINEAR NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many dynamical theories of shortcut learning use fresh samples or population gradients, whereas most practical training repeatedly uses a finite dataset. We study full-batch gradient descent with square loss in a synthetic linear model of spurious correlations. Our analysis uses a non-rigorous dynamical mean-field calculation from statistical physics. The resulting equations separate shortcut and core learning from fitting directions specific to the training dataset. A perturbative analysis predicts a reuse crossover on a timescale proportional to $\sqrt{P/D}$, where P is dataset size and D is input dimension. Without regularisation, the interpolation peak at $P = D$ can degrade test performance without impairing core or shortcut acquisition. For $P < D$, overlap between the features can amplify the shortcut coefficient. Ridge suppresses this peak but can increase the ratio of shortcut to core coefficients. Simulations validate the theory, while image benchmarks provide qualitative evidence beyond the solvable setting.

1 INTRODUCTION

How does training time affect robustness to spurious correlations? A classifier can predict training labels from features whose association with the label does not persist at deployment (Geirhos et al., 2020). For example, a bird classifier can rely on image backgrounds and fail when the relation between background and bird class changes (Sagawa et al., 2020b). Therefore, high training accuracy does not necessarily imply reliable prediction under distribution shift.

This dependence on spurious features develops during training. A shortcut can be learned before a more reliable core feature, with acquisition order depending on feature complexity and noise (Qiu et al., 2024; Yang et al., 2024). Dynamical analyses have characterised early shortcut dominance and identified distinct shortcut and core relaxation times (Yang et al., 2024; Jain et al., 2024). Instead, more stable core features may be harder to extract, favouring shortcut features (Hermann et al., 2024). Understanding the timescale of shortcut learning has practical consequences as some mitigation methods require identifying shortcut usage from initial training dynamics (Liu et al., 2021).

Reusing a finite dataset adds another complication: gradient descent also fits fluctuations specific to the sampled examples. Benign-overfitting theory shows that fitting noisy observations can be compatible with small excess risk (Bartlett et al., 2020). With spurious correlations, however, the ability to fit individual examples can also support a shortcut solution: the predictor relies on the shortcut and fits the conflicting examples individually (Sagawa et al., 2020b). Analyses of linear-regression dynamics show how the empirical covariance spectrum controls relaxation and the interpolation peak (Advani et al., 2020). Understanding generalisation therefore requires accounting for both the cost of sample-specific fitting and the learned core and shortcut coefficients.

Existing finite-dataset theories describe how corrections to online learning accumulate during training (Bordelon et al., 2024). For shortcut learning, the question is how these corrections relate to the different feature-learning timescales:

> *When does data reuse affect the dynamics relative to core and shortcut learning, and how does it change generalisation?*

We address this question in a high-dimensional linear model with core and shortcut features, trained by full-batch gradient descent with square loss. Using dynamical mean-field theory (DMFT), we

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.

derive equations for the core and shortcut coefficients and the squared norm outside their span. These quantities separate changes in feature dependence from sample-specific fitting and allow us to analyse their dependence on dataset size, training time, and ridge regularisation. The DMFT derivation is formal but non-rigorous, and we validate its predictions against simulations throughout. Our **main contributions** are:

- Early in training, while residuals remain strongly correlated across updates, the leading reuse correction grows as $t(t-1)/\alpha$, where t counts updates, $\alpha = P/D$, P is dataset size, and D is input dimension. This gives a perturbative estimate $t_{\mathrm{reuse}} \propto \sqrt{\alpha}$ for the reuse crossover.
- For ridgless regression, at the interpolation threshold $\alpha = 1$, the predictor overfits sample-specific directions orthogonal to the shortcut–core plane (Fig. 4). For $\alpha < 1$, the minimum-norm solution can retain the shortcut direction more strongly and, when shortcut and core directions overlap, increase the shortcut coefficient (Result 2).
- Ridge suppresses sample-specific fitting near interpolation (Fig. 4), but can increase $c_{\mathrm{sc}}/c_{\mathrm{core}}$, the ratio of shortcut to core coefficients (Result 3).
- Models trained with stronger shortcut–label correlations can achieve lower risk on test data where shortcut and label are uncorrelated, despite larger shortcut reliance. We identify a range of dataset sizes where the stronger-correlation model initially has higher test risk but lower risk later in training (Result 4).
- Finite-dimensional simulations validate the DMFT. Results on Waterbirds, Spawrious, and Colored MNIST support its qualitative predictions beyond the solvable setting (Sec. 4).

1.1 Related Work

Shortcut selection and learning dynamics. The preference for a shortcut depends on both the data and the training objective. Geometric and statistical skews can cause out-of-distribution failure in linear classification (Nagarajan et al., 2021), and cross-entropy training can favour a shortcut even when a core feature perfectly predicts the label (Puli et al., 2023). Feature predictivity alone therefore does not determine which feature is learned: availability, or ease of extraction, also matters (Hermann et al., 2024). Shortcut reliance can persist even when core information is encoded in the learned representation (Izmailov et al., 2022; Kirichenko et al., 2023). Dynamical analyses identify several mechanisms behind this preference. Gradient starvation describes how learning a stronger feature can suppress learning of another (Pezeshki et al., 2021), while differences in feature complexity can favour shortcuts and delay core learning (Qiu et al., 2024). For a two-layer network trained on a fixed dataset, Yang et al. (2024) characterise early shortcut dominance through a near-linear approximation. Related spectral analyses connect learning order to the leading eigenfunctions of the neural tangent kernel (Lim et al., 2025). Class imbalance also changes gradient competition and stochastic-gradient noise (Francazi et al., 2023). Closer to our setting, Jain et al. (2024) derive distinct shortcut and core relaxation times for online linear learning; recent theory establishes early shortcut growth under online SGD in a nonlinear XOR model (LaBonte & Muthukumar, 2026). We investigate how these feature-learning timescales compare with the onset of data-reuse effects, and when reduced sample-specific fitting outweighs worse population decorrelated risk.

Interpolation, regularisation, and spurious correlations. Benign-overfitting identifies conditions under which minimum-norm interpolation of noisy observations has small excess risk (Bartlett et al., 2020); extensions to ridge regression show how regularisation balances signal retention and noise fitting (Tsigler & Bartlett, 2023). Such guarantees can also hold under structured covariate shifts (Tang et al., 2025), but interpolation does not necessarily preserve robustness to spurious correlations. Indeed, Sagawa et al. (2020b) show how overparameterised classifiers can rely on a shortcut while fitting shortcut-conflicting examples individually. Wald et al. (2023) construct settings where learned linear classifiers that interpolate with positive margin fail under an environment shift although a non-interpolating learner succeeds. Precise high-dimensional analyses further characterise spurious dependence in ridge regression (Bombari & Mondelli, 2025) and bias amplification in group prediction risks (Subramonian et al., 2025). The geometry of heterogeneous data can itself produce group disparities, as shown in a teacher–mixture model (Sarao Mannelli et al., 2025). A different form of spurious dependence is memorisation of task-irrelevant features: Bombari & Mondelli (2024) relate it to sample stability and feature alignment in random-feature and neural tangent

kernel regression. Our analysis follows both the core and shortcut coefficients and fitting outside their span. This separates the interpolation peak from changes in feature dependence and explains how ridge can suppress sample-specific fitting while increasing the shortcut-to-core coefficient ratio.

Learning dynamics on finite datasets. Reusing a finite dataset introduces correlations between the predictor and the training examples. In linear regression, spectral analyses characterise how the empirical covariance controls relaxation and generalisation under gradient descent and SGD (Advani et al., 2020; Paquette et al., 2021). DMFT accounts for the history dependence induced by reuse in Gaussian-mixture classification, random-feature models, and deep linear networks (Mignacco et al., 2020; Bordelon et al., 2024; Bordelon & Pehlevan, 2025). Rigorous high-dimensional limits have also been established for broad classes of first-order algorithms with random data (Celentano et al., 2021; Gerbelot et al., 2024). Closer to our question, Montanari & Urbani (2025) use a Gaussian approximation and DMFT to describe early feature learning followed by overfitting and feature unlearning in a nonlinear two-layer model. Following the generating-functional formulation of Bordelon & Pehlevan (2025), we use a non-rigorous calculation, validated against finite-dimensional simulations, to separate core and shortcut learning from sample-specific fitting. This allows us to compare the onset of reuse effects with feature-learning timescales and identify increases in test risk that do not require a decrease in the core coefficient.

2 Model and Observables

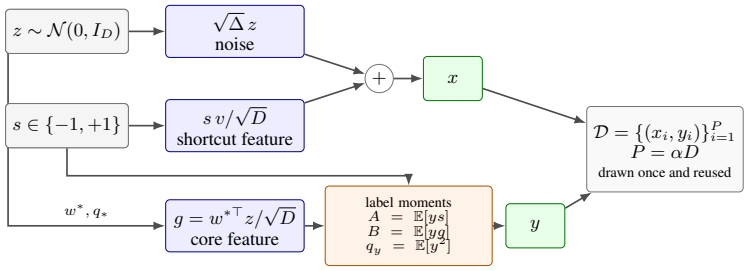

Fig. 1: Finite-reuse shortcut model with $\mu = 0$: s and z generate the input, s and the core field g determine the label moments, and $P = \alpha D$ input–label pairs are drawn once and reused. The parameter μ in (1) rescales the core component without changing this graph.

We study a teacher–student model with shortcut direction v, core direction w^*, input dimension D, and a training set of $P = \alpha D$ examples drawn once and reused. The proportional limit sends $D, P \to \infty$ at fixed α. For each example, draw a component sign $s_i \overset{\text{iid}}{\sim} \text{Unif}\{-1, +1\}$ and an isotropic Gaussian vector $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_D)$, then construct

$$g_i = \frac{w^{*\top} z_i}{\sqrt{D}}, \qquad z_{\perp,i} = z_i - \frac{g_i}{q_*} \frac{w^*}{\sqrt{D}}, \qquad x_i = s_i \frac{v}{\sqrt{D}} + \sqrt{\Delta}\, z_{\perp,i} + \sqrt{\Delta + \mu}\, \frac{g_i}{q_*} \frac{w^*}{\sqrt{D}}, \tag{1}$$

were $\Delta > 0$ is the background variance and $\mu \geq 0$ is the additional variance along the core direction. The scalar g_i is the latent core field and $z_{\perp,i}$ contains the Gaussian variation orthogonal to w^*. Thus s_i controls the shortcut coordinate, while μ changes the variance along the core direction. Fig. 1 illustrates the model design for $\mu = 0$ for readability, the general $\mu \geq 0$ case rescales the core component without changing the intuition.

The shortcut and core directions span a two-dimensional plane—later referred to as *signal plane*—whose geometry is summarised by

$$\nu = \frac{\|v\|^2}{D}, \qquad \kappa = \frac{v^\top w^*}{D}, \qquad q_* = \frac{\|w^*\|^2}{D}, \qquad G = \begin{pmatrix} \nu & \kappa \\ \kappa & q_* \end{pmatrix}, \qquad \det G > 0. \tag{2}$$

Here ν fixes the shortcut norm, q_* fixes the core norm, and κ measures their overlap. The condition $\det G > 0$ excludes collinear directions.

Notice that for square loss, shortcut and core predictivity enter through only three moments:

$$A = \mathbb{E}[ys], \quad B = \mathbb{E}[yg], \quad q_y = \mathbb{E}[y^2]. \tag{3}$$

Labels are defined by

$$y_i = \mathrm{sign}({w^*}^\top x_i/\sqrt{D}). \tag{4}$$

This fixes $q_y = 1$, while the expression for A and B can be easily derived (see App. D.5).

Our main results concern shallow linear predictors $f(x,t) = q(t)^\top x/\sqrt{D}$ with $q(t) \in \mathbb{R}^D$, although the framework extends to deeper architectures, as discussed in App. C.4.1.

2.1 Training Dynamics and Observables

We train with square loss with ridge penalty λ_{L2}. Writing $\ell(q;x,y) = (y - q^\top x/\sqrt{D})^2$, the empirical and population risks are

$$\widehat{\mathcal{L}}(t) = \frac{1}{P}\sum_{i=1}^{P} \ell(q(t);x_i,y_i), \quad \text{and} \quad \mathcal{L}(t) = \mathbb{E}\ell(q(t);x,y), \tag{5}$$

where the expectation is over the generative model in (1). We define decorrelated risk using test moments $A_e = 0$, $B_e = B$, and $q_y = 1$, while keeping the input distribution unchanged. This evaluation does not preserve the training label rule, however the labels used for training remain unchanged.

Define the residual of training example i by $r_i(t) = y_i - q(t)^\top x_i/\sqrt{D}$. One gradient descent update with step size ηD on $\widehat{\mathcal{L}}/2 + \lambda_{\mathrm{L2}}\|q\|^2/(2D)$ is

$$q(t+1) = (1 - \eta\lambda_{\mathrm{L2}})q(t) + \eta\frac{\sqrt{D}}{P}\sum_{i=1}^{P} x_i r_i(t), \tag{6}$$

Since the same P examples are used at every step, the gradient at time t is correlated with the entire previous trajectory. We adapt the Martin–Siggia–Rose construction of Bordelon & Pehlevan (2025) to the core–shortcut mixture. DMFT makes reuse-induced temporal correlations explicit and supports extensions to deep linear networks and persistent SGD, as shown in App. C.4.

The theory outputs the shortcut and core overlaps $m(t) = q(t)^\top v/D$ and $\theta(t) = q(t)^\top w^*/D$, together with the squared norm $Q(t) = \|q(t)\|^2/D$. When v and w^* are not orthogonal, the overlaps mix the two directions. We therefore write

$$q(t) = c_{\mathrm{sc}}(t)v + c_{\mathrm{core}}(t)w^* + q_{\mathrm{out}}(t), \tag{7}$$

with $q_{\mathrm{out}}(t) \perp \mathrm{span}\{v, w^*\}$. The coefficients and remaining normalized squared norm are

$$\begin{pmatrix} c_{\mathrm{sc}}(t) \\ c_{\mathrm{core}}(t) \end{pmatrix} = G^{-1}\begin{pmatrix} m(t) \\ \theta(t) \end{pmatrix}, \quad Q_{\mathrm{out}}(t) = \frac{\|q_{\mathrm{out}}(t)\|^2}{D} = Q(t) - \begin{pmatrix} m(t) \\ \theta(t) \end{pmatrix}^T G^{-1}\begin{pmatrix} m(t) \\ \theta(t) \end{pmatrix}. \tag{8}$$

Thus c_{sc} measures shortcut use after removing the core direction, c_{core} measures core recovery, and Q_{out} is the normalized squared norm outside the population signal plane.

These quantities determine the population risk in environment e,

$$\mathcal{L}_e(t) = q_y - 2A_e m(t) - 2\frac{\sqrt{\Delta+\mu}}{q_*}B_e\theta(t) + m(t)^2 + \Delta Q(t) + \frac{\mu}{q_*}\theta(t)^2. \tag{9}$$

The two negative terms reward alignment with shortcut and core evidence; the remaining terms measure prediction variance, including the contribution ΔQ_{out} from sample-specific directions. The choices $A_e = A$ and $A_e = 0$, with $B_e = B$ and $q_y = 1$, give the in-distribution and moment-defined decorrelated risks, respectively. Training risk is computed from the reused residuals in (6).

3 Results

3.1 Relevant Timescales and Online-to-Reuse Crossover

In Fig. 2, the test risk initially follows online learning, which uses a fresh batch of P examples at every step. As correlations across updates accumulate, finite-dataset training deviates from the

online trajectory, and eventually training and test risks approach the replica predictions (Engel et al., 2001). For shortcut learning, the relative timing of the effect of reuse and the relaxation of the shortcut and core features is crucial.

We first consider relaxation without ridge regularisation ($\lambda_{\mathrm{L2}} = 0$). The mean online gradient is governed by the population covariance $\Sigma_x = \mathbb{E}[xx^\top] = \Delta I_D + vv^\top/D + \mu w^* w^{*\top}/(Dq_*)$. A stable mode with curvature γ contracts by $|1-\eta\gamma|$ per step, giving the relaxation time $t_\gamma = -1/\log|1-\eta\gamma|$.

For $\kappa = 0$, the shortcut and core directions are covariance eigenvectors. Their relaxation times are

$$t_{\mathrm{sc}} = t_{\Delta+\nu}, \qquad t_{\mathrm{core}} = t_{\Delta+\mu}. \tag{10}$$

The extension to $\kappa \neq 0$ is given in App. D.3. On a finite dataset, the lower nonzero bulk edge of the empirical covariance gives the bulk relaxation time

$$t_{\mathrm{edge}} = t_{\Delta(1-\alpha^{-1/2})^2}. \tag{11}$$

Reusing the same examples correlates gradient fluctuations across updates, producing a correction to sample-specific fitting $\chi_{\mathrm{reuse}}(t) = Q_{\mathrm{out}}^{\mathrm{reuse}}(t) - Q_{\mathrm{out}}^{\mathrm{on}}(t)$ ((8)). This correction tracks the departure from online learning in our simulations. Its contribution to test risk is $\Delta\chi_{\mathrm{reuse}}$. The full decomposition, including changes in the core and shortcut coefficients, is given in (149).

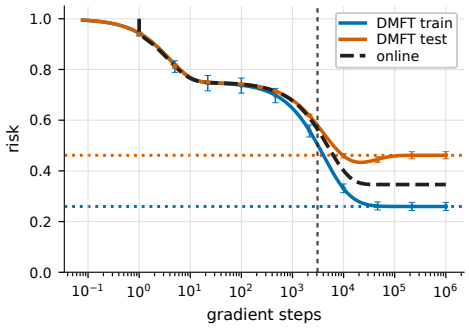

Fig. 2: **Departure from online learning.** Solid blue/orange curves: theory training/test risks; points: simulation means ± standard deviation. Black dashed: online risk; vertical dotted: reuse crossover; horizontal dotted: replica predictions. *Parameters:* App. B.

Starting from $q(0) = 0$, we expand the covariance updates in App. C.3—specifically Eqs. (61), (62), and (69)—for $\eta\Delta t \ll 1$. Neglecting response damping while retaining the evolving residual correlations along the online learning path gives

$$\chi_{\mathrm{reuse}}(t) \simeq \frac{\eta^2\Delta}{\alpha} \sum_{\substack{0\le t',t''<t \\ t'\neq t''}} C_r^{\mathrm{on}}(t',t''), \qquad C_r^{\mathrm{on}}(t',t'') = \mathbb{E}[r^{\mathrm{on}}(t')r^{\mathrm{on}}(t'')]\,. \tag{12}$$

The two residuals are evaluated on the same independent test sample. At this order, equal-time contributions cancel between reused and fresh batches. The remaining $t(t-1)$ ordered pairs accumulate coherently while correlations across updates remain positive and of comparable magnitude.

Result 1 (Perturbative reuse crossover)**.** *Fix $\alpha, \Delta > 0$ and suppose the average residual correlation between distinct updates remains comparable to the evolving population risk along the online learning path. For a relative tolerance $\eta^2 \ll \varepsilon \ll 1$, define the reuse crossover through the sample-*

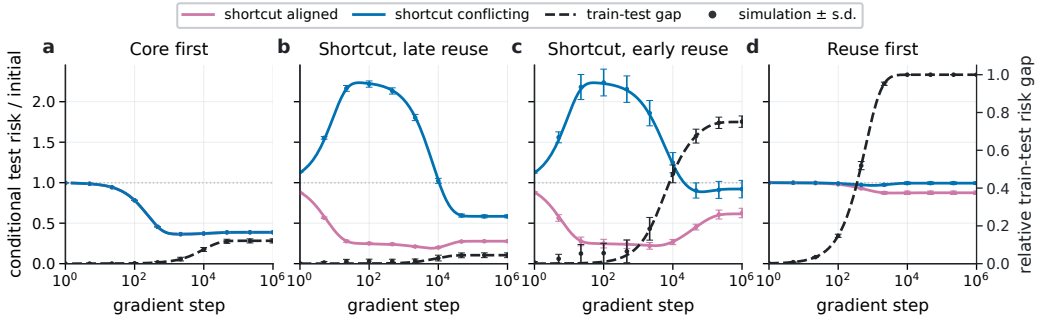

Fig. 3: **Ordering of shortcut relaxation, core recovery, and finite-dataset reuse lead to a rich taxonomy.** Solid curves: theory; points: simulations. Blue: test risk on shortcut-conflicting examples; pink: test risk on shortcut-aligned examples; black: train–test risk gap. *Parameters:* App. B.

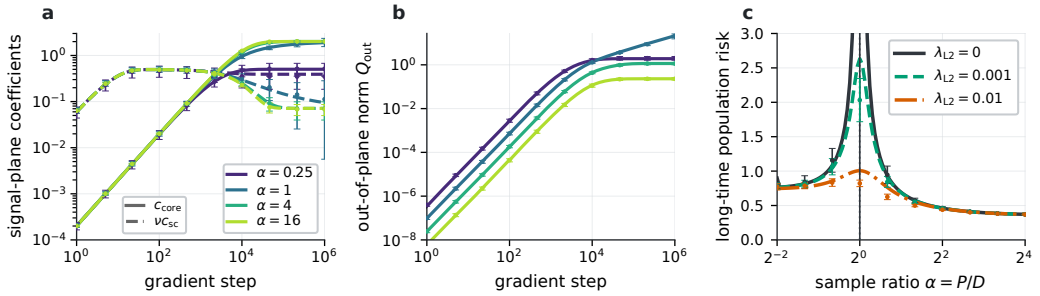

Fig. 4: **Effect of dataset size and regularization.** Panels (a,b) show DMFT trajectories for $\alpha \in \{0.25, 1, 4, 16\}$, reporting theory and simulations. Panel (a) shows the core and rescaled shortcut coefficients. Panel (b) shows the squared norm outside the signal plane. Panel (c) shows long-time proportional-limit predictions and finite-dataset spectral calculations for three ridge values; the line marks $\alpha = 1$. Further manipulations in App. E.2.2. *Parameters:* App. B.

specific population-risk contribution $\Delta\chi_{\text{reuse}}$*, following* (9)*:*

$$t^{(\varepsilon)}_{\text{reuse}} := \inf\{t : \Delta\chi_{\text{reuse}}(t) > \varepsilon\mathcal{L}_{\text{on}}(t)\} \propto \frac{\sqrt{\alpha\varepsilon}}{\eta\Delta} \tag{13}$$

when $1 \ll t^{(\varepsilon)}_{\text{reuse}} \ll 1/(\eta\Delta)$.

The ordering of t_{sc}, t_{core}, and t_{reuse} distinguishes the four regimes in Fig. 3. In the *core-first* regime (a), $t_{\text{core}} < \min(t_{\text{sc}}, t_{\text{reuse}})$: test risks on both shortcut-aligned and shortcut-conflicting samples decrease before reuse becomes appreciable. When the shortcut is learned first, the conflicting-sample risk initially rises. With *late reuse* (b), $t_{\text{sc}} < t_{\text{core}} < t_{\text{reuse}}$, core recovery largely reverses this rise before reuse becomes appreciable. With *early reuse* (c), $t_{\text{sc}} < t_{\text{reuse}} < t_{\text{core}}$, the train–test gap grows during core recovery, and the conflicting-sample risk remains high. In the *reuse-first* regime (d), $t_{\text{reuse}} < \min(t_{\text{sc}}, t_{\text{core}})$: a large train–test gap develops while test risks remain close to their initial values, indicating overfitting before either feature is learned. Further details in App. D.2.

3.2 Reuse Produces Sample-Specific Fitting and Can Amplify the Shortcut

In ridgeless least squares, the population risk asymptotically gives the well-known interpolation peak at $\alpha = 1$ (Advani et al., 2020; Nakkiran et al., 2021) shown in Fig. 4c. The singularity comes from ΔQ_{out} in Eqs. (8) and (9), whereas the shortcut and core coefficients c_{sc} and c_{core} remain finite (Fig. 4a,b). Nevertheless, data reuse can change these coefficients relative to online learning.

Write $c^{\text{pop}}_{\text{sc}}, c^{\text{pop}}_{\text{core}}$ for the population-risk minimiser's coefficients, and $c^{\text{fin}}_{\text{sc}}, c^{\text{fin}}_{\text{core}}$ for the proportional limits of the minimum-norm finite-dataset coefficients, taking training to convergence first. These static limits follow directly from least squares, independently of the formal DMFT derivation (App. D.4). For $0 < \alpha < 1$ and $\mu = 0$, the training-input span retains the population components along v and $w^*_\perp = w^* - (\kappa/\nu)v$ by unequal factors r_{sc} and α. Hence

$$c^{\text{fin}}_{\text{sc}} - c^{\text{pop}}_{\text{sc}} = -(1 - r_{\text{sc}})c^{\text{pop}}_{\text{sc}} + \frac{\kappa}{\nu}(r_{\text{sc}} - \alpha)c^{\text{pop}}_{\text{core}}, \qquad r_{\text{sc}} = \frac{\alpha(\Delta + \nu)}{\Delta + \alpha\nu} > \alpha. \tag{14}$$

Unequal retention can increase the shortcut coefficient despite its direct attenuation. The two-dimensional calculation in App. D.4 gives the condition for general $\mu \geq 0$:

Result 2 (Finite-dataset training can amplify the shortcut)**.** *Consider ridgeless square-loss training with* $q(0) = 0$*, a stable step size,* $0 < \alpha < 1$*,* $\mu \geq 0$*,* $\kappa > 0$*, and positive population coefficients. In the proportional limit after convergence, the shortcut coefficient is amplified if and only if*

$$c^{\text{fin}}_{\text{sc}} > c^{\text{pop}}_{\text{sc}} \quad \Longleftrightarrow \quad \alpha\kappa\, c^{\text{pop}}_{\text{core}} > (\Delta + \alpha\mu)c^{\text{pop}}_{\text{sc}}.$$

Amplification requires overlap: at $\kappa = 0$, the two directions are attenuated separately. For fixed $\alpha > 1$, the ridgeless coefficients recover asymptotically their population values, however Q_{out} remains nonzero and large near interpolation.

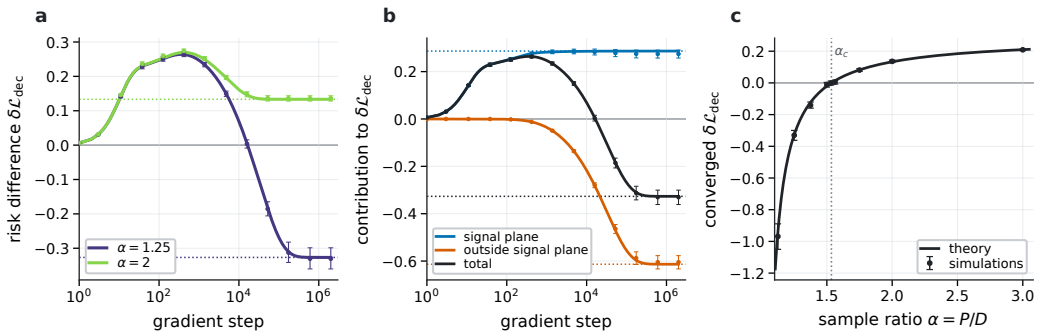

Fig. 5: **Stronger shortcut correlations can reduce decorrelated risk.** Comparison with $\kappa_2 > \kappa_1$. (a) Risk difference during training below and above α_c. (b) Signal-plane and sample-specific contributions below α_c: reduced fitting outside the signal plane outweighs the increased signal-plane risk. (c) Converged risk difference versus α, with the predicted crossing at $\alpha_c = 1 + d/p$. Each problem uses its own B, with $A_e = 0,\ B_e = B$. *Parameters:* App. B.

Ridge suppresses the growth of Q_{out} near interpolation, but does not necessarily favour the core feature. Its effect on relative feature weights is already visible from the population solution.

Result 3 (Ridge can increase the shortcut-to-core coefficient ratio)**.** *For $\mu = 0$, $A, B > 0$, and $\kappa \geq 0$, the population coefficients satisfy the exact identity*

$$\frac{\partial}{\partial \lambda_{\text{L2}}} \frac{c_{\text{sc}}^{\text{pop}}}{c_{\text{core}}^{\text{pop}}} = \frac{A\nu q_* / (\sqrt{\Delta} B) + \kappa}{(\Delta + \nu + \lambda_{\text{L2}})^2} > 0.$$

For every fixed $\alpha > 0$, the proportional-limit finite-dataset ratio also increases with $\lambda_{\text{L2}} > 0$.

The finite-dataset coefficients asymptotically equal the population coefficients at an effective ridge parameter that increases with λ_{L2} (App. D.4). Thus ridge can suppress Q_{out} while increasing relative shortcut reliance, without increasing the absolute shortcut coefficient. This complements Bombari & Mondelli (2025), who show that minimising in-distribution risk and minimising their measure of spurious correlation can favour different ridge strengths.

For $\mu > 0$, ridge can instead reduce the population shortcut-to-core ratio. In a covariance mode of curvature γ, ridge retains the fraction $\gamma/(\gamma + \lambda_{\text{L2}})$ of the unregularised coefficient, suppressing lower-curvature modes more strongly. At $\mu = 0$, the higher-curvature mode lies along the shortcut. Increasing μ adds variance along the core direction and can make ridge preserve the core contribution more strongly instead. The sign also depends on the overlap and the moments (3); see (139).

3.3 Stronger Shortcut Correlations Can Reduce Decorrelated Risk

Consider two realisations of the problem with different shortcut–core overlaps $\kappa_1 < \kappa_2$ at fixed (Δ, μ, ν, q_*). Notice that increasing κ changes both feature geometry and the core–label correlation—alternative targets could avoid this coupling (see App. D.5.1). Define the risk difference

$$\delta\mathcal{L}_{\text{dec}}(t) = \mathcal{L}_{\text{dec}}(t; \kappa_2) - \mathcal{L}_{\text{dec}}(t; \kappa_1). \tag{15}$$

Data reuse can reverse the decorrelated-risk ordering during training (Fig. 5a). For the parameter pairs considered below, the population predictor has higher decorrelated risk but lower training residual variance at larger κ. The smaller residual reduces sample-specific fitting, which can outweigh the population-risk increase.

To quantify this, we separate the risk of the population predictor from the additional risk due to fitting a finite dataset. For ridgeless training with $\alpha > 1$, the signal coefficients converge to their population least-squares values. Let $\mathcal{L}_{\text{dec}}^{\text{pop}}(\kappa)$ be the decorrelated risk of this population predictor and $\sigma_{\text{res}}^2(\kappa)$ its residual variance on the training distribution. After training converges, the risks satisfy

$$\mathcal{L}_{\text{dec}}(\infty; \kappa) = \mathcal{L}_{\text{dec}}^{\text{pop}}(\kappa) + \frac{\sigma_{\text{res}}^2(\kappa)}{\alpha - 1}, \qquad \Delta Q_{\text{out}}(\infty; \kappa) = \frac{\sigma_{\text{res}}^2(\kappa)}{\alpha - 1}. \tag{16}$$

Reducing this residual reduces sample-specific fitting, even if the population decorrelated risk increases.

Result 4 (Reversal of decorrelated-risk ordering)**.** *Consider stable ridgeless gradient descent from* $q(0) = 0$*, with* $\alpha > 1$*,* $\mu \geq 0$*, and* $0 \leq \kappa_1 < \kappa_2 < \sqrt{\nu q_*}$*, at fixed* (Δ, μ, ν, q_*)*. Define the population-risk increase* p *and residual-variance reduction* d *by*

$$p = \mathcal{L}_{\mathrm{dec}}^{\mathrm{pop}}(\kappa_2) - \mathcal{L}_{\mathrm{dec}}^{\mathrm{pop}}(\kappa_1), \qquad d = \sigma_{\mathrm{res}}^2(\kappa_1) - \sigma_{\mathrm{res}}^2(\kappa_2).$$

If $p, d > 0$*, the proportional-limit risk difference* $\delta\mathcal{L}_{\mathrm{dec}}(t) = \mathcal{L}_{\mathrm{dec}}(t; \kappa_2) - \mathcal{L}_{\mathrm{dec}}(t; \kappa_1)$ *obeys*

$$\delta\mathcal{L}_{\mathrm{dec}}(1) > 0, \qquad \delta\mathcal{L}_{\mathrm{dec}}(\infty) = p\,\frac{\alpha - \alpha_c}{\alpha - 1}, \qquad \alpha_c = 1 + \frac{d}{p}. \tag{17}$$

Thus the risk ordering reverses during training for $1 < \alpha < \alpha_c$*.*

Fig. 5b shows how the two contributions change during training at $\alpha = 1.25$. At convergence, the stronger shortcut increases the signal-plane contribution by p, but reduces the risk from fitting outside the signal plane by $d/(\alpha-1)$. The latter reduction is larger, giving a lower total decorrelated risk despite greater shortcut use. Fig. 5c instead shows the converged total risk difference as a function of dataset size. As follows from (16), increasing α reduces the contribution from sample-specific fitting, while the population-risk difference remains p. The benefit of the stronger shortcut therefore weakens: the risk difference crosses zero at $\alpha_c = 1 + d/p$ and becomes positive above it. Thus a stronger shortcut can lower decorrelated risk near interpolation without improving the population predictor. Derivation is provided in App. D.5.

4 Experiments on Image Benchmarks

We test consequences of the theory in standard suprious correlation benchmarks that violate the assumptions of the solvable model. First, we use random-feature models on Waterbirds and Spawrious, which retain a linear readout dynamics while introducing real-image feature statistics. Second, we train a CNN on Colored MNIST, therefore relaxing the fixed-feature and square-loss assumptions.

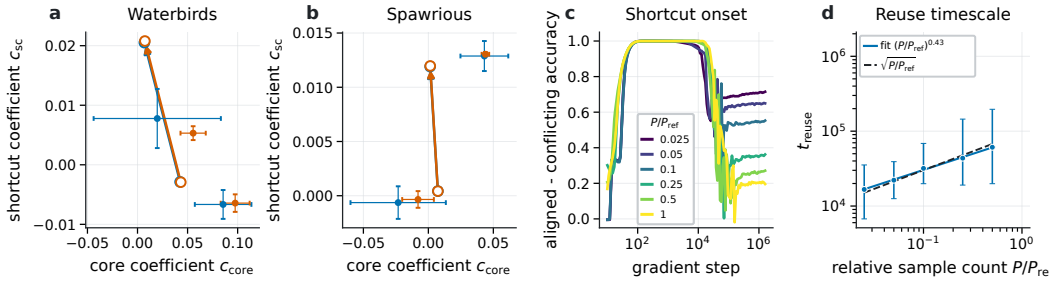

Fig. 6: **Feature coefficients and dataset-size effects on image benchmarks. (a,b)** Final core and shortcut coefficients for Waterbirds and Spawrious. Arrows increase p_{train} from 0.50 to 0.95; blue/orange denote $\alpha = 0.5/4$, and open/filled markers denote DMFT predictions and simulation means $\pm$ standard deviation over five independently sampled training datasets. **(c)** In Colored MNIST at $p_{\mathrm{train}} = 0.80$, the aligned-minus-conflicting accuracy gap starts to grow at similar updates across dataset sizes P, while the final value changes. **(d)** Measured departure times from the $P_{\mathrm{ref}} = 6000$ risk trajectory (labelled t_{reuse}) and power-law fit with exponent 0.43. The dashed line is a square-root reference, expected from Result 1. *Parameters:* App. B.

4.1 Random-Feature Regression on Waterbirds and Spawrious

In Waterbirds, bird type is the target and background is the shortcut (Sagawa et al., 2020a); in Spawrious, the target is Labrador versus Bulldog and the shortcut is snow versus jungle (Lynch et al., 2025). We use a random-feature model with $D = 512$ random features and train only the linear readout on MSE. The parameters of the Gaussian model used for the DMFT predictions are estimated from a 15% calibration split.

Fig. 6(a,b) compares the resulting DMFT predictions with benchmark training. Increasing the shortcut agreement probability, $p_{\rm train} = \Pr(y = s)$, raises $c_{\rm sc}$ in both datasets, for both the underdetermined $\alpha = 0.5$ and overdetermined $\alpha = 4$ settings. The theory reproduces the direction of change, whereas the magnitude of the core-coordinate response is less accurately predicted. Notice that these are tests of the signal-plane decomposition under non-Gaussian image statistics, neither the trajectories nor their large-time coefficients are fitted in the analysis.

4.2 CNN ON COLORED MNIST

On Colored MNIST, the label distinguishes digits 0–4 from 5–9, and color agrees with it with probability $p_{\rm train}$ (Arjovsky et al., 2019). We train a BatchNorm CNN from scratch with cross-entropy and full-batch gradient descent, varying dataset size P at fixed $p_{\rm train}$.

The test-accuracy gap between examples whose color agrees or conflicts with the label starts to grow at similar training steps across dataset sizes. At $p_{\rm train} = 0.80$, the gap initially approaches one for all P, before falling to smaller values on larger datasets (Fig. 6(c)). The common onset and different late-training gaps show that dataset size does not simply rescale the learning trajectory in time.

To quantify dataset-size effects on the risk trajectory, we compare each run with a reference trained on $P_{\rm ref} = 6000$ examples. We average the cross-entropy losses of the aligned and conflicting test groups, giving them equal weight, and define $t_{\rm dep}$ as the first persistent 10% relative difference from the reference. Larger datasets depart later. Fitting departure times across agreement probabilities $p_{\rm train} \geq 0.60$ gives $t_{\rm dep} \propto (P/P_{\rm ref})^{0.43}$ (Fig. 6(d)). For the reference itself $t_{\rm dep}$ is trivially infinity by definition, so it is omitted from panel (d).

Together, the similar shortcut-onset times and the delayed departure support the theory's separation of early shortcut learning and later finite-dataset effects, here with learned features and cross-entropy training. Unlike the random-feature comparison, this test uses no Gaussian-mixture parameter fit to the network's representation. The comparison with the square-root prediction in (13) remains qualitative because $t_{\rm dep}$ is measured against a finite reference. App. E.3.2 gives the protocol, fit, and finite-reference correction.

5 DISCUSSION

Our DMFT separates core and shortcut learning from sample-specific fitting, with a perturbative reuse timescale $t_{\rm reuse} \propto \sqrt{\alpha}$ that determines when reuse competes with core recovery. We then articulated the different ways reuse can affect core acquisition. In the ridgeless scenario, where sample-specific fitting leads to the interpolation peak, we show that below interpolation, feature overlap can amplify the shortcut coefficient. Adding ridge regularisation suppresses sample-specific fitting, but can increase the shortcut-to-core ratio. Finally we show that stronger shortcut–label correlations can also lower decorrelated risk despite greater shortcut reliance by reducing sample-specific fitting, therefore reversing the risk ordering during training.

Our analysis, validated numerically, focuses on shallow linear predictors trained by full-batch square-loss learning in a Gaussian-mixture model. The method can be generalised to deep and nonlinear networks (Bordelon & Pehlevan, 2025; 2022), as well as other optimisers and learning rules (Sarao Mannelli & Urbani, 2021; Mignacco et al., 2020; Bordelon & Pehlevan, 2023), see Apps. C.4.1, C.4.2 and E.2.1. In particular, smaller batches can regularise Gaussian-mixture classification, providing a potential additional mechanism for suppressing $Q_{\rm out}$ and shortcut reliance.

Nonlinear feature learning can change the role of reuse itself. Repeated batches can enable feature acquisition (Dandi et al., 2024), while later training can cause feature unlearning (Montanari & Urbani, 2025). Studying core and shortcut learning separately would identify which features are acquired, reinforced, or lost. Cross-entropy introduces further competition through gradient starvation (Pezeshki et al., 2021). These extensions generally require numerical self-consistency over effective trajectories (Bordelon & Pehlevan, 2022; Roy et al., 2019; Mignacco et al., 2020), making explicit timescale estimates harder to obtain.

Group geometry and proportions affect learning times (Jain et al., 2024). Distinguishing minority-signal acquisition from sample-specific fitting could help explain why majority subsampling can outperform minority upweighting (Sagawa et al., 2020b). Reweighting changes gradient contributions, whereas subsampling also changes dataset size and hence reuse; comparing them would clarify how each affects core learning.

Reproducibility Statement

The model and assumptions are specified in Sec. 2, with derivations in Apps. C and D. The experimental setup is described in Sec. 4, with numerical methods and detailed protocols in Apps. E.1 to E.3. Notation is summarised in App. A, and figure parameters are reoprted in App. B.

AI Use Statement

Generative AI tools were used for language editing, programming assistance, and exploratory analysis. The authors reviewed and verified the mathematical derivations, numerical results, citations, and scientific claims and take full responsibility for the submitted work.

References

Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020. doi: 10.1016/j.neunet.2020.08.022. URL https://doi.org/10.1016/j.neunet.2020.08.022.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. doi: 10.48550/arXiv.1907.02893. URL https://arxiv.org/abs/1907.02893.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Simone Bombari and Marco Mondelli. How spurious features are memorized: Precise analysis for random and NTK features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4267–4299. PMLR, 2024. URL https://proceedings.mlr.press/v235/bombari24a.html.

Simone Bombari and Marco Mondelli. Spurious correlations in high dimensional regression: The roles of regularization, simplicity bias and over-parameterization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4839–4873. PMLR, 2025. URL https://proceedings.mlr.press/v267/bombari25a.html.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32240–32256. Curran Associates, Inc., 2022. doi: 10.52202/068431-2336. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d027a5c93d484a4312cc486d399c62c1-Abstract-Conference.html.

Blake Bordelon and Cengiz Pehlevan. The influence of learning rule on representation dynamics in wide neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nZ2NtpolC5-.

Blake Bordelon and Cengiz Pehlevan. Deep linear network training dynamics from random initialization: Data, width, depth, and hyperparameter transfer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4968–4997. PMLR, 2025. URL https://proceedings.mlr.press/v267/bordelon25a.html.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382. PMLR, 2024. URL https://proceedings.mlr.press/v235/bordelon24a.html.

Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021. doi: 10.48550/arXiv.2112.07572. URL https://arxiv.org/abs/2112.07572.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9991–10016. PMLR, 2024. URL https://proceedings.mlr.press/v235/dandi24a.html.

Andreas Engel, Christian Broeck, et al. *Statistical mechanics of learning*, volume 1. Cambridge University Press Cambridge, 2001.

Emanuele Francazi, Marco Baity-Jesi, and Aurélien Lucchi. A theoretical analysis of the learning dynamics under class imbalance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10285–10322. PMLR, 2023. URL https://proceedings.mlr.press/v202/francazi23a.html.

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(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Cédric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborová. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024. doi: 10.1137/23M1594388.

Moritz Helias and David Dahmen. *Statistical Field Theory for Neural Networks*, volume 970 of *Lecture Notes in Physics*. Springer, Cham, 2020. doi: 10.1007/978-3-030-46444-8. URL https://doi.org/10.1007/978-3-030-46444-8.

Katherine Hermann, Hossein Mobahi, Thomas Fel, and Michael C. Mozer. On the foundations of shortcut learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/bf215fa7fe70a38c5e967e59c44a99d0-Abstract-Conference.html.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/fb64a552feda3d981dbe43527a80a07e-Abstract-Conference.html.

Anchit Jain, Rozhin Nobahari, Aristide Baratin, and Stefano Sarao Mannelli. Bias in motion: Theoretical insights into the dynamics of bias in sgd training. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24435–24471, 2024. doi: 10.52202/079017-0770. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2ba538ff6a168e937d07e360045c0d6b-Abstract-Conference.html.

Hans-Karl Janssen. On a lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties. *Zeitschrift für Physik B Condensed Matter*, 23:377–380, 1976. doi: 10.1007/BF01316547.

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

Tyler LaBonte and Vidya Muthukumar. SGD provably prioritizes a shortcut spurious feature in the XOR model. *arXiv preprint arXiv:2606.30444*, 2026. URL https://arxiv.org/abs/2606.30444.

Johannes Lang, Vincenzo Citro, Luca Leuzzi, and Federico Ricci-Tersenghi. DYNAMITE: A high-performance framework for solving dynamical mean-field equations. *arXiv preprint arXiv:2604.06309*, 2026. doi: 10.48550/arXiv.2604.06309. URL https://arxiv.org/abs/2604.06309.

Jinwoo Lim, Suhyun Kim, and Soo-Mook Moon. Shortcut features as top eigenfunctions of NTK: A linear neural network case and more. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/87571720167f7e88827c40e468e3101f-Abstract-Conference.html.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International conference on machine learning*, pp. 6781–6792. PMLR, 2021.

Aengus Lynch, Gbètondji J.-S. Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. In *ICLR Workshop on Spurious Correlation and Shortcut Learning*, 2025. URL https://openreview.net/forum?id=0S0oITNTCz.

Vladimir A. Marchenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994. URL https://doi.org/10.1070/SM1967v001n04ABEH001994.

Paul C. Martin, E. D. Siggia, and H. A. Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423–437, 1973.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/file/6c81c83c4bd0b58850495f603ab45a93-Paper.pdf.

Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30086–30164, 2025. doi: 10.52202/085713-0895. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/261a154e3de61a0a8a1303c8a6c81ac7-Abstract-Conference.html.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=fSTD6NFIW_b.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the large: Average-case analysis, asymptotics, and stepsize criticality. In Mikhail Belkin and Sam Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3548–3626. PMLR, 2021. URL https://proceedings.mlr.press/v134/paquette21a.html.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Aahlad Manas Puli, Lily H. Zhang, Yoav Wald, and Rajesh Ranganath. Don't blame dataset shift! shortcut learning due to gradients and cross entropy. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3146. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e35460304fdf6df523f068a59aaf8829-Abstract-Conference.html.

Guanwen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Feature learning in the presence of spurious correlations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41658–41697. PMLR, 2024. URL https://proceedings.mlr.press/v235/qiu24e.html.

Felix Roy, Giulio Biroli, Guy Bunin, and Chiara Cammarota. Numerical implementation of dynamical mean field theory for disordered systems: Application to the Lotka–Volterra model of ecosystems. *Journal of Physics A: Mathematical and Theoretical*, 52(48):484001, 2019. doi: 10.1088/1751-8121/ab1f32. URL https://arxiv.org/abs/1901.10036.

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*, 2020a. URL https://openreview.net/forum?id=ryxGuJrFvS.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 2020b. URL https://proceedings.mlr.press/v119/sagawa20a.html.

Stefano Sarao Mannelli and Pierfrancesco Urbani. Analytical study of momentum-based acceleration methods in paradigmatic high-dimensional non-convex problems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 187–199. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/019f8b946a256d9357eadc5ace2c8678-Abstract.html.

Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020. doi: 10.1103/PhysRevX.10.011057. URL https://doi.org/10.1103/PhysRevX.10.011057.

Stefano Sarao Mannelli, Federica Gerace, Negar Rostamzadeh, and Luca Saglietti. Bias-inducing geometries: An exactly solvable data model with fairness implications. *Physical Review E*, 112 (2):025304, 2025. doi: 10.1103/nlfl-35t6. URL https://doi.org/10.1103/nlfl-35t6.

Arjun Subramonian, Samuel J. Bell, Levent Sagun, and Elvis Dohmatob. An effective theory of bias amplification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b0bd56b0fede63dade8ffc66dcebb2bd-Abstract-Conference.html.

Shange Tang, Jiayun Wu, Jianqing Fan, and Chi Jin. Benign overfitting in out-of-distribution generalization of linear models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dcf51b2eef1af32e594177b326f8cff9-Abstract-Conference.html.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023. URL https://www.jmlr.org/papers/v24/22-1398.html.

Yoav Wald, Gal Yona, Uri Shalit, and Yair Carmon. Malign overfitting: Interpolation can provably preclude invariance. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dQNL7Zsta3.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2953–2961. PMLR, 2024. URL https://proceedings.mlr.press/v238/yang24c.html.

APP. CONTENTS

A Notation **15**

B Figure Parameters **16**

C Dynamical Mean-Field Theory: Derivation and Extensions **18**

C.1 Common Reused-Data Module 18

C.2 Shallow Linear Predictor 22

C.3 DMFT equations summary for the shallow linear estimator 23

C.4 Extensions and generalisations 25

C.4.1 Architecture Modules 25

C.4.2 Persistent-SGD Extension 27

D Online Baseline, Reuse Crossover, and Feature Controls **29**

D.1 Shallow Online-SGD Mean Field 29

D.2 Linearized Reuse Crossover Around the Online Learning Path 30

D.3 Independent Core Salience and the Generalized Reuse Correction 32

D.4 Finite reuse, interpolation, and anisotropic retention 35

D.5 Risk Reversal 37

D.5.1 Risk Reversal at Fixed Geometry and Core–Label Correlation 39

E Numerical Methods and Experiments **42**

E.1 Solvers and Computational Cost 42

E.1.1 Causal Forward Solver 42

E.1.2 Validation of the Closed Shallow Equations 42

E.1.3 Adaptive Time-Grid Validation 42

E.1.4 Finite-Dimensional Simulations 44

E.1.5 Complexity 44

E.2 Synthetic-Data Validation 44

E.2.1 Validation Across Depth and Persistent SGD 44

E.2.2 Complete Finite-Reuse Sweeps 45

E.3 Image-Benchmark Experiments 46

E.3.1 Random-Feature Regression on Waterbirds and Spawrious 47

E.3.2 Controlled Colored MNIST 49

A NOTATION

Symbol	Meaning
Data, geometry, and labels	
D, P, $\alpha = P/D$	Input dimension, number of reused training examples, and sample ratio.
$a \in \{-1, +1\}$	Value of the shortcut sign after conditioning. The two signs are equiprobable, with $P_a/P \to 1/2$ in the proportional limit.
i, ℓ in (a, ℓ)	Example indices.
s_i, z_i, g_i, $z_{\perp,i}$	Component sign, isotropic Gaussian vector, latent core field, and Gaussian component orthogonal to w^*; see (1).
x_i, y_i	Input and target of example i.
$v_a = av$, $w_a^* = w^*$	Scaled shortcut centroid and shared core direction; the conditional input mean is $v_a/\sqrt{D}$ in (1).
Δ	Isotropic Gaussian variance. The additional finite-rank core variance is μ.
$\mu \geq 0$	Additional population variance along the core direction; it controls core salience in (1).
q_*, ν, κ	Core self-overlap $\|w^*\|^2/D$, shortcut self-overlap $\|v\|^2/D$, and shortcut–core overlap $v^\top w^*/D$.
G, Σ_x	Signal-plane Gram matrix from (2) and population input covariance $\Sigma_x = \mathbb{E}[xx^\top]$.
A, B, q_y	Shortcut label moment $\mathbb{E}[ys]$, core label moment $\mathbb{E}[yg]$, and label second moment $q_y = \mathbb{E}[y^2]$, equal to one for the labels in the main text.
ρ	Shortcut agreement probability $\Pr(y = s) = (1 + A)/2$ for binary labels.
α_a^{y}, β_a^{y}	Component mean label and Gaussian integration-by-parts coefficient entering the effective data block.
Training dynamics and observables	
t, s, T, η, λ_{L2}	Discrete times, training horizon, gradient step, and ridge strength.
$q(t)$, $f(x, t)$	Effective linear predictor and output $q(t)^\top x/\sqrt{D}$.
$r_i(t)$, $\Delta_{a\ell}(t)$	Prediction residual in the main text and its componentwise effective-process notation in the appendix. Indexed $\Delta_{a\ell}(t)$ is distinct from the bulk variance Δ.
$M_a(t)$, $\Theta_a(t)$, $Q(t)$	Predictor overlap with v_a, overlap with w_a^*, and normalized squared norm. For the model in the main text, $M_a(t) = a\, m(t)$ and $\Theta_a(t) = \theta(t)$.
$C_X(t, s)$	Two-time correlation of the effective field X; for example $C_q(t, s) = q(t)^\top q(s)/D$.
$R_{XY}(t, s)$	Causal response of field $X(t)$ to a source coupled to $Y(s)$.
$h_a(t)$, $h_0(t)$	Centered component-gradient field and the total input-gradient field supplied by the data block.
$L_a(t)$	Label-memory factor generated by finite reuse in the component data block.
$c_{\mathrm{sc}}(t)$, $c_{\mathrm{core}}(t)$, $Q_{\mathrm{out}}(t)$	Shortcut and core coefficients and squared norm outside their signal plane, divided by D; see (8).
$\mathcal{L}_e(t)$, $\widehat{\mathcal{L}}(t)$	Population risk in deployment environment e and empirical training risk.
t_{sc}, t_{core}, t_{edge}	Shortcut, core, and slow spectral-edge relaxation timescales.
$\chi_{\mathrm{reuse}}(t)$, $t_{\mathrm{reuse}}^{(\varepsilon)}$	Reuse-induced change in Q_{out} and the first time $\Delta\chi_{\mathrm{reuse}}(t)/\mathcal{L}_{\mathrm{on}}(t) > \varepsilon$; see (13).
Architecture and persistent-SGD extensions	
N, $\omega = N/D$, L	Hidden width, hidden-width ratio, and network depth.
$W_\ell(t)$, γ_0	Layer-ℓ weight matrix and architecture scaling used in the factorized linear predictor.
$h_\ell(t)$, $g_\ell(t)$	Forward and backward hidden fields at layer ℓ.
$A_{a\ell}(t)$, f_a	Normalized activity mask of example (a, ℓ) and its mini-batch activity fraction. Indexed $A_{a\ell}(t)$ is distinct from the label moment A; the active residual $Z_a(t) = A_a(t)\Delta_a(t)$ enters the gradient.
τ_a, γ_a, $\Phi_a(t, s)$	Mask persistence time, one-step persistence, and two-time activity kernel.

B FIGURE PARAMETERS

Parameters are listed in figure order, with one entry per figure. Here T is the number of gradient steps and n the number of independently sampled training datasets.

Fig.	Parameter protocol
Fig. 1	Not applicable.
Fig. 2	Shallow full-batch GD; $\mu = 0, q_* = 1, \Delta = 0.1, \rho = 0.75, \nu = 128, \alpha = 4, D = 128, \eta = 10^{-3}, \lambda_{\mathrm{L2}} = 0, q(0) = 0, T = 10^6, n = 24$. Reuse threshold $\varepsilon = 0.02$.
Fig. 3	$\Delta = 0.1, q_* = 1, \eta = 10^{-3}, \lambda_{\mathrm{L2}} = 0, q(0) = 0, T = 10^6, \varepsilon = 0.02$. Panel (a): $\alpha = 16, \nu = 0.02, \mu = 2, \kappa = 0, \rho = 1/2, D = 512, n = 24$, giving $A = 0$ and $B = \sqrt{2/\pi}$. Panels (b–d) use $\mu = 0$, $(\alpha, \nu) = (44, 128), (2, 128), (0.125, 0.0555)$ and $(D, n) = (128, 24), (128, 24), (512, 128)$, respectively; $(A, B, \kappa) \simeq (0.5, 0.6356, 0.2133)$. Simulations use $P/2$ examples per component; error bars are sample standard deviations.
Fig. 4	$\mu = 0, q_* = 1, \Delta = 0.1, \rho = 0.75, \nu = 128, \eta = 10^{-3}, q(0) = 0$. Panels (a,b): $\alpha \in \{0.25, 1, 4, 16\}$, $\lambda_{\mathrm{L2}} = 0, D = 128, T = 10^6, n = 24$. Panel (c): large-time predictions on 100 log-spaced $\alpha \in [0.25, 16]$, with $\lambda_{\mathrm{L2}} \in \{0, 10^{-3}, 10^{-2}\}$, compared with exact finite-dataset calculations.
Fig. 5	Full-batch GD; $\Delta = q_* = 1, \nu = 128, \mu = 3$, $(\kappa_1, \kappa_2) = (3, 5)$, $\eta = 10^{-3}, \lambda_{\mathrm{L2}} = 0, q(0) = 0$. Panel (a): $\alpha \in \{1.25, 2\}$; (b): $\alpha = 1.25$; (c): converged risk difference versus α. Evaluation uses each problem's own B, with $A_e = 0, B_e = B, q_y = 1$. The analytical crossing is $\alpha_c \simeq 1.534874$.
Fig. 6	*(a,b)* Waterbirds/Spawrious, full-image Gaussian projections of 64×64 images to $D = 512$; group-stratified calibration/evaluation/training-pool split $15/15/70\%$. Zero-initialized, no-intercept square-loss linear predictor, full-batch GD, $\eta = 0.0016520, T = 10^6$, no ridge or momentum; $\alpha \in \{0.5, 4\}$, $p_{\mathrm{train}} \in \{0.50, 0.95\}$, $n = 5$. Calibration and Gaussian moment approximation: App. E.3.1. *(c,d)* Colored MNIST, BatchNorm CNN, cross-entropy and full-batch GD; $\eta = 6.25 \times 10^{-4}, T = 1.6 \times 10^6$, no momentum or weight decay, three paired seeds. $P \in \{150, 300, 600, 1500, 3000, 6000\}$, balanced target classes; 5010 aligned and 4990 conflicting held-out examples. Panel (c) uses $p_{\mathrm{train}} = 0.80$; (d) pools $p_{\mathrm{train}} \in \{0.60, 0.70, 0.80, 0.90, 0.95, 0.99\}$. Reference $P_{\mathrm{ref}} = 6000$, excluded from the fit; persistent 10% departure threshold defined in App. E.3.2.
Fig. 7	$\mu = 0, q_* = 1, \Delta = 0.1, \rho = 0.75, \nu = 4, \kappa \simeq 0.2133, D = 512, \alpha = 2, \eta = 0.02, q(0) = 0$. A fresh batch of $P = \alpha D$ examples per step; 120 online steps and 12 independent trajectories.
Fig. 8	$D = 128, T = 10^6, n = 24, \varepsilon = 0.02$; $\alpha \in \{0.25, 0.5, 1, 2, 4, 8, 16, 32\}$. Baseline $\nu = 128, \rho = 0.75, \Delta = 0.1, \eta = 10^{-3}$. One parameter is varied at a time over $\nu \in \{32, 128, 512\}$, $\rho \in \{0.6, 0.75, 0.9\}$, $\Delta \in \{0.05, 0.1, 0.2\}$, or $\eta \in \{5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}\}$. Panel (a) uses the baseline; (b) includes the parameter variations.
Fig. 9	Auxiliary targets of (189); $\Delta = q_* = q_y = 1, \nu = 4, B = 0.5, (A_1, A_2) = (0.2, 0.7), \mu = \kappa = 0$. Full-batch GD; $D = 512, n = 64$ paired datasets, $\eta = 10^{-2}, \lambda_{\mathrm{L2}} = 0, q(0) = 0$. Panels (a,b): fixed q_y, with $\alpha \in \{1.5, 3\}$ in (a) and $\alpha = 1.5$ in (b). Panel (c): converged comparison at fixed q_y, and the fixed-noise control $\sigma^2 = q_y - A_1^2 - B^2/q_* = 0.71$. Both predictors use the common test law (190); error bars are paired 95% bootstrap intervals.
Fig. 10	Dense/adaptive shallow-DMFT comparison; $T = 12000, \eta = 10^{-3}$. Panel (a): $\alpha = 1/8, \epsilon_{\mathrm{rel}} = 2.5 \times 10^{-3}, \epsilon_{\mathrm{abs}} = 2.5 \times 10^{-7}$. Panel (b): eleven sample ratios in $[1/64, 1]$ and $\epsilon_{\mathrm{rel}} \in \{0.02, 0.01, 0.005, 0.0025\}$; the ordinate is the maximum population-risk error over time.

Fig.	Parameter protocol
Fig. 11	Shallow predictor and networks with one or three hidden layers, under full-batch GD and persistent SGD. $D = N = 128, \alpha = 4, \nu = 128, \Delta = 0.1, q_* = 1, \rho = 0.75, \kappa \simeq 0.2133, \gamma_0 = 1$; 64 finite simulations, $\eta = 10^{-3}$, physical time $\eta t \leq 4$, DMFT quadrature spacing 2.5×10^{-3}. The reported initialization is Gaussian input/hidden factors and a zero final readout, with $q(0) = 0$. Persistent SGD: active fraction $f = 0.25$, persistence $\eta\tau = 0.1$.
Fig. 12	Separate sweeps of $\alpha \in \{0.25, 1, 4, 16\}$, $\rho \in \{0.55, 0.75, 0.95\}$, and $\nu \in \{4, 32, 512\}$. Baseline $(\alpha, \rho, \nu, \Delta, q_*) = (4, 0.75, 128, 0.1, 1)$, $\eta = 10^{-3}, T = 10^6, n = 24$.
Fig. 13	Not applicable.
Fig. 14	Colored MNIST protocol of Fig. 6(c,d), restricted to $p_{\text{train}} = 0.80$: $P \in \{150, 300, 600, 1500, 3000, 6000\}$, three paired seeds, $\eta = 6.25 \times 10^{-4}, T = 1.6 \times 10^6$. Relative condition-balanced-risk departure from $P_{\text{ref}} = 6000$; persistent 10% threshold, with smoothing and denominator floor specified in App. E.3.2. The zero reference self-departure is not plotted.

C Dynamical Mean-Field Theory: Derivation and Extensions

DMFT describes the effects of reuse in random-feature models and deep linear networks (Bordelon et al., 2024; Bordelon & Pehlevan, 2025). Here we derive the data module for Gaussian mixtures with distinct core and shortcut directions. We use the Martin–Siggia–Rose (MSR) generating functional (Martin et al., 1973; Janssen, 1976) in the modular formulation of Bordelon & Pehlevan (2025). The data distribution, architecture, and optimizer contribute separately to the action, distinguishing memory from data reuse from the responses generated by the network architecture.

We give the derivation for the common data module and the shallow predictor, then collect the closed equations in App. C.3. App. C.4 states the changes needed for factorized linear architectures and persistent sampling, with their equations collected in separate summaries.

Conditioning on the shortcut sign gives two data components, $a = s \in \{-1, +1\}$. We impose the residual and gradient constraints, average the centered Gaussian inputs, and identify the causal responses and effective Gaussian fields. For the shallow predictor, these fields can be eliminated to obtain deterministic correlation–response equations.

C.1 Common Reused-Data Module

Conditioning the model of (1) on $s = a \in \{-1, +1\}$ gives a Gaussian component with centroid $v_a = av$, shared core direction $w_a^* = w^*$, and P_a reused examples, with $P_a/P \to 1/2$ in the proportional limit. Write its centered input as

$$\xi_{a\ell} = \sqrt{\Delta}\, z_{\perp,a\ell} + \sqrt{\Delta + \mu}\, \frac{G_{a\ell}}{q_*} \frac{w^*}{\sqrt{D}}, \qquad \Sigma_0 = \mathbb{E}[\xi_{a\ell}\xi_{a\ell}^\top] = \Delta I_D + \frac{\mu}{q_*} \frac{w^* w^{*\top}}{D}, \tag{18}$$

where $G_{a\ell} \sim \mathcal{N}(0, q_*)$, and $z_{\perp,a\ell} \sim \mathcal{N}(0, I_D - w^* w^{*\top}/(Dq_*))$ is independent of $G_{a\ell}$. Hence $x_{a\ell} = v_a/\sqrt{D} + \xi_{a\ell}$. Eq. (18) retains arbitrary $\mu \geq 0$ and arbitrary shortcut–core overlap κ, subject to $\kappa^2 < \nu q_*$.

For one training example define the centered prediction field

$$U_{a\ell}(t) = \frac{\xi_{a\ell}^\top q(t)}{\sqrt{D}}, \qquad M_a(t) = \frac{q(t)^\top v_a}{D} = a\, m(t), \qquad \Theta_a(t) = \frac{q(t)^\top w_a^*}{D} = \theta(t).$$

The label is

$$Y_{a\ell} = \operatorname{sign}\Big(a\kappa + \sqrt{\Delta + \mu}\, G_{a\ell}\Big).$$

The residual and centered component gradient obey

$$\Delta_{a\ell}(t) = Y_{a\ell} - M_a(t) - U_{a\ell}(t), \tag{19}$$

$$h_{ai}(t) = \frac{\sqrt{D}}{P_a} \sum_{\ell=1}^{P_a} \xi_{a\ell i} \Delta_{a\ell}(t). \tag{20}$$

The centroid contribution is added after the centered-design average.

The Gaussian cavity field u_a, obtained by omitting the example from training, has covariance

$$K_q(t,s) := \mathbb{E}[u_a(t)u_a(s)] = \Delta C_q(t,s) + \frac{\mu}{q_*}\Theta_a(t)\Theta_a(s), \qquad C_q(t,s) = \frac{q(t)^\top q(s)}{D}, \tag{21}$$

and $\mathbb{E}[G_a u_a(t)] = \sqrt{\Delta + \mu}\,\Theta_a(t)$. Unlike u_a, the training field U_a also contains the memory term in (39). The label quantities needed by square loss are

$$\alpha_a^{\mathrm{y}} = \mathbb{E}[Y_a \mid s = a] = aA, \qquad \beta_a^{\mathrm{y}} = \frac{\sqrt{\Delta + \mu}}{q_*}\mathbb{E}[Y_a G_a \mid s = a], \qquad q_y = \mathbb{E}[Y_a^2] = 1. \tag{22}$$

Evaluating these expectations gives

$$A = \operatorname{erf}\left(\frac{\kappa}{\sqrt{2(\Delta + \mu)q_*}}\right), \qquad B = \sqrt{\frac{2q_*}{\pi}} \exp\left[-\frac{\kappa^2}{2(\Delta + \mu)q_*}\right],$$

$$\beta_a^{\mathrm{y}} = \sqrt{\frac{2(\Delta + \mu)}{\pi q_*}} \exp\left[-\frac{\kappa^2}{2(\Delta + \mu)q_*}\right]. \tag{23}$$

These moments are independent of a. We write $C^a_\Delta(t,s) = \langle \Delta_a(t)\Delta_a(s)\rangle$ for the effective residual correlation.

Step 1: write the component generating functional. In the MSR representation, $q(t)$, $\Delta_{a\ell}(t)$, and $h_a(t)$ are integration variables. Using the Fourier convention $\delta(F) = \int \mathrm{d}\widehat{F}\, e^{\mathrm{i}\widehat{F}F}/(2\pi)$, we impose (19)–(20) with conjugate variables $\widehat{\Delta}_{a\ell}$ and $\widehat{h}_{ai}$. The full generating functional also contains the architecture update and probe sources used to generate observables. Suppressing those terms and normalization constants, the component action is

$$\mathcal{Z}_a = \mathbb{E}_{\{\xi_{a\ell}, G_{a\ell}\}} \int \prod_{\ell,t} \frac{\mathrm{d}\widehat{\Delta}_{a\ell}(t)}{2\pi} \prod_{i,t} \frac{\mathrm{d}\widehat{h}_{ai}(t)}{2\pi} \exp\{\mathcal{S}_a[\xi_a]\}, \tag{24}$$

$$\begin{aligned}\mathcal{S}_a[\xi_a] = \mathrm{i}\sum_{\ell,t} \widehat{\Delta}_{a\ell}(t) \left[\Delta_{a\ell}(t) - Y_{a\ell} + M_a(t) + \frac{\xi_{a\ell}^\top q(t)}{\sqrt{D}}\right] \\ + \mathrm{i}\sum_{i,t} \widehat{h}_{ai}(t) \left[h_{ai}(t) - \frac{\sqrt{D}}{P_a}\sum_\ell \xi_{a\ell i}\Delta_{a\ell}(t)\right].\end{aligned} \tag{25}$$

Eq. (25) is the component contribution to the dynamical generating functional. The expectation retains the joint law of the centered input and the core field because the latter also determines the label.

Step 2: average the centered Gaussian design. To retain the dependence of $Y_{a\ell}$ on the core field, introduce $\widehat{G}_{a\ell}$ conjugate to $G_{a\ell} - w^{*\top}\xi_{a\ell}/\sqrt{D(\Delta+\mu)}$. The centered-design average gives

$$\overline{\mathcal{Z}}_a = \int \mathcal{D}G_a\, \mathcal{D}\widehat{G}_a\, \mathcal{D}\widehat{\Delta}_a\, \mathcal{D}\widehat{h}_a\, \exp\left\{\mathcal{S}_a^{(0)} + \mathcal{S}_{\widehat{\Delta}\widehat{\Delta}} + \mathcal{S}_{\widehat{h}\widehat{h}} + \mathcal{S}_{\widehat{\Delta}\widehat{h}} + \mathcal{S}_G\right\}, \tag{26}$$

$$\mathcal{S}_a^{(0)} = \mathrm{i}\sum_{\ell,t} \widehat{\Delta}_{a\ell}(t)\left[\Delta_{a\ell}(t) - Y_{a\ell} + M_a(t)\right] + \mathrm{i}\sum_{i,t}\widehat{h}_{ai}(t)h_{ai}(t),$$

$$\mathcal{S}_{\widehat{\Delta}\widehat{\Delta}} = -\frac{1}{2}\sum_{\ell,t,s} \widehat{\Delta}_{a\ell}(t)K_q(t,s)\widehat{\Delta}_{a\ell}(s), \tag{27}$$

$$\mathcal{S}_{\widehat{h}\widehat{h}} = -\frac{1}{\alpha}\sum_{i,j,t,s} \widehat{h}_{ai}(t)(\Sigma_0)_{ij}\widehat{h}_{aj}(s)C^a_\Delta(t,s), \tag{28}$$

$$\mathcal{S}_{\widehat{\Delta}\widehat{h}} = \frac{1}{P_a}\sum_{\ell,t,s} \widehat{\Delta}_{a\ell}(t)\Delta_{a\ell}(s)q(t)^\top\Sigma_0\widehat{h}_a(s). \tag{29}$$

$$\begin{aligned}\mathcal{S}_G = \sum_\ell \left(\mathrm{i}\widehat{G}_{a\ell}G_{a\ell} - \frac{q_*}{2}\widehat{G}^2_{a\ell}\right) \\ + \sqrt{\Delta+\mu}\sum_{\ell,t}\widehat{G}_{a\ell}\left[\widehat{\Delta}_{a\ell}(t)\Theta_a(t) - \frac{\Delta_{a\ell}(t)}{P_a}w^{*\top}\widehat{h}_a(t)\right].\end{aligned} \tag{30}$$

Here

$$C^a_\Delta(t,s) = \frac{1}{P_a}\sum_\ell \Delta_{a\ell}(t)\Delta_{a\ell}(s), \qquad q(t)^\top\Sigma_0 q(s)/D = K_q(t,s).$$

In (26), the first two quadratic terms become the prediction and gradient Gaussian fields. In the mixed term,

$$q(t)^\top\Sigma_0\widehat{h}_a(s) = \Delta q(t)^\top\widehat{h}_a(s) + \frac{\mu}{q_*}\Theta_a(t)w^{*\top}\widehat{h}_a(s).$$

The extensive isotropic part produces the scalar Onsager memory. The finite-rank part remains in the core projection and produces the additional core-memory contraction derived in step 5.

Step 3: introduce the response through an update functional. To interpret the mixed term, the network update must be included. Write the architecture-dependent update abstractly as $q(t+1) =$

$\Phi_t(q(0:t), h_0(0:t))$. Add a source to the raw component field, $h_a(t) \mapsto h_a(t) + J_a(t)$. Since $h_0 = \frac{1}{2}\sum_{a=\pm1}[B_a v_a + h_a]$, this is equivalent to

$$h_0(t) \longmapsto h_0(t) + \frac{1}{2}J_a(t),$$

and the factor $1/2$ fixes the response normalization. In particular, it produces the direct $\delta_{ts}/2$ term in (65). Impose the update by

$$\mathcal{U}_q[J_a] = \int \mathcal{D}q\,\mathcal{D}\widehat{q}\,\exp\left\{\mathrm{i}\sum_{i,t}\widehat{q}_i(t+1)\left[q_i(t+1) - \Phi_{t,i}\big(q(0:t), h_0(0:t) + J_a(0:t)/2\big)\right]\right\}. \tag{31}$$

The component response is then the derivative of the student trajectory with respect to this source after setting $J_a = 0$:

$$R^a_{qh}(t,s) = \frac{1}{D}\sum_i \left\langle \frac{\delta q_i(t)}{\delta J_{a,i}(s)} \right\rangle \Bigg|_{J_a=0}. \tag{32}$$

Shifting $h'_a = h_a + J_a$ removes the source from the update and leaves $-\mathrm{i}\sum_{i,s}\widehat{h}_{ai}(s)J_{a,i}(s)$ in the exponent. Differentiation gives the saddle identity

$$R^a_{qh}(t,s) = -\frac{\mathrm{i}}{D}\sum_i \left\langle q_i(t)\widehat{h}_{ai}(s)\right\rangle. \tag{33}$$

Causality of the update implies $R^a_{qh}(t,s) = 0$ for $s \geq t$ in the strict gradient-update memory. Substituting (33) into the isotropic part of (29) gives

$$\mathcal{S}^{\mathrm{iso}}_{\widehat{\Delta}\widehat{h}} = +\mathrm{i}\frac{2\Delta}{\alpha}\sum_{\ell,t,s<t}\widehat{\Delta}_{a\ell}(t)R^a_{qh}(t,s)\Delta_{a\ell}(s). \tag{34}$$

Thus the residual constraint in the averaged functional becomes

$$\Delta_{a\ell}(t) - Y_{a\ell} + M_a(t) + \frac{2\Delta}{\alpha}\sum_{s<t}R^a_{qh}(t,s)\Delta_{a\ell}(s) + \text{Gaussian cavity field} = 0. \tag{35}$$

Only the normalized trace of Σ_0 enters this scalar Onsager term; $D^{-1}\operatorname{Tr}\Sigma_0 \to \Delta$. The rank-one part has zero normalized trace but a finite projection on w^*. It is therefore absent from (34) and retained explicitly in the core drift below.

Step 4: write the effective functional and identify the noises. Applying a joint Hubbard–Stratonovich transform to the core and prediction fields gives, for one representative example,

$$\mathcal{Z}_{\Delta,a} = \mathbb{E}_{G_a,u_a}\int \mathcal{D}\Delta_a\,\mathcal{D}\widehat{\Delta}_a\,\exp\left\{\mathrm{i}\sum_t \widehat{\Delta}_a(t)\left[\Delta_a(t) - Y_a(G_a) + M_a(t) + u_a(t) + \frac{2\Delta}{\alpha}\sum_{s<t}R^a_{qh}(t,s)\Delta_a(s)\right]\right\}, \tag{36}$$

where (G_a, u_a) is centered jointly Gaussian with

$$\mathbb{E}[G_a^2] = q_*, \qquad \mathbb{E}[G_a u_a(t)] = \sqrt{\Delta+\mu}\,\Theta_a(t), \qquad \mathbb{E}[u_a(t)u_a(s)] = K_q(t,s). \tag{37}$$

Integrating over $\widehat{\Delta}$ gives the scalar residual process

$$\Delta_a(t) = Y_a - M_a(t) - u_a(t) - \frac{2\Delta}{\alpha}\sum_{s<t}R^a_{qh}(t,s)\Delta_a(s), \qquad Y_a = \operatorname{sign}\Big(a\kappa + \sqrt{\Delta+\mu}\,G_a\Big), \tag{38}$$

which is the residual equation later collected as (48). Comparing with (19) gives the training field

$$U_a(t) = u_a(t) + \frac{2\Delta}{\alpha}\sum_{s<t}R^a_{qh}(t,s)\Delta_a(s). \tag{39}$$

The quadratic gradient-noise sector is analogous:

$$\mathcal{Z}_{h,a} = \int \mathcal{D}h_a\, \mathcal{D}\widehat{h}_a\, \exp\left\{ \mathrm{i} \sum_{i,t} \widehat{h}_{ai}(t) h_{ai}(t) - \frac{1}{\alpha} \sum_{i,j,t,s} \widehat{h}_{ai}(t)(\Sigma_0)_{ij} C^a_\Delta(t,s) \widehat{h}_{aj}(s) \right\}. \tag{40}$$

Linearizing with $\mathbb{E}\exp[-\mathrm{i}\sum_{i,t} \widehat{h}_{ai}(t) u^0_{a,i}(t)]$ identifies a centered vector Gaussian process

$$\mathbb{E}[u^0_{a,i}(t) u^0_{a,j}(s)] = \frac{2(\Sigma_0)_{ij}}{\alpha} C^a_\Delta(t,s). \tag{41}$$

Its rank-one contribution has vanishing normalized trace and vanishing macroscopic overlap with either fixed signal direction. Consequently, the closed norm and response equations retain $D^{-1}\mathbb{E}[u^0_a(t)^\top u^0_a(s)] = 2\Delta C^a_\Delta(t,s)/\alpha$. The deterministic rank-one drift remains finite and is computed next.

Step 5: compute the gradient drift. Let $F_t(Y_a, u_a(0:t)) = \Delta_a(t)$ denote the scalar residual process in (38). The coefficient $L_a(t)$ of the label Y_a in $\Delta_a(t)$ obeys the triangular recursion

$$L_a(t) = 1 - \frac{2\Delta}{\alpha} \sum_{s<t} R^a_{qh}(t,s) L_a(s), \qquad L_a(0) = 1. \tag{42}$$

Next, define the residual response

$$R^a_\Delta(t,s) = \left\langle \frac{\partial \Delta_a(t)}{\partial u_a(s)} \right\rangle, \qquad R^a_\Delta(t,t) = -1, \tag{43}$$

where the equal-time value follows directly from the term $-u_a(t)$ in (38). Gaussian integration by parts for the centered input gives, coordinate by coordinate,

$$\sqrt{D}\,\mathbb{E}[\xi_{a,i}\Delta_a(t)] = \beta^{\mathrm{y}}_a L_a(t) w^*_i + \sum_{s\le t} R^a_\Delta(t,s) \left[\Delta q_i(s) + \frac{\mu}{q_*} \Theta_a(s) w^*_i \right]. \tag{44}$$

The second term in (44) is the student self-interaction generated by the centered Gaussian. Its equal-time part is

$$R^a_\Delta(t,t)\Sigma_0 q(t) = -\Sigma_0 q(t).$$

This is the damping of the current predictor by the input covariance. The remaining $s < t$ terms describe its retarded self-interaction.

The label term does not require differentiating the sign function. Eq. (22) gives

$$\sqrt{D}\,\mathbb{E}[\xi_{a,i} Y_a \mid s = a] = \beta^{\mathrm{y}}_a w^*_i,$$

and since $L_a(t)$ is the coefficient of Y_a in the residual, the label contribution is $\beta^{\mathrm{y}}_a L_a(t) w^\star$. Define the core-memory contraction and the resulting coefficient along w^* by

$$\mathcal{T}_a(t) = \sum_{s\le t} R^a_\Delta(t,s)\Theta_a(s), \tag{45}$$

$$\zeta_a(t) = \beta^{\mathrm{y}}_a L_a(t) + \frac{\mu}{q_*} \mathcal{T}_a(t). \tag{46}$$

Combining the label drift, the covariance response, and the noise from (41) yields

$$h_a(t) = \zeta_a(t) w^* + u^0_a(t) + \Delta \sum_{s\le t} R^a_\Delta(t,s) q(s). \tag{47}$$

Eq. (47) is the centered-gradient process collected below in (49).

The conditional centroid $v_a = av$ contributes in two places. First, it adds the deterministic student field

$$M_a(t) = \frac{q(t)^\top v_a}{D}$$

to the residual, already present in (38). Second, it gives a deterministic contribution to the empirical gradient:

$$\frac{\sqrt{D}}{P} X_a^\top \Delta_a(t) \supset \frac{1}{P} v_a \sum_{\ell \in a} \Delta_{a\ell}(t) = \frac{1}{2} B_a(t) v_a, \qquad B_a(t) = \frac{1}{P_a} \sum_{\ell \in a} \Delta_{a\ell}(t).$$

Eq. (47) gives the centered field $h_a(t)$. Adding the two component contributions with equal weight forms the total input-gradient field $h_0(t)$. The final data block is

$$\Delta_a(t) = Y_a - M_a(t) - u_a(t) - \frac{2\Delta}{\alpha} \sum_{s<t} R^a_{qh}(t,s) \Delta_a(s), \tag{48}$$

$$h_a(t) = \zeta_a(t) w^* + u^0_a(t) + \Delta \sum_{s \le t} R^a_\Delta(t,s) q(s), \tag{49}$$

$$h_0(t) = \frac{1}{2} \sum_{a=\pm 1} \left[B_a(t) v_a + h_a(t) \right]. \tag{50}$$

The effective noises are those identified in Step 4: u_a has covariance K_q and core covariance $\mathbb{E}[G_a u_a(t)] = \sqrt{\Delta + \mu}\, \Theta_a(t)$, while the normalized covariance of u^0_a is $2\Delta C^a_\Delta / \alpha$. Eqs. (48)–(50), together with (45)–(46), are valid for simultaneous $\mu > 0$ and $\kappa \neq 0$ and are used by all architectures below.

C.2 Shallow Linear Predictor

The spectral solution (155) gives the shallow trajectory for a fixed training set. DMFT averages over datasets to obtain closed equations for feature overlaps and correlations in the proportional limit. We close these equations below; expanding them around online learning gives the reuse correction derived in App. D.2.

For the shallow linear predictor, the finite-dimensional update is

$$q(t+1) = a_\lambda q(t) + \eta h_0(t), \qquad a_\lambda = 1 - \eta \lambda_{\mathrm{L2}}, \qquad h_0(t) = \frac{\sqrt{D}}{P} \sum_{a,\ell} x_{a\ell} \Delta_{a\ell}(t).$$

Summing the update gives the strictly retarded effective-process memory

$$q(t) = a^t_\lambda q_0 + \eta \sum_{s<t} a^{t-1-s}_\lambda h_0(s). \tag{51}$$

Substituting the reused shortcut–core data block from App. C.1 closes the shallow linear predictor. The risk formula follows from the Gaussian identities

$$\mathbb{E}_a \left[y \frac{q^\top x}{\sqrt{D}} \right] = \alpha^{\mathrm{y}}_a M_a + \beta^{\mathrm{y}}_a \Theta_a,$$

and

$$\mathbb{E}_a \left[\left(\frac{q^\top x}{\sqrt{D}} \right)^2 \right] = M_a^2 + \Delta C_q(t,t) + \frac{\mu}{q_*} \Theta_a(t)^2.$$

Closed CHSCK equations for the shallow model. For the shallow architecture one can eliminate the effective processes and write deterministic causal equations for the order parameters, substantially reducing the computational cost. We first display the exact fixed-step recursion used by the solver. Replacing its causal sums by integrals gives the continuous-time limit used for the analytical discussion. We refer to these closed correlation–response equations as the Crisanti–Horner–Sommers–Cugliandolo–Kurchan (CHSCK) equations (Sarao Mannelli et al., 2020; Helias & Dahmen, 2020). The summary collects the complete shallow system for arbitrary $\mu \geq 0$, arbitrary $\kappa^2 < \nu q_*$, and ridge strength λ_{L2}.

C.3 DMFT EQUATIONS SUMMARY FOR THE SHALLOW LINEAR ESTIMATOR

Fixed quantities and moments

$$a, b \in \{-1, +1\}, \qquad \frac{P_a}{P} \to \frac{1}{2}, \qquad v_a = av, \qquad w_a^* = w^*, \qquad a_\lambda = 1 - \eta\lambda_{\mathrm{L2}}.$$

$$M_a(t) = \frac{q(t)^\top v_a}{D}, \qquad \Theta_a(t) = \frac{q(t)^\top w_a^*}{D}, \qquad C_q(t,s) = \frac{q(t)^\top q(s)}{D}.$$

$$V_{ab} = \frac{v_a^\top v_b}{D} = ab\nu, \qquad A_{ab} = \frac{v_a^\top w_b^*}{D} = a\kappa, \qquad T_{ab} = \frac{w_a^{*\top} w_b^*}{D} = q_*. \tag{52}$$

$$\alpha_a^{\mathrm{y}} = \mathbb{E}[Y_a \mid s = a], \qquad \beta_a^{\mathrm{y}} = \frac{\sqrt{\Delta+\mu}}{q_*}\mathbb{E}[Y_a G_a \mid s = a], \qquad q_y = \mathbb{E}[Y_a^2] = 1, \qquad G_a \sim \mathcal{N}(0, q_*).$$

$$Y_a = \operatorname{sign}\Big(a\kappa + \sqrt{\Delta+\mu}\, G_a\Big), \qquad q_y = 1,$$

$$\alpha_a^{\mathrm{y}} = a\ \operatorname{erf}\left(\frac{\kappa}{\sqrt{2(\Delta+\mu)q_*}}\right), \qquad \beta_a^{\mathrm{y}} = \sqrt{\frac{2(\Delta+\mu)}{\pi q_*}}\, e^{-\kappa^2/[2(\Delta+\mu)q_*]}.$$

Initialization and row order

$$M_a(0) = \Theta_a(0) = 0, \qquad C_q(0,0) = \sigma_q^2, \qquad R_{qh}^a(0,s) = 0, \qquad R_{qh}^a(t,s) = 0 \ \ (s \geq t).$$

$$(R_\Delta, L, B, \Gamma, \mathcal{T}, \zeta) \ \longrightarrow\ (\Upsilon, C_\Delta, \Xi) \ \longrightarrow\ (C_{hq}, C_{hh}, D) \ \longrightarrow\ (M, \Theta, C_q, R_{qh}).$$

Effective process

$$K_q(t,s) = \Delta C_q(t,s) + \frac{\mu}{q_*}\Theta_a(t)\Theta_a(s),$$

$$\mathcal{T}_a(t) = \sum_{r \leq t} R_\Delta^a(t,r)\Theta_a(r), \qquad \zeta_a(t) = \beta_a^{\mathrm{y}} L_a(t) + \frac{\mu}{q_*}\mathcal{T}_a(t). \tag{53}$$

$$\Delta_a(t) = Y_a - M_a(t) - u_a(t) - \frac{2\Delta}{\alpha}\sum_{r<t} R_{qh}^a(t,r)\Delta_a(r),$$

$$h_a(t) = \zeta_a(t)w^* + u_a^0(t) + \Delta\sum_{r \leq t} R_\Delta^a(t,r)q(r), \qquad h_0(t) = \frac{1}{2}\sum_{a=\pm 1}[B_a(t)v_a + h_a(t)], \tag{54}$$

$$\mathbb{E}[u_a(t)u_a(s)] = K_q(t,s), \qquad \mathbb{E}[G_a u_a(t)] = \sqrt{\Delta+\mu}\,\Theta_a(t),$$

$$\frac{1}{D}\mathbb{E}[u_a^0(t)^\top u_b^0(s)] = \delta_{ab}\frac{2\Delta}{\alpha}C_\Delta^a(t,s), \qquad C_\Delta^a(t,s) = \langle \Delta_a(t)\Delta_a(s)\rangle.$$

Causal residual row

$$R_\Delta^a(t,s) = -\delta_{ts} - \frac{2\Delta}{\alpha}\sum_{r<t} R_{qh}^a(t,r)R_\Delta^a(r,s), \tag{55}$$

$$L_a(t) = 1 - \frac{2\Delta}{\alpha}\sum_{r<t} R_{qh}^a(t,r)L_a(r), \tag{56}$$

$$B_a(t) = \alpha_a^{\mathrm{y}} - M_a(t) - \frac{2\Delta}{\alpha}\sum_{r<t} R_{qh}^a(t,r)B_a(r), \tag{57}$$

$$\Gamma_a(t) = q_y - \alpha_a^{\mathrm{y}} M_a(t) - \beta_a^{\mathrm{y}}\Theta_a(t) - \frac{2\Delta}{\alpha}\sum_{r<t} R_{qh}^a(t,r)\Gamma_a(r),$$

$$B_a(t) = \langle \Delta_a(t)\rangle, \qquad \Gamma_a(t) = \langle Y_a\Delta_a(t)\rangle. \tag{58}$$

$$\Upsilon_a(t,s) = \langle u_a(t)\Delta_a(s)\rangle = \Theta_a(t)\zeta_a(s) + \Delta\sum_{r \leq s} R_\Delta^a(s,r)C_q(t,r). \tag{59}$$

$$C_\Delta^a(t,s) = \Gamma_a(s) - M_a(t)B_a(s) - \Upsilon_a(t,s) - \frac{2\Delta}{\alpha}\sum_{r<t} R_{qh}^a(t,r)C_\Delta^a(r,s), \qquad s \leq t, \tag{60}$$

$$C_\Delta^a(s,t) = C_\Delta^a(t,s).$$

$$\Xi_a(t,s) = \frac{1}{D}\langle u_a^0(t)^\top q(s)\rangle = \frac{2\Delta}{\alpha}\sum_{r<s} C_\Delta^a(t,r) R_{qh}^a(s,r). \tag{61}$$

Gradient correlations

$$\begin{aligned} C_{hq}(t,s) &= \frac{1}{D}\langle h_0(t)^\top q(s)\rangle \\ &= \frac{1}{2}\sum_{a=\pm 1}\left[B_a(t)M_a(s) + \zeta_a(t)\Theta_a(s) + \Xi_a(t,s) + \Delta\sum_{r\le t} R_\Delta^a(t,r)C_q(r,s)\right]. \end{aligned} \tag{62}$$

$$\Psi_a(t,s) = B_a(t)M_a(s) + \zeta_a(t)\Theta_a(s), \qquad C_{hh}(t,s) = \frac{1}{D}\langle h_0(t)^\top h_0(s)\rangle.$$

$$\begin{aligned} C_{hh}(t,s) = \frac{1}{4}\sum_{a,b=\pm 1}\Bigg\{ &\Omega_{ab}(t,s) + \Delta\sum_{\ell\le s} R_\Delta^b(s,\ell)\Psi_a(t,\ell) \\ &+ \Delta\sum_{r\le t} R_\Delta^a(t,r)\Psi_b(s,r) + \Delta^2\sum_{r\le t}\sum_{\ell\le s} R_\Delta^a(t,r)R_\Delta^b(s,\ell)C_q(r,\ell) \\ &+ \delta_{ab}\frac{2\Delta}{\alpha}C_\Delta^a(t,s) + \Delta\sum_{\ell\le s} R_\Delta^b(s,\ell)\Xi_a(t,\ell) \\ &+ \Delta\sum_{r\le t} R_\Delta^a(t,r)\Xi_b(s,r)\Bigg\}, \end{aligned} \tag{63}$$

$$\begin{aligned} \Omega_{ab}(t,s) = B_a(t)B_b(s)V_{ab} + B_a(t)\zeta_b(s)A_{ab} + \zeta_a(t)B_b(s)A_{ba} \\ + \zeta_a(t)\zeta_b(s)T_{ab}. \end{aligned} \tag{64}$$

State and response updates

$$D_b(t,s) = \frac{1}{D}\left\langle \frac{\partial h_0(t)}{\partial u_b^0(s)}\right\rangle = \frac{1}{2}\delta_{ts} + \frac{\Delta}{2}\sum_{a=\pm 1}\sum_{r\le t} R_\Delta^a(t,r)R_{qh}^b(r,s). \tag{65}$$

$$M_b(t+1) = a_\lambda M_b(t) + \frac{\eta}{2}\sum_{a=\pm 1}\left[B_a(t)V_{ab} + \zeta_a(t)A_{ba} + \Delta\sum_{r\le t} R_\Delta^a(t,r)M_b(r)\right], \tag{66}$$

$$\Theta_b(t+1) = a_\lambda \Theta_b(t) + \frac{\eta}{2}\sum_{a=\pm 1}\left[B_a(t)A_{ab} + \zeta_a(t)T_{ab} + \Delta\sum_{r\le t} R_\Delta^a(t,r)\Theta_b(r)\right], \tag{67}$$

$$C_q(t+1,s) = a_\lambda C_q(t,s) + \eta C_{hq}(t,s), \qquad 0\le s\le t, \tag{68}$$

$$C_q(t+1,t+1) = a_\lambda^2 C_q(t,t) + 2a_\lambda\eta C_{hq}(t,t) + \eta^2 C_{hh}(t,t). \tag{69}$$

$$R_{qh}^b(t+1,s) = a_\lambda R_{qh}^b(t,s) + \eta D_b(t,s), \qquad R_{qh}^b(0,s) = 0, \tag{70}$$

Risks

$$\widehat{\mathcal{L}}(t) = \frac{1}{2}\sum_{a=\pm 1} C_\Delta^a(t,t), \tag{71}$$

$$\mathcal{L}(t) = \frac{1}{2}\sum_{a=\pm 1}\left[q_y - 2\alpha_a^{\mathrm{y}} M_a(t) - 2\beta_a^{\mathrm{y}}\Theta_a(t) + M_a(t)^2 + \Delta C_q(t,t) + \frac{\mu}{q_*}\Theta_a(t)^2\right]. \tag{72}$$

Setting $\mu = 0$ recovers the isotropic reused-data equations, while $\kappa = 0$ makes $A_{ab} = 0$ and decouples the shortcut and core directions. For simultaneous $\mu > 0$ and $\kappa \neq 0$, the same coefficient ζ_a enters both overlap equations and all low-rank terms in C_{hh}. No additional two-time kernel is required.

App. E.1.2 compares the accuracy and runtime of these closed equations with Monte Carlo solutions of the effective processes.

C.4 Extensions and generalisations

C.4.1 Architecture Modules

Following Bordelon & Pehlevan (2025), we describe a factorized linear predictor through coupled forward and backward fields instead of a direct update of q. Averaging the initial weight matrices adds Gaussian sources and their causal responses; the reused-data module is unchanged. We collect the resulting equations for L hidden layers below. The one-hidden-layer model is the case $L = 1$: there are no intermediate backward fields, and the output equations apply directly to layer one.

All hidden layers have width $N = \omega D$. The input matrix is $W_0 \in \mathbb{R}^{N\times D}$, the other matrices are $W_\ell \in \mathbb{R}^{N\times N}$, and their initial entries and those of the output vector $g_L(0)$ are independent standard Gaussians. The parameter γ_0 sets the output normalization and layer update scale. Here $\lambda_{\mathrm{L2}} = 0$; the data parameters μ and κ enter through the general data block in App. C.3. The finite network is specified by

$$h_1 = \frac{W_0 h_0}{\sqrt{D}}, \qquad h_{\ell+1} = \frac{W_\ell h_\ell}{\sqrt{N}}, \qquad g_\ell = \frac{W_\ell^\top g_{\ell+1}}{\sqrt{N}}, \qquad q = \frac{\sqrt{D}}{N\gamma_0} W_0^\top g_1,$$

$$W_0(t+1) = W_0(t) + \frac{\eta\gamma_0}{\sqrt{D}} g_1(t) h_0(t)^\top, \qquad W_\ell(t+1) = W_\ell(t) + \frac{\eta\gamma_0}{\sqrt{N}} g_{\ell+1}(t) h_\ell(t)^\top,$$

$$g_L(t+1) = g_L(t) + \eta\gamma_0 h_L(t), \qquad 1 \le \ell < L.$$

The normalized correlations are $C_h^0(t,s) = D^{-1}\langle h_0(t)^\top h_0(s)\rangle$, $C_h^\ell(t,s) = N^{-1}\langle h_\ell(t)^\top h_\ell(s)\rangle$, and $C_g^\ell(t,s) = N^{-1}\langle g_\ell(t)^\top g_\ell(s)\rangle$. The responses $R_{hu}^\ell, R_{gu}^\ell, R_{hr}^\ell, R_{gr}^\ell$ are the averaged single-coordinate derivatives of h_ℓ, g_ℓ with respect to their sources u_ℓ, r_ℓ. At the input boundary, R_{qr} and R_{hr}^0 are the normalized trace responses of q and h_0 to r_0. All responses vanish for $s > t$. For $g_L^0 = g_L(0)$, the static responses are $E_L(t) = \langle \partial g_L(t)/\partial g_L^0\rangle$ and $F_L(t) = \langle \partial h_L(t)/\partial g_L^0\rangle$. The input projections are $\Phi_b^v(t) = D^{-1}\langle h_0(t)^\top v_b\rangle$ and $\Phi_b^w(t) = D^{-1}\langle h_0(t)^\top w_b^*\rangle$.

In the summary, $Q(t,r)$ is an architecture kernel, not the predictor norm $Q(t) = C_q(t,t)$. The quantities $R_\Delta, L, B, \Gamma, C_\Delta, \Xi, C_{hq}, C_{hh}, D_a$ are given by the data-block equations in App. C.3, with $C_h^0 = C_{hh}$. The architecture equations replace only the shallow updates of M, Θ, C_q, R_{qh}. The correlations and responses are propagated causally; intermediate layers require a 2×2 linear solve at equal time.

DMFT equations summary for linear architectures

Initialization and effective fields

$$M_a(0) = \Theta_a(0) = 0, \qquad C_g^\ell(0,0) = 1, \qquad C_q(0,0) = \frac{1}{\omega\gamma_0^2}, \qquad E_L(0) = 1.$$

$$q(t) = r_0(t) + \sum_{r\le t} Q(t,r) h_0(r), \qquad r_0 \sim \mathcal{GP}\left(0, \frac{C_g^1}{\omega\gamma_0^2}\right), \tag{73}$$

$$h_\ell(t) = u_\ell(t) + \sum_{r\le t} H_\ell(t,r) g_\ell(r), \qquad u_\ell \sim \mathcal{GP}(0, C_h^{\ell-1}), \quad 1 \le \ell \le L, \tag{74}$$

$$g_\ell(t) = r_\ell(t) + \sum_{r\le t} G_\ell(t,r) h_\ell(r), \qquad r_\ell \sim \mathcal{GP}(0, C_g^{\ell+1}), \quad 1 \le \ell < L, \tag{75}$$

$$g_L(t) = g_L^0 + \eta\gamma_0 \sum_{r<t} h_L(r), \qquad g_L^0 \sim \mathcal{N}(0,1). \tag{76}$$

Architecture kernels

$$H_1(t,r) = \frac{R_{hr}^0(t,r)}{\omega\gamma_0} + \eta\gamma_0 \mathbf{1}_{\{r<t\}} C_h^0(t,r), \tag{77}$$

$$H_\ell(t,r) = R_{hr}^{\ell-1}(t,r) + \eta\gamma_0 \mathbf{1}_{\{r<t\}} C_h^{\ell-1}(t,r), \qquad 2 \le \ell \le L, \tag{78}$$

$$G_\ell(t,r) = R_{gu}^{\ell+1}(t,r) + \eta\gamma_0 \mathbf{1}_{\{r<t\}} C_g^{\ell+1}(t,r), \qquad 1 \le \ell < L, \tag{79}$$

$$Q(t,r) = \frac{R_{gu}^1(t,r)}{\gamma_0} + \eta \mathbf{1}_{\{r<t\}} C_g^1(t,r). \tag{80}$$

Input projections and responses

$$\Phi_b^v(t) = \frac{1}{2}\sum_{a=\pm 1}\left[B_a(t)V_{ab} + \zeta_a(t)A_{ba} + \Delta\sum_{r\le t} R_\Delta^a(t,r)M_b(r)\right], \tag{81}$$

$$\Phi_b^w(t) = \frac{1}{2}\sum_{a=\pm 1}\left[B_a(t)A_{ab} + \zeta_a(t)T_{ab} + \Delta\sum_{r\le t} R_\Delta^a(t,r)\Theta_b(r)\right], \tag{82}$$

$$R_{qr}(t,s) = \delta_{ts} + \sum_{r\le t} Q(t,r)R_{hr}^0(r,s), \tag{83}$$

$$R_{hr}^0(t,s) = \frac{\Delta}{2}\sum_{a=\pm 1}\sum_{r\le t} R_\Delta^a(t,r)R_{qr}(r,s), \tag{84}$$

$$R_{qh}^a(t,s) = \sum_{r\le t} Q(t,r)D_a(r,s). \tag{85}$$

Layer responses

$$R_{hu}^\ell(t,s) = \delta_{ts} + \sum_{r\le t} H_\ell(t,r)R_{gu}^\ell(r,s), \qquad 1\le \ell\le L, \tag{86}$$

$$R_{gu}^\ell(t,s) = \sum_{r\le t} G_\ell(t,r)R_{hu}^\ell(r,s), \qquad 1\le \ell< L, \tag{87}$$

$$R_{hr}^\ell(t,s) = \sum_{r\le t} H_\ell(t,r)R_{gr}^\ell(r,s), \qquad 1\le \ell< L, \tag{88}$$

$$R_{gr}^\ell(t,s) = \delta_{ts} + \sum_{r\le t} G_\ell(t,r)R_{hr}^\ell(r,s), \qquad 1\le \ell< L. \tag{89}$$

$$\Lambda_\ell(t) = 1 - H_\ell(t,t)G_\ell(t,t), \qquad \bar R_{hr}^\ell(t,s) = \sum_{r<t} H_\ell(t,r)R_{gr}^\ell(r,s),$$

$$\bar R_{gr}^\ell(t,s) = \delta_{ts} + \sum_{r<t} G_\ell(t,r)R_{hr}^\ell(r,s).$$

$$R_{hr}^\ell(t,s) = \frac{\bar R_{hr}^\ell(t,s) + H_\ell(t,t)\bar R_{gr}^\ell(t,s)}{\Lambda_\ell(t)}, \qquad R_{gr}^\ell(t,s) = \bar R_{gr}^\ell(t,s) + G_\ell(t,t)R_{hr}^\ell(t,s), \quad 1\le\ell<L. \tag{90}$$

$$R_{gu}^L(t,s) = \eta\gamma_0\sum_{r<t} R_{hu}^L(r,s), \tag{91}$$

$$E_L(t) = 1 + \eta\gamma_0\sum_{r<t} F_L(r), \qquad F_L(t) = \sum_{r\le t} H_L(t,r)E_L(r). \tag{92}$$

Layer correlations

$$U_\ell(t,s) = \sum_{r\le s} C_h^{\ell-1}(t,r)R_{gu}^\ell(s,r), \qquad 1\le\ell\le L,$$

$$V_\ell(t,s) = \sum_{r\le s} C_g^{\ell+1}(t,r)R_{hr}^\ell(s,r), \qquad 1\le\ell< L, \tag{93}$$

$$\begin{aligned} C_h^\ell(t,s) &= C_h^{\ell-1}(t,s) + \sum_{m\le s} H_\ell(s,m)U_\ell(t,m) + \sum_{r\le t} H_\ell(t,r)U_\ell(s,r) \\ &\quad + \sum_{r\le t}\sum_{m\le s} H_\ell(t,r)H_\ell(s,m)C_g^\ell(r,m), \qquad 1\le\ell\le L, \end{aligned} \tag{94}$$

$$\begin{aligned} C_g^\ell(t,s) &= C_g^{\ell+1}(t,s) + \sum_{m\le s} G_\ell(s,m)V_\ell(t,m) + \sum_{r\le t} G_\ell(t,r)V_\ell(s,r) \\ &\quad + \sum_{r\le t}\sum_{m\le s} G_\ell(t,r)G_\ell(s,m)C_h^\ell(r,m), \qquad 1\le\ell< L, \end{aligned} \tag{95}$$

$$\begin{aligned} C_g^L(t,s) &= 1 + \eta\gamma_0\sum_{r<t} F_L(r) + \eta\gamma_0\sum_{m<s} F_L(m) \\ &\quad + \eta^2\gamma_0^2\sum_{r<t}\sum_{m<s} C_h^L(r,m). \end{aligned} \tag{96}$$

Predictor overlaps and correlation

$$M_b(t) = \sum_{r \le t} Q(t,r)\Phi_b^v(r), \qquad \Theta_b(t) = \sum_{r \le t} Q(t,r)\Phi_b^w(r). \tag{97}$$

$$\Sigma_0(t,s) = \frac{C_g^1(t,s)}{\omega\gamma_0^2}, \qquad S_0(t,s) = \sum_{m \le s} \Sigma_0(t,m) R_{hr}^0(s,m).$$

$$\begin{aligned} C_q(t,s) = \Sigma_0(t,s) &+ \sum_{m \le s} Q(s,m) S_0(t,m) + \sum_{r \le t} Q(t,r) S_0(s,r) \\ &+ \sum_{r \le t} \sum_{m \le s} Q(t,r) Q(s,m) C_h^0(r,m). \end{aligned} \tag{98}$$

C.4.2 Persistent-SGD Extension

Following Mignacco et al. (2020), we represent persistent mini-batches through correlated activity masks. Persistent mini-batches change the data module, not the architecture equations. Conditioning on the activity history replaces the gradient residual by $Z_a(t) = A_a(t)\Delta_a(t)$. The Gaussian average is otherwise the full-batch calculation. The stationary two-state Markov law then closes the activity average through its two-time correlation and conditional expectation. Active moments drive the gradient; full residual moments give the training risk. The equations below are for $\mu = 0$, $\lambda_{\mathrm{L2}} = 0$, with arbitrary allowed κ. Related DMFT formulations treat momentum (Sarao Mannelli & Urbani, 2021) and alternative learning rules, including error-modulated Hebbian learning (Bordelon & Pehlevan, 2023).

For component a, let $m_a(t) \in \{0,1\}$ have stationary active fraction f_a, and define $A_a(t) = m_a(t)/f_a$, $\gamma_a = \exp(-1/\tau_a)$, and $\chi = 2\Delta/\alpha$. Here τ_a is the persistence time in gradient steps; $\gamma_a = 0$ denotes independent masks. Independent activity masks still reuse a finite dataset; they do not give online learning with fresh examples. The unweighted moments are $B_\Delta = \langle \Delta \rangle$, $\Gamma_\Delta = \langle Y\Delta \rangle$, $L_\Delta = \langle \partial\Delta/\partial Y \rangle$, and $R_\Delta(t,s) = \langle \partial\Delta(t)/\partial u(s) \rangle$. Their active counterparts B_Z, Γ_Z, L_Z, R_Z insert the factor $A(t)$ in the same expectations. We use $C_Z(t,s) = \langle Z(t)Z(s) \rangle$, $C_{Z\Delta}(t,s) = \langle Z(t)\Delta(s) \rangle$, and $C_{\Delta Z}(t,s) = \langle \Delta(t)Z(s) \rangle$. The component index is suppressed in the moment equations.

The architecture is coupled to this block by replacing $(B, L, R_\Delta, C_\Delta)$ with (B_Z, L_Z, R_Z, C_Z) in every gradient correlation, response, and overlap equation. Thus $\zeta_a = \beta_a^y L_Z^a$ in App. C.3, and the centered gradient-noise covariance is χC_Z^a. The architecture initialization is unchanged. Training risk uses C_Δ, not C_Z; the population-risk formula remains (72) with $\mu = 0$.

DMFT equations summary for persistent SGD

Activity and effective process

$$\Phi_a(t,s) = 1 + (f_a^{-1} - 1)\gamma_a^{|t-s|}, \tag{99}$$

$$\mathcal{P}_{t \leftarrow s}^a[U,V] = U + \gamma_a^{t-s}(V - U), \qquad t \ge s, \tag{100}$$

$$\Delta_a(t) = Y_a - M_a(t) - u_a(t) - \chi \sum_{s<t} R_{qh}^a(t,s) Z_a(s), \tag{101}$$

$$h_a(t) = \beta_a^y L_Z^a(t) w^* + u_a^0(t) + \Delta \sum_{s \le t} R_Z^a(t,s) q(s), \tag{102}$$

$$h_0(t) = \frac{1}{2} \sum_{a=\pm 1} [B_Z^a(t) v_a + h_a(t)]. \tag{103}$$

$$\langle u_a(t) u_a(s) \rangle = \Delta C_q(t,s), \qquad \langle G_a u_a(t) \rangle = \sqrt{\Delta}\, \Theta_a(t), \qquad \frac{1}{D} \langle u_a^0(t)^\top u_b^0(s) \rangle = \delta_{ab} \chi C_Z^a(t,s).$$

Active moments and responses

$$L_Z(t) = 1 - \chi \sum_{r<t} R_{qh}(t,r)\Phi(t,r)L_Z(r), \tag{104}$$

$$B_Z(t) = \alpha^{\mathrm{y}} - M(t) - \chi \sum_{r<t} R_{qh}(t,r)\Phi(t,r)B_Z(r), \tag{105}$$

$$\Gamma_Z(t) = 1 - \alpha^{\mathrm{y}} M(t) - \beta^{\mathrm{y}}\Theta(t) - \chi \sum_{r<t} R_{qh}(t,r)\Phi(t,r)\Gamma_Z(r), \tag{106}$$

$$R_Z(t,s) = -\delta_{ts} - \chi \sum_{r<t} R_{qh}(t,r)\Phi(t,r)R_Z(r,s), \tag{107}$$

$$\mathcal{U}_Z(t,s) = \beta^{\mathrm{y}}\Theta(t)L_Z(s) + \Delta \sum_{\ell \le s} C_q(t,\ell)R_Z(s,\ell), \tag{108}$$

$$\begin{aligned} C_Z(t,s) = {} & \Phi(t,s)[\Gamma_Z(s) - M(t)B_Z(s) - \mathcal{U}_Z(t,s)] \\ & - \chi \sum_{r<t} R_{qh}(t,r)\Phi\big(t,\max\{r,s\}\big)C_Z(r,s), \qquad s \le t. \end{aligned} \tag{109}$$

$$C_Z(s,t) = C_Z(t,s), \qquad R_Z(t,s) = R_\Delta(t,s) = 0 \quad (s > t).$$

Full moments and mixed correlations

$$\begin{aligned} L_\Delta(t) &= 1 - \chi \sum_{r<t} R_{qh}(t,r)L_Z(r), \\ B_\Delta(t) &= \alpha^{\mathrm{y}} - M(t) - \chi \sum_{r<t} R_{qh}(t,r)B_Z(r), \\ \Gamma_\Delta(t) &= 1 - \alpha^{\mathrm{y}} M(t) - \beta^{\mathrm{y}}\Theta(t) - \chi \sum_{r<t} R_{qh}(t,r)\Gamma_Z(r), \end{aligned} \tag{110}$$

$$R_\Delta(t,s) = -\delta_{ts} - \chi \sum_{r<t} R_{qh}(t,r)R_Z(r,s), \tag{111}$$

$$\begin{aligned} C_{\Delta Z}(t,s) = {} & \Gamma_Z(s) - M(t)B_Z(s) - \beta^{\mathrm{y}}\Theta(t)L_Z(s) - \Delta \sum_{\ell \le s} C_q(t,\ell)R_Z(s,\ell) \\ & - \chi \sum_{r<t} R_{qh}(t,r)C_Z(r,s), \qquad s \le t, \end{aligned} \tag{112}$$

$$\begin{aligned} C_{Z\Delta}(t,s) = {} & \mathcal{P}_{t\leftarrow s}[\Gamma_\Delta(s), \Gamma_Z(s)] - M(t)\mathcal{P}_{t\leftarrow s}[B_\Delta(s), B_Z(s)] \\ & - \beta^{\mathrm{y}}\Theta(t)\mathcal{P}_{t\leftarrow s}[L_\Delta(s), L_Z(s)] - \Delta \sum_{\ell \le s} C_q(t,\ell)\mathcal{P}_{t\leftarrow s}[R_\Delta(s,\ell), R_Z(s,\ell)] \\ & - \chi \sum_{r<s} R_{qh}(t,r)\mathcal{P}_{t\leftarrow s}[C_{\Delta Z}(s,r), C_Z(s,r)] \\ & - \chi \sum_{s \le r < t} R_{qh}(t,r)\Phi(t,r)C_{Z\Delta}(r,s), \qquad s \le t, \end{aligned} \tag{113}$$

$$\begin{aligned} C_\Delta(t,s) = {} & \Gamma_\Delta(s) - M(t)B_\Delta(s) - \beta^{\mathrm{y}}\Theta(t)L_\Delta(s) - \Delta \sum_{\ell \le s} C_q(t,\ell)R_\Delta(s,\ell) \\ & - \chi \sum_{r<t} R_{qh}(t,r)\mathcal{C}_{Z\Delta}(r,s), \qquad s \le t. \end{aligned} \tag{114}$$

$$\mathcal{C}_{Z\Delta}(r,s) = \begin{cases} C_{Z\Delta}(r,s), & r \ge s, \\ C_{\Delta Z}(s,r), & r < s, \end{cases} \qquad C_\Delta(s,t) = C_\Delta(t,s).$$

Shallow overlaps and training risk

$$M_b(t+1) = M_b(t) + \frac{\eta}{2} \sum_{a=\pm 1} \left[B_Z^a(t) V_{ab} + \beta_a^{\mathrm{y}} L_Z^a(t) A_{ba} + \Delta \sum_{r \le t} R_Z^a(t,r) M_b(r) \right], \tag{115}$$

$$\Theta_b(t+1) = \Theta_b(t) + \frac{\eta}{2} \sum_{a=\pm 1} \left[B_Z^a(t) A_{ab} + \beta_a^{\mathrm{y}} L_Z^a(t) T_{ab} + \Delta \sum_{r \le t} R_Z^a(t,r) \Theta_b(r) \right], \tag{116}$$

$$\widehat{\mathcal{L}}(t) = \frac{1}{2} \sum_{a=\pm 1} C_\Delta^a(t,t). \tag{117}$$

Numerical checks of both extensions are reported in App. E.2.1.

D Online Baseline, Reuse Crossover, and Feature Controls

The online theory of Jain et al. (2024) gives distinct core and shortcut relaxation times. Here we compare these times with the reuse crossover, then study independent core salience, finite-dataset interpolation, ridge, and the dependence of regression on shortcut–label agreement.

D.1 Shallow Online-SGD Mean Field

With a fixed dataset, the residuals depend on earlier predictor values through repeated updates on the same examples. For online learning, draw a new batch of $P = \alpha D$ examples from the shortcut–core model and update

$$q(t+1) = q(t) + \eta \frac{\sqrt{D}}{P} \sum_{i=1}^{P} x_i r_i(t), \qquad r_i(t) = y_i - \frac{q(t)^\top x_i}{\sqrt{D}}.$$

We give the $\mu = 0$ specialization used by the fresh-batch online baseline in the main text; the anisotropic extension is derived in App. D.3. Because every batch is fresh, the order parameters close on

$$m(t) = \frac{q(t)^\top v}{D}, \qquad \theta(t) = \frac{q(t)^\top w^\star}{D}, \qquad Q(t) = \frac{\|q(t)\|^2}{D}.$$

Let $A = \mathbb{E}[ys]$, $B = \mathbb{E}[yg]$, and $q_y = 1$, as in Sec. 2. The population risk and mean gradient are

$$\mathcal{L}(t) = q_y - 2 \left[A m(t) + \sqrt{\Delta} \frac{B}{q_*} \theta(t) \right] + m(t)^2 + \Delta Q(t), \tag{118}$$

$$G(t) = [A - m(t)]\, v + \sqrt{\Delta} \frac{B}{q_*} w^\star - \Delta q(t). \tag{119}$$

Projecting (119) onto the two signal directions yields the closed overlap recursions

$$m(t+1) = m(t) + \eta \left[\nu \big(A - m(t)\big) + \kappa \sqrt{\Delta} \frac{B}{q_*} - \Delta m(t) \right], \tag{120}$$

$$\theta(t+1) = \theta(t) + \eta \left[\kappa \big(A - m(t)\big) + \sqrt{\Delta} B - \Delta \theta(t) \right]. \tag{121}$$

The norm contains the finite-step correction

$$Q(t+1) = Q(t) + 2\eta \frac{q(t)^\top G(t)}{D} + \eta^2 \left[\frac{\|G(t)\|^2}{D} + \frac{\Delta}{\alpha} \mathcal{L}(t) \right]. \tag{122}$$

The final term is fresh-batch gradient noise. It vanishes as $\alpha \to \infty$, leaving the population-gradient recursion. These online equations match direct fresh-batch simulations, as shown in Fig. 7.

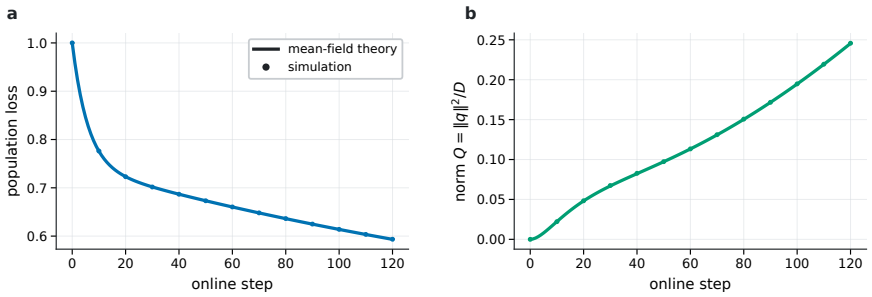

Fig. 7: Online-learning validation for the shortcut–core model. Panels report population risk and the squared predictor norm per dimension. Solid curves are the closed recursions (120)–(122); filled circles and error bars are finite simulations, shown as means ± one standard error over independent trajectories. *Parameters:* App. B.

D.2 Linearized Reuse Crossover Around the Online Learning Path

We estimate the reuse crossover by expanding the fixed-dataset dynamics around the closed online solution. We first consider the isotropic two-component model of Jain et al. (2024). The expansion estimates when the sample-specific contribution $\Delta\chi_{\mathrm{reuse}}$ reaches a specified fraction of the online risk. Changes in the signal overlaps also contribute to total-risk departure, as quantified in (149).

Step 1: specialize the online trajectory. Set $q_* = 1$ in the online equations of App. D.1, and write $q_t = q(t)$, $m_t = m(t)$, $\theta_t = \theta(t)$, and $Q_t = Q(t)$. The overlaps obey (120) and (121), with online risk $\mathcal{L}_{\mathrm{on}}(t)$ given by (118). Writing $b_t = A - m_t$, the norm recursion (122) becomes

$$Q_{t+1} = Q_t + 2\eta\left(m_t b_t + \sqrt{\Delta}B\theta_t - \Delta Q_t\right) + \eta^2\left[\Gamma_t + \frac{\Delta}{\alpha}\mathcal{L}_{\mathrm{on}}(t)\right], \tag{123}$$

$$\Gamma_t = b_t^2\nu + \Delta B^2 + \Delta^2 Q_t + 2b_t\sqrt{\Delta}B\kappa - 2\Delta(b_t m_t + \sqrt{\Delta}B\theta_t). \tag{124}$$

The overlap recursions are affine and can be solved by matrix powers; Q_t then obeys a scalar forced linear recursion. In the continuous online scaling of Jain et al. (2024), the matrix powers become sums of exponentials with shortcut and core relaxation times t_v and $t_\star$. Our plots use the discrete recursions.

Step 2: isolate the fixed-dataset perturbation. We first isolate the gradient fluctuations orthogonal to $\mathrm{span}\{v, w^\star\}$. Their accumulated contribution u_t satisfies the linearized recursion

$$u_{t+1} = \lambda u_t + \eta\zeta_t, \qquad \lambda = 1 - \eta\Delta, \qquad u_0 = 0. \tag{125}$$

Here ζ_t is the orthogonal projection of the quenched fluctuation of the empirical gradient around the closed online gradient, evaluated on the online trajectory. Because we reuse the same examples, ζ_t and ζ_s are correlated. To leading order in $1/\alpha$,

$$\frac{1}{D}\mathbb{E}[\zeta_t^\top\zeta_s] = \frac{\Delta}{\alpha}C_\Delta^{\mathrm{on}}(t,s), \tag{126}$$

where the online residual correlation is

$$C_\Delta^{\mathrm{on}}(t,s) = 1 - A(m_t + m_s) - \sqrt{\Delta}B(\theta_t + \theta_s) + m_t m_s + \Delta Q_{ts}^{\mathrm{on}}. \tag{127}$$

The two-time overlap $Q_{ts}^{\mathrm{on}} = q_t^\top q_s/D$ is also explicit. Its projection onto $\mathrm{span}\{v, w^\star\}$ is

$$Q_{ts}^{\parallel} = \frac{\nu\theta_t\theta_s - \kappa(\theta_t m_s + m_t\theta_s) + m_t m_s}{\nu - \kappa^2}, \tag{128}$$

and the orthogonal online noise gives $Q_{ts}^{\perp} = \lambda^{|t-s|}Q_{\min(t,s)}^{\perp}$, with $Q_t^{\perp} = Q_t - Q_{tt}^{\parallel}$. Thus $Q_{ts}^{\mathrm{on}} = Q_{ts}^{\parallel} + Q_{ts}^{\perp}$.

Step 3: remove the fresh-batch diagonal. The diagonal part $t = s$ of (126) is already included in the closed online norm recursion through the $\Delta\mathcal{L}_{\rm on}(t)/\alpha$ term in (123). The reuse-specific correction is therefore the off-diagonal memory only:

$$\chi_{\rm reuse}(t) \simeq \frac{\eta^2\Delta}{\alpha} \sum_{\substack{0\le r,s<t \\ r\neq s}} \lambda^{2t-2-r-s} C_\Delta^{\rm on}(r,s). \tag{129}$$

This estimates $\chi_{\rm reuse} = Q_{\rm out}^{\rm reuse} - Q_{\rm out}^{\rm on}$ after subtracting the fresh-batch contribution. Its population-risk contribution is $\Delta\chi_{\rm reuse}$, distinct from the contribution of changes in m and θ. The corresponding linearized reuse crossover time is

$$t_{\rm reuse,lin}^{(\varepsilon)}(\alpha) = \inf\left\{ t : \frac{\Delta\chi_{\rm reuse}(t)}{\mathcal{L}_{\rm on}(t)} > \varepsilon \right\}. \tag{130}$$

If the residual correlation is approximately constant over the first few steps, (129) grows as $t(t-1)/\alpha$. Larger datasets therefore delay the reuse crossover in this regime. The damping factor λ and the evolving online residual correlation limit this early quadratic growth.

More explicitly, if $C_\Delta^{\rm on}(r,s) \simeq \mathcal{L}_{\rm on}(t)$ over the memory window and $\eta\Delta t \ll 1$, then $\lambda^{2t-2-r-s} \simeq 1$. The off-diagonal sum contains $t(t-1)$ ordered pairs, so

$$\frac{\Delta\chi_{\rm reuse}(t)}{\mathcal{L}_{\rm on}(t)} \simeq \frac{\eta^2\Delta^2}{\alpha} t(t-1). \tag{131}$$

Solving the threshold condition gives the compact crossover estimate

$$t_{\rm reuse,early}^{(\varepsilon)} = \left\lceil \frac{1+\sqrt{1+4\varepsilon\alpha/(\eta^2\Delta^2)}}{2} \right\rceil, \tag{132}$$

$$t_{\rm reuse,early}^{(\varepsilon)} \simeq \frac{\sqrt{\varepsilon\alpha}}{\eta\Delta} \qquad \left[\frac{\varepsilon\alpha}{\eta^2\Delta^2} \gg 1\right]. \tag{133}$$

Since $t_{\rm core} \simeq 1/(\eta\Delta)$, the early approximation predicts $t_{\rm reuse}/t_{\rm core} \simeq \sqrt{\varepsilon\alpha}$. Its formal reuse–core boundary is therefore $\alpha \simeq 1/\varepsilon$, equal to 50 at the 2% convention. This boundary lies precisely where $\eta\Delta t_{\rm reuse} \ll 1$ ceases to hold, so its location is only a rough estimate; the evolving residual correlation and damping in (129) must be retained near the crossing.

Step 4: evaluate the linearized crossover. We evaluated (130) on the same sample-ratio setting as Fig. 4: $q_* = 1$, $\rho = 0.75$, $\nu = 128$, $\Delta = 0.1$, $\eta = 10^{-3}$, and horizon 10^6. A check of the online baseline gives a maximum risk discrepancy of 2.2×10^{-16} between the closed recursions and the general online implementation. The damped linearized estimate gives

$$t_{\rm reuse,lin}^{2\%} = (721, 1469, 3086, 7223) \quad \text{for} \quad \alpha = (0.25, 1, 4, 16).$$

These estimates concern the $Q_{\rm out}$-based criterion. In the early-time regime of (151)–(152), both $\Delta\chi_{\rm reuse}$ and the total-risk difference grow as $t(t-1)/\alpha$, but with different coefficients. Thus, a common square-root dependence on α does not imply equal crossover times.

Step 5: compare the timescale–crossover ordering. We evaluated the damped linearized correction and finite-dataset spectral dynamics at the same 72 one-factor parameter settings. The scan uses $\alpha \in \{0.25, 0.5, 1, 2, 4, 8, 16, 32\}$, and separately varies $\nu \in \{32, 128, 512\}$, $\rho \in \{0.6, 0.75, 0.9\}$, $\Delta \in \{0.05, 0.1, 0.2\}$, and $\eta \in \{5\times10^{-4}, 10^{-3}, 2\times10^{-3}\}$. Every finite-dataset point averages 24 independently sampled $D = 128$ training datasets and is propagated to 10^6 steps with the exact spectral formula.

Fig. 8 compares the relaxation timescales with the damped linearized crossover. At the baseline, this estimate gives $t_{\rm sc} < t_{\rm reuse} < t_{\rm core} \le t_{\rm edge}$ through $\alpha = 16$, and $t_{\rm sc} < t_{\rm core} < t_{\rm reuse} < t_{\rm edge}$ at $\alpha = 32$. Panel (b) shows the corresponding predicted ratio $t_{\rm reuse}/t_{\rm core}$. Across the full parameter scan, the early approximation (132) has 7.2% median relative error against the damped linearized estimate (130), not against full DMFT. The error grows at large α, where the short-time assumption fails. The numerical departure times in Fig. 8 illustrate dataset-size dependence. These comparisons do not quantify the accuracy of the $Q_{\rm out}$-based crossover against full DMFT using the same threshold and normalization.

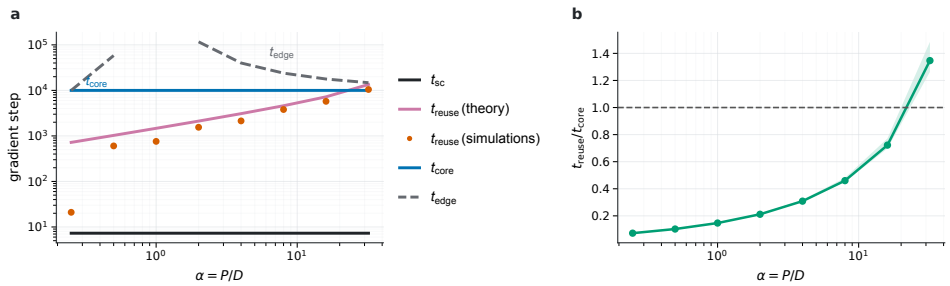

Fig. 8: Relaxation timescales and reuse crossover across sample ratio. (a) Modal timescales and the damped linearized prediction for the Q_{out}-based crossover (pink line); mean numerical departure times (orange points) are shown for qualitative comparison. The blue line is t_{core}; the grey dashed curve is t_{edge}, which diverges at $\alpha = 1$. (b) Predicted ratio $t_{\text{reuse}}/t_{\text{core}}$; the band spans separate variations of ν, ρ, Δ, and η, and the dashed line marks equal timescales. *Parameters:* App. B.

D.3 Independent Core Salience and the Generalized Reuse Correction

The model in (1) allows independent control of shortcut and core salience. Changing μ changes both the population relaxation modes and the label moments. At $\kappa = 0$, the shortcut moment is $A = 0$, and its population coefficient remains zero from zero initialization. This section derives the online recurrence and reuse correction for arbitrary $\mu \geq 0$ in the same discrete-time convention as the preceding calculation. It extends our fresh-batch, finite-step analogue of the one-example online dynamics in Jain et al. (2024) to anisotropic core salience.

Step 1: input distribution and labels. Let $z \sim \mathcal{N}(0, I_D)$, $g = w^{\star\top} z/\sqrt{D}$, and remove the core component through $z_\perp = z - (g/q_*)w^\star/\sqrt{D}$. We draw

$$x = \frac{sv}{\sqrt{D}} + \sqrt{\Delta}\, z_\perp + \sqrt{\Delta + \mu}\, \frac{g}{q_*} \frac{w^\star}{\sqrt{D}}, \qquad y = \operatorname{sign}\Big(\kappa s + \sqrt{\Delta + \mu}\, g\Big), \tag{134}$$

where $s \in \{-1, +1\}$ is uniform and independent of z. The moments are fixed by

$$A = \operatorname{erf}\left(\frac{\kappa}{\sqrt{2(\Delta + \mu)q_*}}\right), \qquad B = \sqrt{\frac{2q_*}{\pi}}\, e^{-\kappa^2/[2(\Delta+\mu)q_*]}, \qquad q_y = 1.$$

Define $\nu = \|v\|^2/D$, $\kappa = v^\top w^\star/D$, and $q_* = \|w^\star\|^2/D$. The signal-plane Gram matrix, population covariance, and label forcing are

$$G = \begin{pmatrix} \nu & \kappa \\ \kappa & q_* \end{pmatrix}, \quad \Sigma_x = \Delta I + \frac{vv^\top}{D} + \frac{\mu}{q_*}\frac{w^\star w^{\star\top}}{D}, \quad b = Av + hw^\star, \qquad h = \frac{\sqrt{\Delta + \mu}\, B}{q_*}. \tag{135}$$

Consequently the population risk of a predictor q, with $m = q^\top v/D$, $\theta = q^\top w^\star/D$, and $Q = \|q\|^2/D$, is

$$\mathcal{L}(q) = q_y - 2Am - 2h\theta + m^2 + \Delta Q + \frac{\mu}{q_*}\theta^2. \tag{136}$$

The core moment B enters the gradient through h.

Population ridge response. Ridge can improve in-distribution prediction while increasing learned spurious correlations (Bombari & Mondelli, 2025). Here we derive its effect on the shortcut-to-core coefficient ratio. Minimise $\mathcal{L}/2 + \lambda_{\text{L2}} Q/2$, with population risk $\mathcal{L}$ given by (136), and write $q = c_{\text{sc}}^{\text{pop}} v + c_{\text{core}}^{\text{pop}} w^\star$ and $\delta_\lambda = \Delta + \lambda_{\text{L2}}$. Using $m = \nu c_{\text{sc}}^{\text{pop}} + \kappa c_{\text{core}}^{\text{pop}}$ and $\theta = \kappa c_{\text{sc}}^{\text{pop}} + q_* c_{\text{core}}^{\text{pop}}$, the stationarity equations are

$$\begin{pmatrix} \delta_\lambda + \nu & \kappa \\ \mu\kappa/q_* & \delta_\lambda + \mu \end{pmatrix} \begin{pmatrix} c_{\text{sc}}^{\text{pop}} \\ c_{\text{core}}^{\text{pop}} \end{pmatrix} = \begin{pmatrix} A \\ h \end{pmatrix}. \tag{137}$$

The determinant $d_\lambda = (\delta_\lambda + \nu)(\delta_\lambda + \mu) - \mu\kappa^2/q_*$ is positive for $\Delta > 0$, $\lambda_{\mathrm{L2}} \geq 0$, and $\det G > 0$. Hence

$$c_{\mathrm{sc}}^{\mathrm{pop}} = \frac{A(\delta_\lambda + \mu) - \kappa h}{d_\lambda}, \qquad c_{\mathrm{core}}^{\mathrm{pop}} = \frac{h(\delta_\lambda + \nu) - \mu\kappa A/q_*}{d_\lambda}. \tag{138}$$

Differentiating their ratio gives

$$\frac{\partial}{\partial \lambda_{\mathrm{L2}}} \frac{c_{\mathrm{sc}}^{\mathrm{pop}}}{c_{\mathrm{core}}^{\mathrm{pop}}} = \frac{Ah(\nu - \mu) + \kappa\big(h^2 - \mu A^2/q_*\big)}{\big[h(\delta_\lambda + \nu) - \mu\kappa A/q_*\big]^2}. \tag{139}$$

For $\mu = 0$ and $B > 0$, positivity reduces to $A\nu q_* + \kappa\sqrt{\Delta}B > 0$, recovering (156). For $\mu > 0$, core salience and geometric mixing can change the sign of the relative ridge response.

Step 2: close the online signal plane. For fresh batches of size $P = \alpha D$, the mean update is $q_{t+1} = q_t + \eta(b - \Sigma_x q_t) + \eta\xi_t$, where the fresh-batch fluctuation ξ_t has zero mean. Projecting onto v and $w^\star$ gives the affine two-dimensional recurrence

$$p_{t+1} = (I - \eta H)p_t + \eta d, \quad p_t = \begin{pmatrix} m_t \\ \theta_t \end{pmatrix}, \quad H = \begin{pmatrix} \Delta + \nu & \mu\kappa/q_* \\ \kappa & \Delta + \mu \end{pmatrix}, \quad d = \begin{pmatrix} A\nu + h\kappa \\ A\kappa + hq_* \end{pmatrix}. \tag{140}$$

Thus $p_t = [I - (I - \eta H)^t]H^{-1}d$ from zero initialization. This is the two-mode propagator for arbitrary κ. For orthogonal features $\kappa = 0$, the eigenvalues are $\lambda_{\mathrm{sc}} = \Delta + \nu$ and $\lambda_{\mathrm{core}} = \Delta + \mu$, with exact discrete timescales

$$t_j = -\frac{1}{\log|1 - \eta\lambda_j|}, \qquad j \in \{\mathrm{sc}, \mathrm{core}\}. \tag{141}$$

For $\kappa = 0$, these are covariance relaxation times; the shortcut mode is unforced because $A = 0$. For $\kappa \neq 0$, both eigenmodes can contribute to the signal trajectory.

Step 3: close the online norm. Set $a_t = A - m_t$ and $c_t = h - (\mu/q_*)\theta_t$. The squared mean gradient per dimension and its contraction with the predictor are

$$\Gamma_t = a_t^2\nu + c_t^2 q_* + 2a_t c_t\kappa + \Delta^2 Q_t - 2\Delta(a_t m_t + c_t\theta_t), \tag{142}$$

$$q_t^\top(b - \Sigma_x q_t)/D = Am_t + h\theta_t - m_t^2 - \Delta Q_t - \frac{\mu}{q_*}\theta_t^2. \tag{143}$$

Along the online trajectory, set $\mathcal{L}_{\mathrm{on}}(t) = \mathcal{L}(q_t)$. The fresh-batch fluctuation contributes $\Delta\mathcal{L}_{\mathrm{on}}(t)/\alpha$ at leading order in D, because the two signal spikes have finite rank. Therefore

$$Q_{t+1} = Q_t + 2\eta\left(Am_t + h\theta_t - m_t^2 - \Delta Q_t - \frac{\mu}{q_*}\theta_t^2\right) + \eta^2\left(\Gamma_t + \frac{\Delta}{\alpha}\mathcal{L}_{\mathrm{on}}(t)\right). \tag{144}$$

Eqs. (136), (140), and (144) close the fresh-batch online dynamics.

Step 4: propagate the residual correlation. Let $Q_{rs} = q_r^\top q_s/D$. The signal-plane contribution is obtained without choosing an orthogonal basis:

$$Q_{rs}^{\parallel} = p_r^\top G^{-1} p_s. \tag{145}$$

Outside this plane the response remains $\lambda_\perp = 1 - \eta\Delta$, so $Q_{rs}^\perp = \lambda_\perp^{|r-s|} Q_{\min(r,s)}^\perp$, where $Q_t^\perp = Q_t - p_t^\top G^{-1} p_t$. The closed online residual correlation entering the reuse perturbation is then

$$C_{\Delta,\mu}^{\mathrm{on}}(r,s) = q_y - A(m_r + m_s) - h(\theta_r + \theta_s) + m_r m_s + \Delta Q_{rs} + \frac{\mu}{q_*}\theta_r\theta_s. \tag{146}$$

The extra μ-term and the $\Delta + \mu$ damping already present in θ_t are the two changes that would be missed by reusing the isotropic formula unchanged.

Step 5: subtract fresh-batch noise and define the reuse crossover. Projecting the gradient fluctuations outside the signal plane and removing the diagonal $r = s$, already counted in (144), gives the approximation

$$\chi^{(\mu)}_{\rm reuse}(t) \simeq \frac{\eta^2\Delta}{\alpha} \sum_{\substack{0\le r,s<t \\ r\neq s}} \lambda_\perp^{2t-2-r-s} C^{\rm on}_{\Delta,\mu}(r,s). \tag{147}$$

For a relative threshold ε, the reuse crossover is

$$t^{(\varepsilon)}_{\rm reuse} = \inf\left\{ t : \Delta\chi^{(\mu)}_{\rm reuse}(t)/\mathcal{L}_{\rm on}(t) > \varepsilon \right\}. \tag{148}$$

It is defined by the relative reuse correction, not by a relaxation timescale. Setting $\mu = 0$ and $q_* = 1$, so that $h = \sqrt{\Delta}B$, recovers the online equations and reuse correction of App. D.2. Numerical checks recover this reduction within 10^{-12}; the finite-dataset spectral propagator agrees with direct gradient updates within 10^{-10}.

Signal-plane contribution to population risk. For the general geometry in Sec. 2, with $\det G > 0$, use $p = (m, \theta)^\top$ and G from (8), and set

$$a_e = \begin{pmatrix} A_e \\ h_e \end{pmatrix}, \qquad h_e = \frac{\sqrt{\Delta+\mu}\, B_e}{q_*}, \qquad K_\parallel = \begin{pmatrix} 1 & 0 \\ 0 & \mu/q_* \end{pmatrix} + \Delta G^{-1}.$$

The main-text risk (9) is $\mathcal{L}_e = 1 - 2a_e^\top p + p^\top K_\parallel p + \Delta Q_{\rm out}$. Thus, with $\delta p = p_{\rm reuse} - p_{\rm on}$ and $\delta\mathcal{L}_e = \mathcal{L}_{e,\rm reuse} - \mathcal{L}_{e,\rm on}$,

$$\delta\mathcal{L}_e(t) = \Delta\chi_{\rm reuse}(t) + \underbrace{2(K_\parallel p_{\rm on}(t) - a_e)^\top \delta p(t) + \delta p(t)^\top K_\parallel \delta p(t)}_{\delta\mathcal{L}_{\parallel,e}(t)}. \tag{149}$$

This identity uses the exact difference in $Q_{\rm out}$, approximated by (147); $\delta\mathcal{L}_{\parallel,e}$ is the signal-plane contribution.

Its order follows directly from the update in (6). Start at $q_0 = 0$ without ridge and define $\widehat{\Sigma} = P^{-1}\sum_i x_i x_i^\top$ and $\widehat{b} = \sqrt{D}\, P^{-1}\sum_i x_i y_i$, with $\mathbb{E}\widehat{b} = b = Av + hw^\star$ as in (135). Expanding the reused update gives $q_t^{\rm reuse} = \eta t\widehat{b} - \eta^2\binom{t}{2}\widehat{\Sigma}\widehat{b} + \cdots$. Independent online batches instead give $\Sigma_x b$ for the mean of the second product. The same-example terms yield

$$\mathbb{E}[\widehat{\Sigma}\widehat{b}] - \Sigma_x b = \frac{\sqrt{D}\,\mathbb{E}[x\|x\|^2 y] - \Sigma_x b}{P} = \frac{\Delta}{\alpha} b + o(\sqrt{D}), \tag{150}$$

since $\|x\|^2/D \to \Delta$. Consequently, after the proportional limit, at fixed t, α and small η, with $a = (A, h)^\top$,

$$\delta p(t) = -\frac{\eta^2\Delta}{2\alpha} t(t-1) G a + O(\eta^3), \tag{151}$$

$$\delta\mathcal{L}_e(t) = \frac{\eta^2\Delta}{\alpha} t(t-1)\big[\Delta + a_e^\top G a\big] + O(\eta^3). \tag{152}$$

The second line uses $\chi_{\rm reuse}(t) = \eta^2\Delta t(t-1)/\alpha + O(\eta^3)$. At $t = 2$, (151) is exact in η. For decorrelated risk, $a_e^\top G a = h(\kappa A + q_* h)$: the signal-plane term need not vanish when $A_e = 0$, and contributes at the same order as $\Delta\chi_{\rm reuse}$.

Neglecting this contribution requires $|\delta\mathcal{L}_{\parallel,e}| \ll |\Delta\chi_{\rm reuse}|$ near the crossing, with both thresholds using the same online-risk normalization. Extending the fixed-t expansion to a time window requires $\eta t \max\{\|\Sigma_x\|_{\rm op}, \|\widehat{\Sigma}\|_{\rm op}\} \ll 1$, where $\|\cdot\|_{\rm op}$ is the spectral norm. The weaker condition $\eta\Delta t \ll 1$ alone therefore does not identify the $Q_{\rm out}$ crossover with the first departure of total risk.

Step 6: connect the online and fixed-dataset equations. The online equations above and the fixed-dataset CHSCK system in App. C.2 use the same anisotropic kernel K_q and signal-plane geometry. In the fixed-dataset system, the core spike enters through $\mathcal{T}_a$ and ζ_a; because $A_{ab} = a\kappa$, this term contributes to both M_b and Θ_b when $\kappa \neq 0$. Taking the large-sample limit removes the reuse memory and recovers the coupled propagator in (140). For $\mu = 0$, the online equations reduce to App. D.1; for $\kappa = 0$, the signal-plane modes decouple, giving the relaxation times in (10).

Step 7: integrate stiff signal modes without changing the closure. For the high-salience panels of Fig. 3, $\nu = 128$ makes an explicit signal-plane update stability-limited even after the memory kernels are smooth. Conditional on a trial kernel row, the finite-dimensional overlaps obey an affine equation $\dot{p} = J_n p + b_n$. We therefore use the embedded pair

$$(I - h_n J_n) p_{n+1}^{\mathrm{IE}} = p_n + h_n b_n, \qquad \left(I - \frac{h_n}{2} J_n\right) p_{n+1}^{\mathrm{IT}} = p_n + \frac{h_n}{2}\left[\dot{p}_n + b_n\right], \tag{153}$$

namely implicit Euler and implicit trapezoidal updates for this small block. All CHSCK rows still use the quadrature and causal closure described in App. E.1.3; the full predicted–corrected state enters the same adaptive error controller. This removes numerical oscillations around the shortcut fixed point but does not alter the DMFT equations. On the common $T = 12000$ validation window, empirical and population risks agree with the explicit small-step reference to relative RMSE 2.1×10^{-4} and 2.0×10^{-4}, respectively. The $\mu = 0$ reduction is also checked against the original adaptive closure to machine precision, and the large-α core-spike trajectory agrees with the closed online recurrence in (140)–(144).

D.4 Finite reuse, interpolation, and anisotropic retention

The reuse crossover measures the growth of sample-specific fitting relative to online learning. We now derive the converged feature coefficients and sample-specific norm used in Sec. 3.2.

Signal and sample-specific coordinates. Write $q_{\|} = c_{\mathrm{sc}} v + c_{\mathrm{core}} w^*$. With $G = \left(\begin{smallmatrix} \nu & \kappa \\ \kappa & q_* \end{smallmatrix}\right)$,

$$\begin{pmatrix} c_{\mathrm{sc}} \\ c_{\mathrm{core}} \end{pmatrix} = G^{-1} \begin{pmatrix} m \\ \theta \end{pmatrix}, \qquad Q_{\mathrm{out}} = Q - \begin{bmatrix} m & \theta \end{bmatrix} G^{-1} \begin{pmatrix} m \\ \theta \end{pmatrix}. \tag{154}$$

Thus Q_{out} is the squared norm per dimension outside the signal plane. A population minimizer lies in the signal plane, whereas a finite interpolating solution can have $Q_{\mathrm{out}} > 0$.

For a fixed training set, let $\widehat{\Sigma} = X^\top X/P$ and $b_X = \sqrt{D} X^\top y/P$. If $\widehat{\Sigma} = U \operatorname{diag}(\lambda_i) U^\top$, zero-initialized full-batch gradient descent with ridge λ_{L2} has the exact integer-time solution

$$q_t = U \operatorname{diag}\left[\frac{1 - \{1 - \eta(\lambda_i + \lambda_{\mathrm{L2}})\}^t}{\lambda_i + \lambda_{\mathrm{L2}}}\right] U^\top b_X. \tag{155}$$

The zero-eigenvalue factor is defined by continuity. This formula explains both the empirical spectral times and the converged solution: at zero ridge and $q_0 = 0$, it converges to $\widehat{\Sigma}^+ b_X$, the minimum-norm empirical solution. The slow lower edge produces the long tail near interpolation, while eigenvectors outside the signal plane generate Q_{out}.

Population reference and ridge. The population ridge coefficients are given by (137)–(138), and their relative ridge response by (139). For $\mu = 0$ and $\kappa > 0$, the relative ridge response is positive:

$$\frac{\partial}{\partial \lambda_{\mathrm{L2}}} \frac{c_{\mathrm{sc}}^{\mathrm{pop}}}{c_{\mathrm{core}}^{\mathrm{pop}}} = \frac{A\nu q_* / (\sqrt{\Delta} B) + \kappa}{(\Delta + \nu + \lambda_{\mathrm{L2}})^2} > 0. \tag{156}$$

The sign does not have to remain positive when core salience μ is large.

Finite-dataset signal coefficients. Use the model in Sec. 2, fixing $(\Delta, \mu, \nu, \kappa, q_*)$ with $\Delta > 0$, $\mu \geq 0$, and $\det G > 0$. Take training to convergence before $P, D \to \infty$ at fixed $\alpha = P/D$. In orthonormal signal-plane coordinates write $X = (S, \sqrt{\Delta} Z)$, where $S \in \mathbb{R}^{P \times 2}$ contains the signal coordinates and the standard Gaussian bulk $Z \in \mathbb{R}^{P \times (D-2)}$ is independent of (S, y). For $\beta = q/\sqrt{D}$, eliminating the bulk coefficients gives the exact finite-dataset identity

$$T_\lambda = (\Delta Z Z^\top + P\lambda I_P)^{-1}, \qquad \widehat{\beta}_{\|} = (I_2 + S^\top T_\lambda S)^{-1} S^\top T_\lambda y, \qquad \lambda = \lambda_{\mathrm{L2}}. \tag{157}$$

This holds for $\lambda > 0$; at zero ridge it requires $P < D - 2$, which holds eventually at fixed $0 < \alpha < 1$.

Let $\Sigma_{\|} = \mathbb{E}[S_i S_i^\top]$ and $\bar{b}_{\|} = \mathbb{E}[S_i y_i]$, with S_i the column vector of signal coordinates for example i. The population residual $\epsilon_i = y_i - S_i^\top \Sigma_{\|}^{-1} \bar{b}_{\|}$ depends on S_i, but satisfies $\mathbb{E}[S_i \epsilon_i] = 0$. Uniform s

also gives $\mathbb{E}[S_i] = \mathbb{E}[\epsilon_i] = 0$. Independence of the rows and of Z from (S, ϵ), together with finite fourth moments, therefore implies

$$\mathbb{E}[S^\top T_\lambda \epsilon \mid T_\lambda] = 0, \qquad \mathrm{Var}[(S^\top T_\lambda \epsilon)_j \mid T_\lambda] \le C \operatorname{Tr} T_\lambda^2 = O_p(P^{-1}), \quad j = 1, 2, \tag{158}$$

for a moment-dependent constant C. For positive ridge the bound follows from $\|T_\lambda\| \le (P\lambda)^{-1}$; at zero ridge and fixed $\alpha < 1$, $\Delta Z Z^\top / P$ has a positive limiting lower spectral edge. Thus $S^\top T_\lambda \epsilon = o_p(1)$, without assuming independence between residuals and signal coordinates. Here $o_p(1)$ denotes a term vanishing in probability.

The same quadratic-form concentration gives $S^\top T_\lambda S = \chi_\lambda \Sigma_\parallel + o_p(1)$, where χ_λ is the limit of $\operatorname{Tr} T_\lambda$. The Gaussian-bulk resolvent equation (Marchenko & Pastur, 1967) gives

$$\frac{1}{\chi_\lambda} = \lambda + \frac{\Delta/\alpha}{1 + \Delta\chi_\lambda}, \qquad \lambda_{\text{eff}} = \chi_\lambda^{-1}. \tag{159}$$

Choose $\chi_\lambda > 0$; at zero ridge and $\alpha < 1$, $\lambda_{\text{eff}} = \Delta(1-\alpha)/\alpha$. Substitution into (157) and conversion to the (v, w^*) coefficients yield

$$\widehat{\mathbf{c}}^{\text{fin}} = \mathbf{c}^{\text{fin}} + o_p(1), \qquad \mathbf{c}^{\text{fin}} = \begin{pmatrix} \Delta + \nu + \lambda_{\text{eff}} & \kappa \\ \mu\kappa/q_* & \Delta + \mu + \lambda_{\text{eff}} \end{pmatrix}^{-1} \begin{pmatrix} A \\ h \end{pmatrix} = \mathbf{c}^{\text{pop}}(\lambda_{\text{eff}}), \tag{160}$$

where $\mathbf{c} = (c_{\text{sc}}, c_{\text{core}})^\top$, a hat denotes empirical coefficients, and $h = \sqrt{\Delta + \mu}\, B/q_*$. The residual contributes $o_p(1)$ to each signal coefficient, but can produce a nonvanishing norm outside the signal plane. For ridgeless $\alpha > 1$, the direct projection calculation in (174)–(175) instead gives $\mathbf{c}^{\text{fin}} = \mathbf{c}^{\text{pop}}(0)$. Ridgeless interpolation $(\alpha = 1)$ is excluded.

Why an underdetermined sample can amplify the shortcut. For zero ridge, $0 < \alpha < 1$, and $\mu = 0$, (160) gives

$$\begin{pmatrix} c_{\text{sc}}^{\text{fin}} \\ c_{\text{core}}^{\text{fin}} \end{pmatrix} = \begin{pmatrix} r_{\text{sc}} & (\kappa/\nu)(r_{\text{sc}} - \alpha) \\ 0 & \alpha \end{pmatrix} \begin{pmatrix} c_{\text{sc}}^{\text{pop}} \\ c_{\text{core}}^{\text{pop}} \end{pmatrix}, \qquad r_{\text{sc}} = \frac{\alpha(\Delta + \nu)}{\Delta + \alpha\nu} > \alpha. \tag{161}$$

For general $\mu \ge 0$, subtracting the unregularised population solution and inverting the two-dimensional matrix instead gives

$$\begin{aligned} c_{\text{sc}}^{\text{fin}} - c_{\text{sc}}^{\text{pop}} &= \frac{(1-\alpha)\Delta}{d_\alpha} \left[\alpha\kappa c_{\text{core}}^{\text{pop}} - (\Delta + \alpha\mu) c_{\text{sc}}^{\text{pop}}\right], \\ d_\alpha &= (\Delta + \alpha\nu)(\Delta + \alpha\mu) - \alpha^2\mu\kappa^2/q_* > 0. \end{aligned} \tag{162}$$

Positivity follows from $\Delta > 0$ and $\nu q_* > \kappa^2$. The bracket therefore determines amplification of the limiting coefficient, proving Result 2; equality gives equal limiting coefficients, not a fixed sign for finite-dataset fluctuations. At $\kappa = 0$, the transfer term vanishes.

Finite-dataset ridge. For a signal-plane covariance eigenvalue τ, (160) retains the fraction

$$r_\tau(\alpha, \lambda) = \frac{\tau\chi_\lambda}{1 + \tau\chi_\lambda} = \frac{\tau}{\tau + \lambda_{\text{eff}}} \tag{163}$$

of the unregularised population coefficient. On the positive-ridge branch, $\lambda_{\text{eff}} > \max\{0, \Delta(1 - \alpha)/\alpha\}$. Differentiating (159) and applying the chain rule to (156) gives

$$\begin{aligned} \frac{\mathrm{d}\lambda_{\text{eff}}}{\mathrm{d}\lambda} &= \left[1 - \frac{\Delta^2}{\alpha(\Delta + \lambda_{\text{eff}})^2}\right]^{-1} > 0, \\ \frac{\mathrm{d}}{\mathrm{d}\lambda} \frac{c_{\text{sc}}^{\text{fin}}}{c_{\text{core}}^{\text{fin}}} &= \frac{A\nu q_*/(\sqrt{\Delta}B) + \kappa}{(\Delta + \nu + \lambda_{\text{eff}})^2} \frac{\mathrm{d}\lambda_{\text{eff}}}{\mathrm{d}\lambda} > 0 \quad (\mu = 0,\ A, B > 0,\ \kappa \ge 0). \end{aligned} \tag{164}$$

This proves the finite-α statement in Result 3 for every fixed $\alpha > 0$ and $\lambda > 0$.

For $\mu = 0$, the shortcut direction is a covariance eigenvector, so its limiting overlap satisfies

$$m_\alpha(\lambda) = \frac{\Delta + \nu}{\Delta + \nu + \lambda_{\text{eff}}} m^{\text{pop}}(0), \qquad \frac{\mathrm{d}m_\alpha}{\mathrm{d}\lambda} \le 0 \quad \text{if } m^{\text{pop}}(0) \ge 0. \tag{165}$$

The decrease is strict when $m^{\text{pop}}(0) > 0$. Thus the shortcut overlap can decrease while the shortcut-to-core coefficient ratio increases; neither alone determines shifted-environment risk.

D.5 Risk Reversal

We derive Result 4 for the labels defined in the main text, with $\Delta > 0$, $\mu \geq 0$, and $0 \leq \kappa < \sqrt{\nu q_*}$. Training uses full-batch gradient descent with $\lambda_{\mathrm{L2}} = 0$, $q(0) = 0$, and a stable step size. The converged calculation first takes the finite-dataset least-squares solution and then the proportional limit at fixed $\alpha = P/D > 1$, away from interpolation.

Step 1: compute the label moments. Write $V = \Delta + \mu$ for the centered variance along the core direction. The label and its moments are

$$\begin{aligned} y &= \mathrm{sign}(\kappa s + \sqrt{V}\, g), \qquad u = \frac{\kappa}{\sqrt{V q_*}}, \qquad A = \mathrm{erf}(u/\sqrt{2}), \\ B &= \sqrt{\frac{2q_*}{\pi}} e^{-u^2/2}, \qquad \bar{B} = \frac{B}{\sqrt{q_*}}, \qquad h = \frac{\sqrt{V}\, B}{q_*}, \qquad q_y = 1. \end{aligned} \tag{166}$$

Here u is the standardized teacher shift, $\bar{B}$ the normalized core moment, and h the core forcing from (135). Shortcut–label agreement is $\rho = \Pr(y = s) = \Phi(u)$, where Φ is the standard normal cumulative distribution function, so $A = 2\rho - 1$. Increasing κ at fixed (Δ, μ, ν, q_*) increases A, decreases B, and changes the overlap between the features.

For each value of κ, decorrelated evaluation retains that problem's input distribution and uses $A_e = 0$, $B_e = B$, $q_y = 1$, as in Sec. 2.1. Thus the two problems do not share a fixed core moment or a common test distribution. This is the moment-defined evaluation in (9), not the unchanged deterministic label rule applied to unchanged inputs. The comparison therefore does not isolate shortcut evidence at fixed core evidence.

Step 2: solve the population regression. Setting $\lambda_{\mathrm{L2}} = 0$ in (137) gives

$$\begin{aligned} c_{\mathrm{sc}}^{\mathrm{pop}} &= \frac{VA - \kappa h}{d_0}, \qquad c_{\mathrm{core}}^{\mathrm{pop}} = \frac{(\Delta + \nu)h - \mu\kappa A/q_*}{d_0}, \\ d_0 &= (\Delta + \nu)V - \frac{\mu\kappa^2}{q_*} > 0. \end{aligned} \tag{167}$$

The overlaps are $m^{\mathrm{pop}} = \nu c_{\mathrm{sc}}^{\mathrm{pop}} + \kappa c_{\mathrm{core}}^{\mathrm{pop}}$ and $\theta^{\mathrm{pop}} = \kappa c_{\mathrm{sc}}^{\mathrm{pop}} + q_* c_{\mathrm{core}}^{\mathrm{pop}}$. The normal equations imply that the variance of the population prediction equals its covariance with the training label. Writing σ_{res}^2 for the variance of its training residual gives

$$\begin{aligned} \sigma_{\mathrm{res}}^2 &= 1 - Am^{\mathrm{pop}} - h\theta^{\mathrm{pop}}, \\ \mathcal{L}_{\mathrm{dec}}^{\mathrm{pop}} &= 1 + Am^{\mathrm{pop}} - h\theta^{\mathrm{pop}} = \sigma_{\mathrm{res}}^2 + 2Am^{\mathrm{pop}}. \end{aligned} \tag{168}$$

These residuals arise because a linear predictor cannot exactly represent the binary labels.

For explicit expressions in the model parameters, define the dimensionless quantity

$$\widetilde{\nu} = \frac{\nu}{\Delta} - \frac{\mu}{\Delta} u^2, \qquad \widetilde{\nu} - u^2 = \frac{\nu - \kappa^2/q_*}{\Delta} > 0. \tag{169}$$

Substituting (167) into (168) yields

$$\begin{aligned} \sigma_{\mathrm{res}}^2(\kappa) &= 1 - \bar{B}^2 - \frac{\widetilde{\nu} A^2 + 2uA\bar{B} - u^2\bar{B}^2}{1 + \widetilde{\nu}}, \\ \mathcal{L}_{\mathrm{dec}}^{\mathrm{pop}}(\kappa) &= 1 - \bar{B}^2 + \frac{\widetilde{\nu} A^2 + u^2\bar{B}^2}{1 + \widetilde{\nu}}. \end{aligned} \tag{170}$$

All quantities on the right are fixed by $(\Delta, \mu, \nu, q_*, \kappa)$. The shortcut coefficient can also be written as

$$c_{\mathrm{sc}}^{\mathrm{pop}} = \frac{A - u\bar{B}}{\Delta(1 + \widetilde{\nu})}. \tag{171}$$

At fixed (Δ, μ, ν, q_*), its numerator increases from zero because $\mathrm{d}(A - u\bar{B})/\mathrm{d}u = u^2\bar{B} > 0$ for $u > 0$, while its positive denominator is non-increasing. Hence increasing κ strictly increases $c_{\mathrm{sc}}^{\mathrm{pop}}$.

Step 3: isolate the finite-dataset contribution. Let $\ell_\perp = \nu - \kappa^2/q_* > 0$, and choose orthonormal coordinates with the second axis along w^* and the first along the component of v orthogonal to w^*. In these coordinates,

$$x_{i1} = \sqrt{\ell_\perp}\, s_i + \sqrt{\Delta}\, z_{i1}, \qquad x_{i2} = \frac{\kappa}{\sqrt{q_*}} s_i + \sqrt{V}\, z_{i2}, \qquad g_i = \sqrt{q_*}\, z_{i2}, \tag{172}$$

where z_{i1}, z_{i2} are independent standard Gaussians. The population covariance and label forcing restricted to this plane are

$$\Sigma_\parallel = \begin{pmatrix} \Delta + \ell_\perp & \kappa\sqrt{\ell_\perp/q_*} \\ \kappa\sqrt{\ell_\perp/q_*} & V + \kappa^2/q_* \end{pmatrix}, \qquad \bar{b}_\parallel = \begin{pmatrix} \sqrt{\ell_\perp}\, A \\ (\kappa A + q_* h)/\sqrt{q_*} \end{pmatrix}. \tag{173}$$

Write the design as $X = (X_\parallel, \sqrt{\Delta} Z)$, with $Z \in \mathbb{R}^{P\times(D-2)}$ standard Gaussian and independent of $(X_\parallel, y)$. For the normalized predictor $\beta = q/\sqrt{D}$, eliminate the bulk coefficients using $M_Z = I_P - Z(Z^\top Z)^{-1} Z^\top$:

$$\widehat{\beta}_\parallel = (X_\parallel^\top M_Z X_\parallel)^{-1} X_\parallel^\top M_Z y. \tag{174}$$

The projector has a uniformly distributed orientation independent of the signal-plane data, and $P^{-1} \operatorname{Tr} M_Z \to 1 - \alpha^{-1}$. Consequently, in probability,

$$\frac{1}{P} X_\parallel^\top M_Z X_\parallel \longrightarrow (1 - \alpha^{-1})\Sigma_\parallel, \qquad \frac{1}{P} X_\parallel^\top M_Z y \longrightarrow (1 - \alpha^{-1})\bar{b}_\parallel. \tag{175}$$

Thus $\widehat{\beta}_\parallel \to \Sigma_\parallel^{-1} \bar{b}_\parallel$, and the two signal coefficients converge to (167).

Let $\epsilon = y - X_\parallel \beta_\parallel^{\mathrm{pop}}$ be the vector of population-regression residuals. It is independent of Z, but can depend on $X_\parallel$, and $P^{-1}\|\epsilon\|^2 \to \sigma_{\mathrm{res}}^2$. The vanishing estimation error in the two signal coefficients contributes $o_p(1)$ to the bulk norm at fixed $\alpha > 1$. Hence

$$Q_{\mathrm{out}}(\infty) = \frac{1}{\Delta} \epsilon^\top Z (Z^\top Z)^{-2} Z^\top \epsilon + o_p(1) \longrightarrow \frac{\sigma_{\mathrm{res}}^2}{\Delta(\alpha - 1)}. \tag{176}$$

The last limit uses the independent Gaussian orientation of Z and $\operatorname{Tr}(Z^\top Z)^{-1} \to 1/(\alpha - 1)$. On a fresh sample from the training distribution, the signal-plane prediction converges to the population solution, and the orthogonal component contributes ΔQ_{out}. Hence

$$\mathcal{L}_{\mathrm{ID}}(\infty; \kappa) = \sigma_{\mathrm{res}}^2(\kappa) + \Delta Q_{\mathrm{out}}(\infty; \kappa) = \frac{\alpha}{\alpha - 1} \sigma_{\mathrm{res}}^2(\kappa). \tag{177}$$

At $\kappa = 0$, this reduces to $\alpha(1 - 2/\pi)/(\alpha - 1)$.

The covariance outside the signal plane is ΔI in both training and evaluation, so

$$\mathcal{L}_{\mathrm{dec}}(\infty; \kappa) = \mathcal{L}_{\mathrm{dec}}^{\mathrm{pop}}(\kappa) + \frac{\sigma_{\mathrm{res}}^2(\kappa)}{\alpha - 1}. \tag{178}$$

Together with (170), this gives the decorrelated risk for arbitrary admissible μ, κ.

Step 4: compare shortcut–label agreements. Fix (Δ, μ, ν, q_*) and take $0 \le \kappa_1 < \kappa_2 < \sqrt{\nu q_*}$. Define

$$\begin{aligned} p &= \mathcal{L}_{\mathrm{dec}}^{\mathrm{pop}}(\kappa_2) - \mathcal{L}_{\mathrm{dec}}^{\mathrm{pop}}(\kappa_1), \\ d &= \sigma_{\mathrm{res}}^2(\kappa_1) - \sigma_{\mathrm{res}}^2(\kappa_2). \end{aligned} \tag{179}$$

Both differences are explicit functions of the parameters through (170). Subtracting (178) between the two problems gives

$$\delta\mathcal{L}_{\mathrm{dec}}(\infty) = p - \frac{d}{\alpha - 1} = p\, \frac{\alpha - \alpha_c}{\alpha - 1}, \qquad \alpha_c = 1 + \frac{d}{p}, \tag{180}$$

where the factored form assumes $p \neq 0$. If $p, d > 0$, the larger-κ problem has worse population decorrelated risk but lower finite-dataset decorrelated risk for $1 < \alpha < \alpha_c$. The reduction in ΔQ_{out} outweighs the population-risk increase. Neither positivity condition is automatic for general μ, κ, and the threshold is not universally 2.

Step 5: establish the first-update ordering. Let $\bar{b} = \mathbb{E}[xy] = (Av + hw^*)/\sqrt{D}$ and $\bar{b}_{\mathrm{dec}} = hw^*/\sqrt{D}$ denote the training and decorrelated label forcings. Since $q(1)/\sqrt{D} = \eta P^{-1}\sum_i x_i y_i$ and $y_i^2 = 1$, averaging over the training dataset gives the exact finite-dimensional identity

$$\mathbb{E}_{\mathcal{D}}[\mathcal{L}_{\mathrm{dec}}(1;\kappa)] = 1 - 2\eta H_\kappa + \eta^2\left[\left(1 - \frac{1}{P}\right)J_\kappa + \frac{\operatorname{Tr}\Sigma_x^2}{P}\right], \tag{181}$$

where the two forcing contractions are

$$\begin{aligned} H_\kappa &= \bar{b}_{\mathrm{dec}}^\top \bar{b} = h(\kappa A + q_* h), \\ J_\kappa &= \bar{b}^\top \Sigma_x \bar{b} \\ &= \Delta(\nu A^2 + 2\kappa A h + q_* h^2) + (\nu A + \kappa h)^2 + \frac{\mu}{q_*}(\kappa A + q_* h)^2, \end{aligned} \tag{182}$$

and

$$\operatorname{Tr}\Sigma_x^2 = D\Delta^2 + 2\Delta(\nu + \mu) + \nu^2 + \mu^2 + \frac{2\mu\kappa^2}{q_*}. \tag{183}$$

In the proportional limit the first-update risk concentrates:

$$\mathcal{L}_{\mathrm{dec}}(1;\kappa) = 1 - 2\eta H_\kappa + \eta^2\left(J_\kappa + \frac{\Delta^2}{\alpha}\right). \tag{184}$$

The common sample-specific term $\eta^2\Delta^2/\alpha$ cancels in the comparison.

To determine the remaining sign, differentiate with respect to u at fixed (Δ, μ, ν, q_*). Since $A' = \bar{B}$ and $\bar{B}' = -u\bar{B}$,

$$\frac{\mathrm{d}H_\kappa}{\mathrm{d}u} = V\bar{B}\big[(1-u^2)A - u\bar{B}\big] < 0 \qquad (u > 0). \tag{185}$$

Indeed, $u\bar{B} - (1-u^2)A$ vanishes at zero and has derivative $2uA > 0$. To check J_κ, the derivatives of its three nonnegative components are controlled by

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}u}(\nu A^2 + 2\kappa A h + q_* h^2) &= 2A\bar{B}\,[\nu + V(1-u^2)], \\ \frac{\mathrm{d}}{\mathrm{d}u}(\nu A + \kappa h) &= \bar{B}\,[\nu + V(1-u^2)], \\ \frac{\mathrm{d}}{\mathrm{d}u}(\kappa A + q_* h) &= \sqrt{V q_*}\,A. \end{aligned} \tag{186}$$

Admissibility gives $Vu^2 = \kappa^2/q_* < \nu$, so these derivatives are positive for $u > 0$ and J_κ increases. It follows that

$$\delta\mathcal{L}_{\mathrm{dec}}(1) = -2\eta(H_{\kappa_2} - H_{\kappa_1}) + \eta^2(J_{\kappa_2} - J_{\kappa_1}) > 0. \tag{187}$$

Combining this inequality with (180) proves the reversal for $p, d > 0$ and $1 < \alpha < \alpha_c$. The first-update and converged limits establish at least one reversal, but not its time or uniqueness.

A nonzero-salience example. For $\Delta = q_* = 1$, $\nu = 128$, $\mu = 3$, and $(\kappa_1, \kappa_2) = (3, 5)$, the standardized shifts are $u_1 = 1.5$ and $u_2 = 2.5$. Eqs. (170) and (179) give

$$p \simeq 0.286690, \qquad d \simeq 0.153343, \qquad \alpha_c \simeq 1.534874. \tag{188}$$

Thus $\delta\mathcal{L}_{\mathrm{dec}}(\infty) \simeq -0.326682$ at $\alpha = 1.25$, but $+0.133347$ at $\alpha = 2$. With $\eta = 10^{-3}$, the first-update difference is $+0.006023$. These values follow directly from the analytical expressions and require neither fitted parameters nor added label noise.

D.5.1 Risk Reversal at Fixed Geometry and Core–Label Correlation

We keep the input distribution (1) fixed and replace the target by

$$y_A = As + \frac{B}{q_*}g + \sigma_\zeta(A)\zeta, \qquad \sigma_\zeta^2(A) = q_y - A^2 - \frac{B^2}{q_*} \geq 0, \qquad \zeta \sim \mathcal{N}(0,1), \tag{189}$$

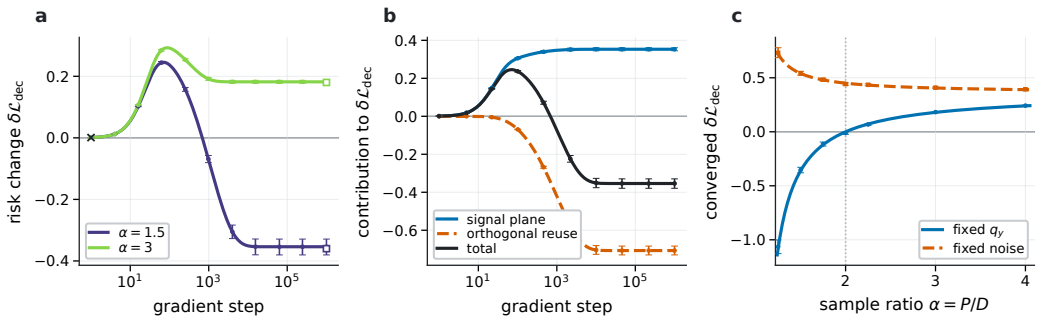

Fig. 9: **Risk reversal at fixed geometry and core–label correlation.** Both predictors use the same decorrelated test distribution. (a) At fixed q_y, the risk ordering reverses for $\alpha = 1.5$, but not for $\alpha = 3$ in the trajectory shown. Lines connect simulation means; the cross and open squares give the analytical first-update and converged predictions. (b) At $\alpha = 1.5$, reduced sample-specific fitting $\Delta\,\delta Q_{\text{out}}$ outweighs the signal-plane cost. (c) Converged analytical predictions (lines) and simulation means (markers): the fixed-q_y comparison crosses zero at $\alpha = 2$, while the fixed independent-noise comparison remains positive. Error bars are paired 95% bootstrap intervals. *Parameters:* App. B.

where ζ is independent of all input variables. This real-valued target has $\mathbb{E}[y_A s] = A$, $\mathbb{E}[y_A g] = B$, and $\mathbb{E}[y_A^2] = q_y$. We compare $0 \le A_1 < A_2$ at fixed B, q_y, with $A_2^2 + B^2/q_* \le q_y$, and use one common test distribution:

$$y_{\text{test}} = \frac{B}{q_*} g + \sqrt{q_y - B^2/q_*}\,\zeta_{\text{test}}, \qquad \zeta_{\text{test}} \sim \mathcal{N}(0,1), \tag{190}$$

with fresh inputs of the same law and independent test noise. Hence $A_e = 0, B_e = B$, and the test label second moment is q_y for both predictors. Increasing A at fixed B, q_y replaces independent target noise with shortcut signal; it does not change feature geometry or the core–label correlation.

Assume $\kappa = 0$, $\Delta, \nu, q_* > 0$, fixed $\mu \ge 0$, $\lambda_{\text{L2}} = 0$, $q(0) = 0$, and stable full-batch gradient descent. As above, the converged calculation takes the finite-dataset least-squares solution and then the proportional limit at fixed $\alpha > 1$.

Core response and population risk. Square-loss dynamics is linear in the training targets. At $\kappa = 0$, reflection of the core coordinate cancels the mean core response to As, and the independent target noise has zero mean response. Thus, even with reuse, $\mathbb{E}_{\mathcal{D}}[c_{\text{core}}(t)]$ is independent of A at every update; this is an expectation over datasets, not an identity for individual datasets. At convergence, the signal coefficients approach

$$c_{\text{sc}}^{\text{pop}} = \frac{A}{\Delta + \nu}, \qquad c_{\text{core}}^{\text{pop}} = \frac{B}{q_* \sqrt{\Delta + \mu}}. \tag{191}$$

The population training residual variance and common-test risk are therefore

$$\begin{aligned} \sigma_{\text{res}}^2(A) &= \sigma_\zeta^2(A) + \frac{A^2\Delta}{\Delta + \nu} = q_y - \frac{B^2}{q_*} - \frac{A^2\nu}{\Delta + \nu}, \\ \mathcal{L}_{\text{dec}}^{\text{pop}}(A) &= q_y - \frac{B^2}{q_*} + \frac{A^2\nu}{\Delta + \nu}. \end{aligned} \tag{192}$$

Stronger shortcut correlation increases the shortcut coefficient and population decorrelated risk, while reducing the training residual variance.

Finite-dataset risk reversal. The auxiliary targets and their population residuals remain independent of the Gaussian coordinates outside the signal plane. The projection argument in (174)–(176) therefore applies unchanged and gives

$$\begin{aligned} \Delta Q_{\text{out}}(\infty; A) &= \frac{\sigma_{\text{res}}^2(A)}{\alpha - 1}, \\ \mathcal{L}_{\text{dec}}(\infty; A) &= q_y - \frac{B^2}{q_*} + \frac{A^2\nu}{\Delta + \nu} + \frac{q_y - B^2/q_* - A^2\nu/(\Delta + \nu)}{\alpha - 1}. \end{aligned} \tag{193}$$

Writing $\delta\mathcal{L}_{\mathrm{dec}}(t) = \mathcal{L}_{\mathrm{dec}}(t; A_2) - \mathcal{L}_{\mathrm{dec}}(t; A_1)$, the population-risk increase p and residual-variance reduction d of (179) satisfy

$$p = d = \frac{(A_2^2 - A_1^2)\nu}{\Delta + \nu} > 0, \qquad \delta\mathcal{L}_{\mathrm{dec}}(\infty) = p\,\frac{\alpha - 2}{\alpha - 1}. \tag{194}$$

Thus stronger shortcut correlation reduces the common-test risk for $1 < \alpha < 2$, despite greater shortcut reliance and an unchanged limiting core coefficient. The two risks agree at $\alpha = 2$; their ordering is reversed for $\alpha > 2$. No claim is made at $\alpha = 1$.

Ordering during training. Both risks equal q_y at initialization. Let $V = \Delta + \mu$. For the first update, the diagonal second-moment contribution is

$$\mathbb{E}[y_A^2 x^\top \Sigma_x x] = q_y\big[(D-2)\Delta^2 + (\Delta + \nu)^2 + V^2\big] + \frac{2V^2B^2}{q_*},$$

which is independent of A. Averaging the squared first-update predictor over the training dataset therefore gives the finite-dimensional identity

$$\mathbb{E}_{\mathcal{D}}[\delta\mathcal{L}_{\mathrm{dec}}(1)] = \left(1 - \frac{1}{P}\right)\eta^2(A_2^2 - A_1^2)\nu(\Delta + \nu). \tag{195}$$

In the proportional limit, the risk difference concentrates at the positive value $\eta^2(A_2^2 - A_1^2)\nu(\Delta + \nu)$. Combined with (194), this establishes a reversal during training for every fixed $1 < \alpha < 2$, without determining its time or uniqueness. Population-gradient descent instead satisfies

$$\delta\mathcal{L}_{\mathrm{dec}}^{\mathrm{pop}}(t) = p\left[1 - \{1 - \eta(\Delta + \nu)\}^t\right]^2 \geq 0, \tag{196}$$

so it does not exhibit this reversal.

Control at fixed independent target-noise variance. To distinguish replacing noise from adding shortcut signal, instead train on $y_A^{(\sigma)} = As + (B/q_*)g + \sigma\zeta$ with σ fixed and retain the same test distribution (190). The training label second moment now increases with A, and the population residual variance is $\sigma^2 + A^2\Delta/(\Delta + \nu)$. Consequently,

$$\delta\mathcal{L}_{\mathrm{dec}}^{(\sigma)}(\infty) = \frac{A_2^2 - A_1^2}{\Delta + \nu}\left(\nu + \frac{\Delta}{\alpha - 1}\right) > 0. \tag{197}$$

The converged benefit therefore depends on replacing independent target noise with shortcut signal, not on less shortcut reliance or improved mean core recovery. Fig. 9 compares the two interventions.

E Numerical Methods and Experiments

We first describe the numerical solvers and their validation, then report finite-dimensional tests of the theoretical results. The image-benchmark protocols and the limits of quantitative transfer follow.

E.1 Solvers and Computational Cost

E.1.1 Causal Forward Solver

The theory curves use the deterministic correlation–response recursions in Apps. C.3 and C.4. At each time, the known predictor history determines the residual row; the gradient moments then advance the predictor or architecture state. Persistent updates use the active moments for the gradient and the full moments for training risk. There is no fixed-point iteration over the complete trajectory.

For an independent numerical check, we also sample the effective Gaussian processes. The sampled reference implementation proceeds as follows:

Input:	$T,\ M,\ K,$ architecture, geometry, $\eta, \alpha, \omega, \gamma_0$.
Initialize:	static geometry samples, initial network fields, Gaussian source states.
For $t = 0, \dots, T$:	
1.	Build $q(t)$ from past $h_0(s)$ and architecture responses.
2.	Update $C_q(t,\cdot)$ and $R^a_{qh}(t,\cdot)$.
3.	For each component $a,$ update $\Delta_a(t), C^a_\Delta(t,\cdot), R^a_\Delta(t,\cdot)$.
4.	If SGD is active, update $Z_a(t), C^a_Z(t,\cdot), R^a_Z(t,\cdot)$.
5.	Sample the component gradient sources and assemble $h_0(t)$.
6.	Update hidden-layer fields and responses, if present.

Causality applies to both implementations. The closed recursions additionally eliminate the Monte Carlo population used by the sampled reference.

E.1.2 Validation of the Closed Shallow Equations

We checked the closed recursion in App. C.2 against the sampled shallow DMFT implementation in five settings: the reference geometry, unequal variances, unequal weights, more correlated teachers, and a larger α. With $T = 80$ and sampled population $N_{\mathrm{pop}} = 65536$, the training-risk relative RMSE lies between 1.09×10^{-3} and 2.11×10^{-3}, the population-risk relative RMSE lies between 6.39×10^{-5} and 1.97×10^{-3}, and the closed recursion is between $55\times$ and $106\times$ faster on the local CPU used for this check.

E.1.3 Adaptive Time-Grid Validation

The fixed-step recursion above is an accurate reference, but its dense two-time arrays make long horizons impractical. We therefore implemented an adaptive time-grid solver based on DYNAMITE (Lang et al., 2026) for the continuous-time limit of the shallow closure. Let $0 = \tau_0 < \tau_1 < \cdots < \tau_N$ denote the accepted physical times, with $\tau = \eta t$. Every causal memory integral is evaluated by composite trapezoidal quadrature,

$$\int_0^{\tau_n} f(\tau_n, s)\,\mathrm{d}s \simeq \sum_{j=0}^{n} w_j^{(n)} f(\tau_n, \tau_j), \qquad w_j^{(n)} = \begin{cases} (\tau_1 - \tau_0)/2, & j = 0, \\ (\tau_{j+1} - \tau_{j-1})/2, & 0 < j < n, \\ (\tau_n - \tau_{n-1})/2, & j = n. \end{cases} \tag{198}$$

The singular equal-time responses must not be interpolated numerically. We instead write the dense response conventions as

$$R^a_{qh}(t,s) = \frac{\eta}{2} G(\tau_t, \tau_s) + o(\eta), \qquad R^a_\Delta(t,s) = -\delta_{ts} + \eta R^a_{\Delta,\mathrm{reg}}(\tau_t, \tau_s) + o(\eta) \qquad (t > s), \tag{199}$$

and impose the causal boundary $G(\tau,\tau) = 1$ analytically. The regular part $R^a_{\Delta,\mathrm{reg}}$ is then the solution of the nonsingular CHSCK equation obtained from (55). This separation is essential: a naive interpolation of the discrete impulse produces an error of order one near the diagonal.

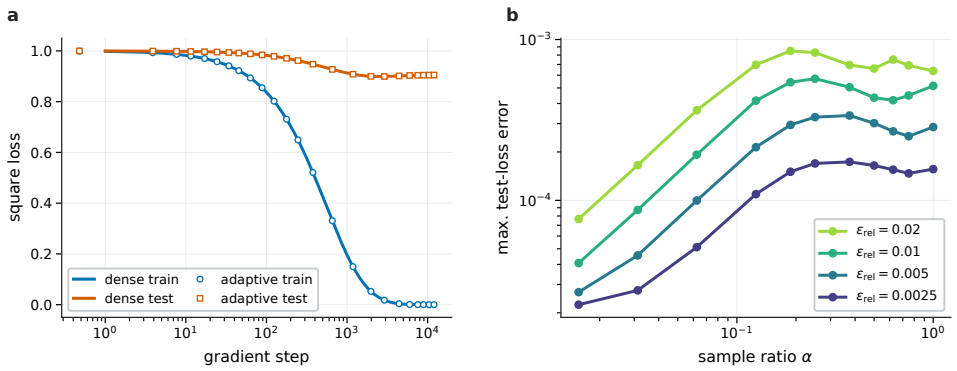

Fig. 10: Validation of the adaptive time grid for the closed shallow full-batch DMFT. (a) Dense fixed-step curves and adaptive values at the stored nodes for the shortcut geometry at $\alpha = 1/8$, with $\epsilon_{\mathrm{rel}} = 2.5 \times 10^{-3}$. (b) Maximum population-risk error over time as the sample ratio α and relative tolerance ϵ_{rel} vary. *Parameters:* App. B.

For each trial interval $h_n = \tau_{n+1} - \tau_n$, we evolve the one-time overlaps and the new rows of C_q and G with an embedded Euler–Heun pair. If Y^{E}_{n+1} and Y^{H}_{n+1} are the predicted and corrected rows, respectively, the accepted error is

$$e_n = \max_i \frac{|Y^{\mathrm{H}}_{n+1,i} - Y^{\mathrm{E}}_{n+1,i}|}{\epsilon_{\mathrm{abs}} + \epsilon_{\mathrm{rel}} \max(|Y_{n,i}|, |Y^{\mathrm{H}}_{n+1,i}|)}. \tag{200}$$

The row is accepted when $e_n \le 1$; the next step is multiplied by $0.9 e_n^{-1/2}$, clipped to the prescribed minimum and maximum factors. After acceptance, the data block is closed on the new row in the same causal order as (55)– (70).

We validated this discretization in four ways. First, we compared five one-time observables (training and population risks, M, Θ, and $C_q(t,t)$) against eleven dense trajectories with $\alpha \in [1/64, 1]$, $\eta = 10^{-3}$, and $T = 12000$, at four adaptive tolerances. At $\epsilon_{\mathrm{rel}} = 2.5 \times 10^{-3}$, the worst absolute error over all 55 curves is 1.59×10^{-3}; the worst population-risk error is 1.73×10^{-4}. Fig. 10 compares one trajectory and the maximum population-risk errors across sample ratios and tolerances. The largest relative RMSE, 6.05×10^{-3}, occurs for the training risk at $\alpha = 1/64$, where the reference value is numerically zero. Second, changing the maximum accepted interval from 0.125 to 1 changes the representative population-risk error by only 9.46×10^{-5}; the controller itself limits the largest attempted interval to 0.323. Third, reducing the dense step from 8×10^{-3} to 10^{-3} reduces its population-risk discrepancy from 1.66×10^{-3} to 2.24×10^{-4}, consistent with convergence of the first-order dense update to the adaptive continuum solution. Finally, we compared the complete C_q, C^a_Δ, G, and $R^a_{\Delta,\mathrm{reg}}$ kernels for five geometries. The worst kernel relative RMSE is 2.64×10^{-3}. For response kernels we omit lags below four dense steps, where the finite-step impulse convention in (199) rather than the regular continuum kernel controls the comparison.

For the representative $\alpha = 1/8$ trajectory, the tight adaptive run stores 158 time nodes and takes 4.77 seconds. The dense reference stores 12001 times, takes 5597 seconds, and reaches 13.5 GiB peak RSS. The complete validation process, including all short dense-kernel checks and plotting, reaches 197 MiB peak RSS with a cap of 1024 adaptive nodes. The adaptive grid reduces memory and runtime on these smooth shallow trajectories. This implementation still stores a triangular history and is cubic in the number of accepted nodes. It does not yet use a fixed $\theta = \tau'/\tau$ grid, high-order two-dimensional interpolation, or residual-based history sparsification, and therefore does not attain the full DYNAMITE complexity result. These checks concern the shallow full-batch equations; they do not validate the adaptive solver for hidden layers or persistent SGD.

E.1.4 FINITE-DIMENSIONAL SIMULATIONS

The finite simulations use the same geometry as the DMFT equations. A generic finite run is:

Input: $D, \alpha, \eta, T, \Delta, v, w^*, \mu, \kappa$.

1. Build the joint Gram matrix for centroids and teachers and check it is PSD.
2. Sample vectors $v_a, w_a^\star \in \mathbb{R}^D$ from that Gram matrix.
3. Draw $P/2 = \alpha D/2$ examples for each $a \in \{-1, +1\}$.
4. Generate $x_{a\ell}$ from (1) and $y_{a\ell} = \mathrm{sign}((w^*)^\top x_{a\ell}/\sqrt{D})$.
5. Run full-batch GD or persistent SGD on the fixed dataset.
6. At checkpoints, compute training and population risks/accuracies for each component.

For shallow full-batch simulations, the finite trajectory can be evaluated exactly at checkpoints by diagonalizing the empirical Hessian (or the dual Gram matrix when $P < D$). This removes time-stepping error from the finite simulation and leaves only finite-dimensional dataset fluctuations.

E.1.5 COMPLEXITY

Let T be the number of gradient steps, M the DMFT population size, K the number of mixture components, and L the number of hidden layers. For a fixed architecture, the dense closed recursions store two-time arrays of order $O((K+L)T^2)$, with causal history sums costing up to $O((K+L)T^3)$. The independently sampled full-batch implementation stores $O(KT^2)$ component order parameters plus $O(MT)$ sampled fields for each active Gaussian process. The causal Gaussian samplers require triangular covariance operations, giving an order $O(S(MT^2 + T^3))$ cost for S Gaussian sources, with additional architecture-dependent response multiplications.

In the sampled persistent-SGD implementation, responses are averaged at fixed activity history. This reference implementation stores

$$O(KMT^2)$$

sample-level response entries. Its dominant additional cost is $O(KMT^2)$ arithmetic for the active residual recursion. The resulting cost limits this implementation to short horizons; longer horizons will require a compressed or checkpointed response representation.

Finite simulations scale with the explicit width. For $N = \omega D$ and $P = \alpha D$, a hidden linear network costs roughly

$$O\big(T\big[\alpha D^2 + \omega D^2 + L\omega^2 D^2\big]\big)$$

per initialization, with the first term from data matrix products and the remaining terms from forward/backward hidden matrix products and rank-one updates.

E.2 SYNTHETIC-DATA VALIDATION

These experiments test the architecture and sampling extensions and the effects of sample ratio, shortcut–label agreement, and shortcut salience.

E.2.1 VALIDATION ACROSS DEPTH AND PERSISTENT SGD

We validate the DMFT after factorizing the predictor into a deep linear network and replacing the full gradient by a persistent active subset. A single causal row sweep combines the architecture kernels of App. C.4.1 with the active/full data moments of App. C.4.2.

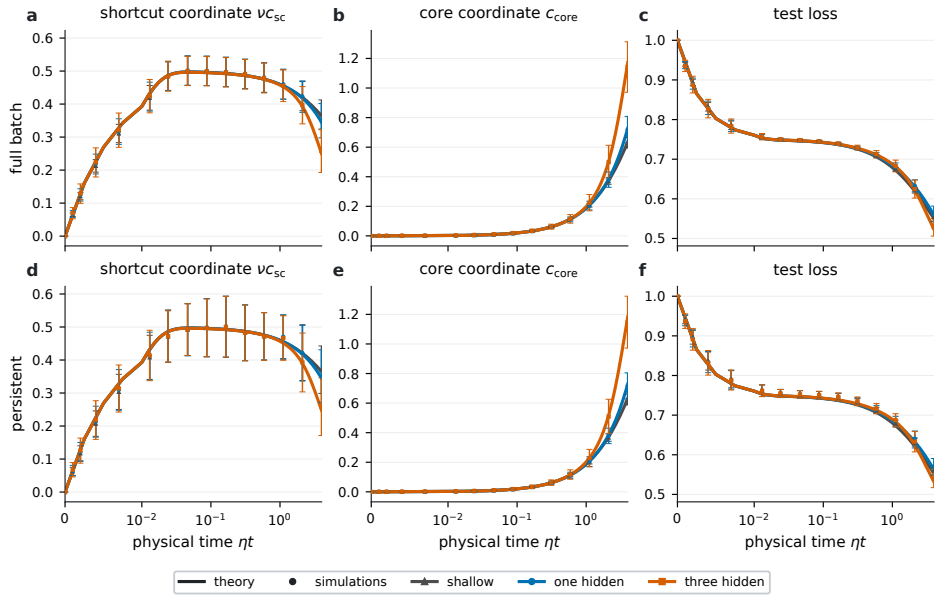

Fig. 11: Closed-DMFT validation across architecture depth and update rule. The top row uses full-batch gradient descent and the bottom row persistent SGD. Columns show the rescaled shortcut coefficient νc_{sc}, core coefficient c_{core}, and population risk. Solid curves are the deterministic causal closure; markers and error bars are means $\pm$ one standard deviation over finite simulations. Gray triangles, blue circles, and orange squares denote a shallow predictor, one hidden layer, and three hidden layers, respectively. All predictors start at $q(0) = 0$; factorized networks use a zero final readout. *Parameters:* App. B.

Quantitative agreement. In Fig. 11, the closed equations agree with the simulations throughout the displayed interval. Population-risk trajectory RMSE is 1.99–2.27×10^{-3} for full-batch dynamics and 6.37–6.84×10^{-3} under persistent SGD. Across all eighteen architecture–observable pairs, final-time discrepancies are below 0.56 simulation standard deviations and at least 99.8% of saved points lie within one standard deviation. The three-hidden persistent closure reaches population risk 0.5313 at time four, consistent with the large-width finite-network value 0.5304.

Effects of depth and persistence. For these parameters, the three-hidden-layer model learns the core faster, reaches the largest core coefficient, and has the lowest population risk at the end of the displayed interval. The early shortcut coefficients are similar across architectures. Persistent SGD leaves the two coefficients near their full-batch values while increasing population risk. These results test the architecture-dependent and activity-dependent kernels in one setting; they do not establish a general effect of depth or persistence.

E.2.2 Complete Finite-Reuse Sweeps

Fig. 12 compares the shortcut and core coefficients with Q_{out} while varying one parameter at a time. The columns vary sample ratio, shortcut–label agreement, and shortcut salience; the rows distinguish changes in the two coefficients from sample-specific fitting outside the signal plane.

Increasing ρ raises A while lowering B, so c_{sc} grows while c_{core} falls. Increasing ν changes the coefficient needed to represent a fixed shortcut contribution but leaves late core recovery, Q_{out}, and the risks nearly unchanged in these experiments.

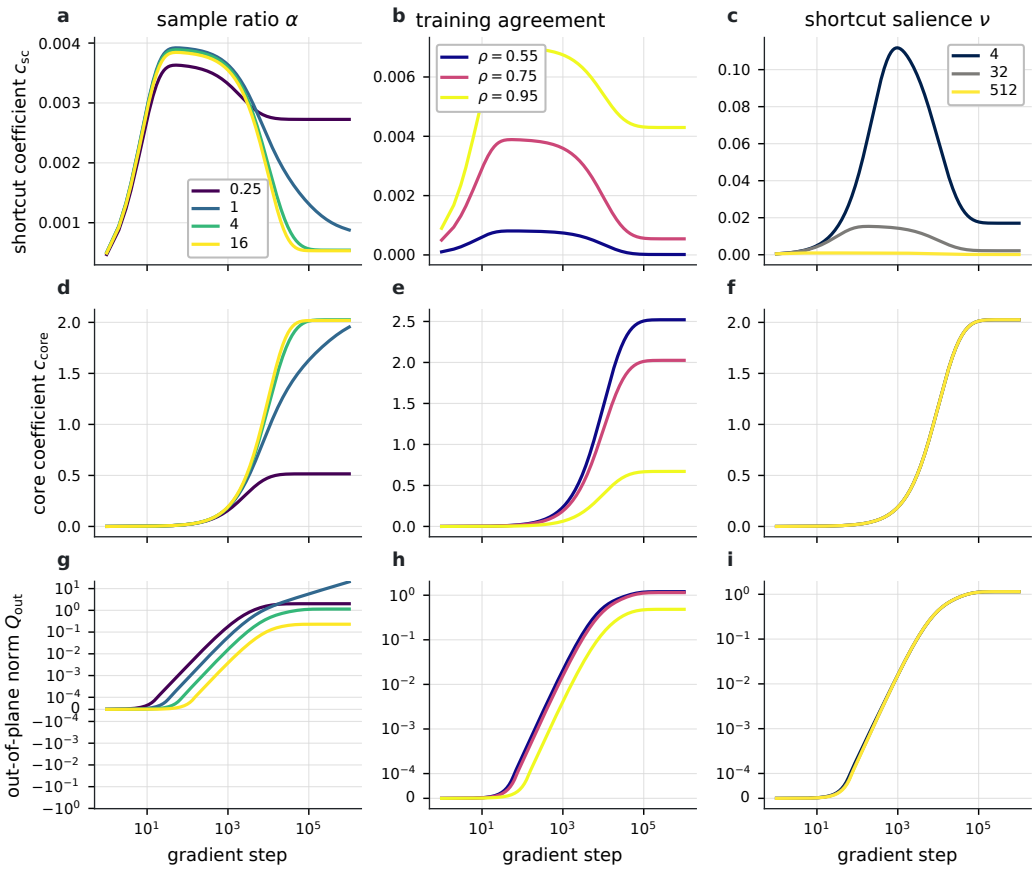

Fig. 12: Coefficient and sample-specific-fitting trajectories. Columns vary $\alpha \in \{0.25, 1, 4, 16\}$, $\rho \in \{0.55, 0.75, 0.95\}$, and $\nu \in \{4, 32, 512\}$, respectively. Rows report c_{sc}, c_{core}, and Q_{out}; curves are means over independently sampled training datasets. *Parameters:* App. B.

E.3 Image-Benchmark Experiments

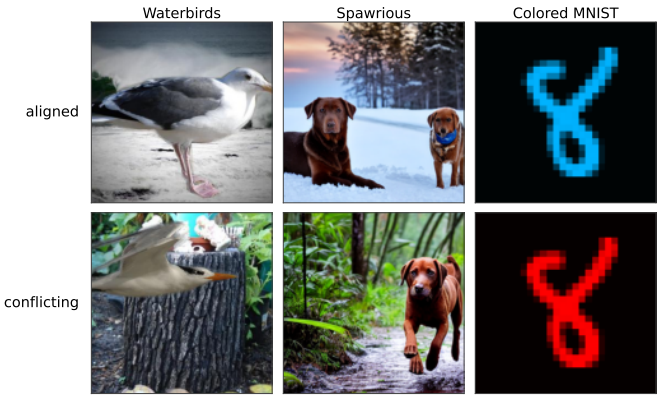

Fig. 13: Held-out image-benchmark examples with the target fixed and the shortcut aligned (first row) or conflicting (second row). Columns show Waterbirds, Spawrious, and Controlled Colored MNIST. The Colored MNIST pair uses the same digit under opposite color assignments. The examples are selected deterministically from the evaluation partitions.

E.3.1 Random-Feature Regression on Waterbirds and Spawrious

For Waterbirds, y is bird type and s is background type. For Spawrious, y distinguishes Labrador from Bulldog and s distinguishes snow from jungle. Each dataset is split within the four (y, s) groups into 15% calibration, 15% evaluation, and 70% training-pool records. Images are resized to 64×64 and mapped to $D = 512$ by a fixed Gaussian projection, either from the full image or from separate foreground and background images. Centering, scaling, and the moments used in the DMFT are estimated from the calibration split alone. The representation and calibration below specify the Gaussian approximation used for these benchmarks; the image labels are not generated by the teacher in Sec. 2.

Each training dataset contains $P = \alpha D$ examples drawn with the prescribed group counts. A zero-initialized, no-intercept linear predictor is trained for 10^6 fixed-rate full-batch square-loss steps, without momentum, regularization, schedules, or early stopping. We use $\alpha \in \{0.5, 1, 2, 4\}$, training agreement $p_{\mathrm{train}} \in \{0.5, 0.9, 0.95\}$, both representations, and five independently selected training datasets, giving 48 conditions and 240 finite trajectories. The full/split step sizes are $0.0016520/0.0023198$. The projection matrices and image partitions are fixed across these conditions. The five realisations vary the training dataset, not the calibration split.

Image representation. Let $I_i \in [0,1]^L$ be the flattened RGB image, with $L = 3 \times 64^2$, and let M_i be its foreground mask, repeated across the three colour channels. The full and split representations, together with their foreground-only counterparts, are

$$\begin{aligned} \phi_i^{\mathrm{full}} &= R^\top I_i, & \phi_i^{\mathrm{fg,full}} &= R^\top (M_i \odot I_i), \\ \phi_i^{\mathrm{split}} &= \begin{pmatrix} R_f^\top (M_i \odot I_i) \\ R_b^\top ((1 - M_i) \odot I_i) \end{pmatrix}, & \phi_i^{\mathrm{fg,split}} &= \begin{pmatrix} R_f^\top (M_i \odot I_i) \\ 0_{D/2} \end{pmatrix}. \end{aligned} \tag{201}$$

Here $\odot$ denotes entrywise multiplication; $R \in \mathbb{R}^{L \times D}$ and $R_f, R_b \in \mathbb{R}^{L \times D/2}$ have independent $\mathcal{N}(0, 1/L)$ entries, generated with seeds $2026080401, 2026080402, 2026080403$, respectively. The frozen DeepLabV3-ResNet50 segmenter supplies bird probabilities for Waterbirds and dog probabilities for Spawrious. Its input is the 224×224 RGB image with ImageNet normalization, and its probabilities are bilinearly resized to 64×64. If their spatial mean is outside $[0.02, 0.70]$, the mask is replaced by the indicator of probabilities at or above their 80th percentile. This fixed per-image fallback does not use the calibration labels.

For each representation separately, let n_{cal} be the number of calibration images. We use

$$\begin{gathered} \bar{\phi} = \frac{1}{n_{\mathrm{cal}}} \sum_{i \in \mathrm{cal}} \phi_i, \qquad a_x = \left[\frac{D}{n_{\mathrm{cal}}^{-1} \sum_{i \in \mathrm{cal}} \|\phi_i - \bar{\phi}\|^2} \right]^{1/2}, \\ x_i = a_x(\phi_i - \bar{\phi}), \qquad x_i^{\mathrm{fg}} = a_x(\phi_i^{\mathrm{fg}} - \bar{\phi}). \end{gathered} \tag{202}$$

Thus the calibration average of $\|x\|^2/D$ is one. These preprocessing averages retain the calibration split's group proportions; the moment estimates below instead reweight its four groups.

Calibration of the signal plane. Write p for the achieved shortcut–label agreement after rounding the training group counts: $p = \mathrm{round}(p_{\mathrm{train}} P/2)/(P/2)$. For any sample quantity F, define

$$\langle F \rangle_p = \sum_{y,s=\pm 1} \omega_{ys}(p) \frac{1}{n_{ys}} \sum_{i \in \mathrm{cal}:(y_i,s_i)=(y,s)} F_i, \qquad \omega_{ys}(p) = \frac{1 + ys(2p-1)}{4}, \tag{203}$$

where n_{ys} is the number of calibration images in group (y, s). Both y and s are balanced under this average. Let $\bar{x}_{ys}^{\mathrm{fg}}$ be the foreground-feature mean in that group. The two directions are estimated as

$$\begin{gathered} \bar{v} = \langle sx \rangle_p, \qquad d_{\mathrm{core}} = \frac{1}{2} \sum_{s=\pm 1} (\bar{x}_{+,s}^{\mathrm{fg}} - \bar{x}_{-,s}^{\mathrm{fg}}), \qquad e = \frac{d_{\mathrm{core}}}{\|d_{\mathrm{core}}\|}, \\ v = \sqrt{D}\, \bar{v}, \qquad w^* = \sqrt{D}\, e, \qquad \nu = \|\bar{v}\|^2, \quad \kappa = \bar{v}^\top e, \quad q_* = 1. \end{gathered} \tag{204}$$

The foreground class contrast fixes the orientation of the core direction. The estimated w^* is a core proxy, not a teacher assumed to reproduce every image label. The shortcut direction is a conditional

image contrast and can contain foreground information. It depends on p, whereas e is fixed within a representation. Within each condition, the DMFT and image predictor use the same calibrated basis. Their coefficient comparison therefore uses the same directions and normalization. Across agreement levels, both the trained predictor and the shortcut direction change: the paired changes compare calibrated conditions, not a fixed pair of features.

Variance and label moments. Define the image residual $\epsilon_i = x_i - s_i\bar{v}$ and its core projection $\zeta_i = \epsilon_i^\top e$. The estimates are

$$\Delta = \max\left\{\frac{\langle\|\epsilon\|^2\rangle_p - \langle\zeta^2\rangle_p}{D-1}, 10^{-8}\right\}, \qquad \mu = \max\{\langle\zeta^2\rangle_p - \Delta, 0\},$$
$$A = \langle ys\rangle_p = 2p-1, \qquad \beta_\parallel = \langle y\zeta\rangle_p, \qquad q_y = 1. \tag{205}$$

These are residual second moments, without an additional recentering after group reweighting. The measured correlation $\beta_\parallel$ is not the latent-field moment B of (3): under the Gaussian model with $q_* = 1$, it would equal $\sqrt{\Delta+\mu}\,B$.

Gaussian approximation used in the dynamics. The image predictions use the component of the measured core direction orthogonal to $\bar{v}$:

$$h_\perp = \sqrt{1-\kappa^2/\nu}, \qquad e_\perp = \frac{e-(\kappa/\nu)\bar{v}}{h_\perp}, \qquad \zeta_{\perp,i} = \epsilon_i^\top e_\perp,$$
$$\mu_\perp = \max\{\langle\zeta_\perp^2\rangle_p - \Delta, 0\}, \qquad \beta_s = \langle y\zeta_\perp \mid s\rangle_p. \tag{206}$$

The construction requires $\nu > 0$ and $h_\perp > 0$. The DMFT uses uniform component signs, means $s\bar{v}$, isotropic variance Δ, and additional variance $\mu_\perp$ along $e_\perp$. Its covariance and supplied label moments are

$$\Sigma_{\text{eff}} = \Delta I + \bar{v}\bar{v}^\top + \mu_\perp e_\perp e_\perp^\top,$$
$$\mathbb{E}_{\text{eff}}[y \mid s] = sA, \qquad \mathbb{E}_{\text{eff}}[y\zeta_\perp \mid s] = \beta_s, \qquad \mathbb{E}_{\text{eff}}[y^2 \mid s] = 1. \tag{207}$$

The moments are substituted in the square-loss closure. This approximation replaces the residual second-moment matrix by $\Delta I + \mu_\perp e_\perp e_\perp^\top$, sets the common reweighted image mean to zero, and discards image–label correlations outside the retained directions. In particular, the covariance spike is along $e_\perp$, not along e as it would be in (1) with $w^* = \sqrt{D}e$. Reconstructing the observables in the original basis does not remove this approximation.

The relations between A, B and the geometry derived in App. D.5 are not imposed on the image moments. In the orthogonal geometry used by this solver, the deterministic teacher would give $A = 0$ and $B = \sqrt{2/\pi}$. Instead, the image calculation uses $A = 2p-1$ and conditional core moments $B_s = \beta_s/\sqrt{\Delta+\mu_\perp}$. Imposing the teacher relations would additionally assume that the measured foreground contrast generates the image labels. The approximation instead retains the measured label correlations, while the image predictor is trained on the original benchmark labels.

For square loss, this choice has a direct connection to the training dynamics. The exact full-batch update, without ridge, is

$$q(t+1) = (I - \eta\widehat{\Sigma}_x)q(t) + \eta\sqrt{D}\,\widehat{b}_x,$$
$$\widehat{\Sigma}_x = \frac{1}{P}\sum_{i=1}^{P} x_i x_i^\top, \qquad \widehat{b}_x = \frac{1}{P}\sum_{i=1}^{P} x_i y_i. \tag{208}$$

Thus the empirical input second moments and input–label correlations determine the trajectory. The calibration approximates their population structure through the two directions, residual variances, and label correlations in (207); the finite-dataset DMFT retains the dependence across updates caused by reusing the same examples.

Readout coefficients. The trained readout has the same convention as Sec. 2.1, $f(x,t) = q(t)^\top x/\sqrt{D}$. For each condition, both the image predictor and its DMFT prediction use the cal-

ibrated Gram matrix to compute

$$m(t) = \frac{q(t)^\top \bar{v}}{\sqrt{D}}, \qquad \theta(t) = \frac{q(t)^\top e}{\sqrt{D}}, \qquad G = \begin{pmatrix} \nu & \kappa \\ \kappa & 1 \end{pmatrix},$$
$$\begin{pmatrix} c_{\mathrm{sc}}(t) \\ c_{\mathrm{core}}(t) \end{pmatrix} = G^{-1} \begin{pmatrix} m(t) \\ \theta(t) \end{pmatrix}, \qquad Q_{\mathrm{out}}(t) = Q(t) - \begin{pmatrix} m(t) \\ \theta(t) \end{pmatrix}^\top G^{-1} \begin{pmatrix} m(t) \\ \theta(t) \end{pmatrix}. \tag{209}$$

The solver returns m, Q, and $\theta_\perp = q^\top e_\perp/\sqrt{D}$; its core overlap is converted back through $\theta = (\kappa/\nu)m + h_\perp \theta_\perp$. Negative values of Q_{out} from this subtraction are clipped to zero. For real images, directions outside this estimated plane can carry population signal; their contribution to Q_{out} need not arise solely from sample-specific fitting.

Population least-squares reference. The population solution of the same approximation separates the effect of the calibrated moments from that of a finite dataset. With $\bar{\beta} = (\beta_{-1} + \beta_{+1})/2$, its forcing is $\mathbb{E}_{\mathrm{eff}}[xy] = A\bar{v} + \bar{\beta} e_\perp$. Hence

$$\frac{q_{\mathrm{pop}}}{\sqrt{D}} = a_{\mathrm{pop}} \bar{v} + b_{\mathrm{pop}} e_\perp, \qquad a_{\mathrm{pop}} = \frac{A}{\Delta + \nu}, \qquad b_{\mathrm{pop}} = \frac{\bar{\beta}}{\Delta + \mu_\perp},$$
$$c_{\mathrm{core}}^{\mathrm{pop}} = \frac{b_{\mathrm{pop}}}{h_\perp}, \qquad c_{\mathrm{sc}}^{\mathrm{pop}} = a_{\mathrm{pop}} - \frac{\kappa b_{\mathrm{pop}}}{\nu h_\perp}, \qquad Q_{\mathrm{out}}^{\mathrm{pop}} = 0. \tag{210}$$

For zero-initialized population gradient descent, the finite-time reference replaces a_{pop} by $a_{\mathrm{pop}}[1 - (1 - \eta(\Delta + \nu))^t]$ and b_{pop} by $b_{\mathrm{pop}}[1 - (1 - \eta(\Delta + \mu_\perp))^t]$. An increase in shortcut use with agreement is already possible in this population solution. The additional finite-dataset prediction concerns the dependence on α: at fixed calibrated moments, the population reference is independent of α and has $Q_{\mathrm{out}} = 0$, whereas reused data produce fitting outside the signal plane and, for $\alpha < 1$, change the signal coefficients through projection onto the span of the training inputs. For $\alpha > 1$, the converged signal coefficients of the Gaussian approximation recover their population values. Agreement of the final coefficients at $\alpha = 4$ therefore tests the calibrated population response, not an improvement over population least squares.

The illustrative examples in Fig. 13 are selected deterministically from the evaluation partitions with target $y = +1$. Waterbirds and Spawrious use one $s = +1$ and one $s = -1$ image. Colored MNIST uses the same held-out digit eight recolored under aligned and conflicting assignments.

Paired comparisons. Table 3 summarises changes at the final checkpoint, first averaging the image results over five training datasets within each condition. The bounded change is then computed between condition means, not separately between training realisations. The agreement comparison uses $p_{\mathrm{train}} : 0.50 \to 0.95$ at each dataset, representation, and $\alpha \in \{0.5, 4\}$; the sample-ratio comparison uses $\alpha : 0.5 \to 4$ at each dataset, representation, and $p_{\mathrm{train}} \in \{0.50, 0.95\}$; the representation comparison uses full $\to$ split at each dataset, sample ratio, and agreement. Each intervention thus contains eight condition pairs. The full-to-split changes compare separately normalized representations and separately calibrated directions.

The table gives a directional summary of these interventions. Both the DMFT and image simulations have a positive median bounded change in the shortcut coefficient when agreement increases, and a negative one in Q_{out} when α increases from 0.5 to 4. These observations are consistent with the calibrated approximation, without establishing that all out-of-plane fitting in images is sample-specific. The raw full-image coefficients are shown in Fig. 6(a,b), without the bounded transformation: the shortcut response has the same direction, while the magnitude of the core response is less accurately predicted.

E.3.2 Controlled Colored MNIST

The target is $y = \mathbf{1}\{\mathrm{digit} \geq 5\}$, and the binary color equals y on shortcut-aligned examples and $1 - y$ on conflicting examples. Training sets are exactly balanced between the two digit classes. The primary BatchNorm CNN is trained from scratch with cross-entropy by full-batch GD at $\eta = 6.25 \times 10^{-4}$ for 1.6×10^6 steps, with no momentum or weight decay. Every run is evaluated on the same 5010 aligned and 4990 conflicting held-out examples. Reported end-of-training values average the final five checkpoints and uncertainty is computed over three seeds.

Intervention	Observable	DMFT median	Simulation median [IQR]
Increase agreement	c_{sc}	$+2.00$	$+2.00\ [2.00, 2.00]$
Increase agreement	c_{core}	-1.38	$-0.43\ [-1.44, 0.29]$
Increase α	Q_{out}	-0.99	$-0.76\ [-0.80, -0.72]$
Full $\to$ split map	c_{core}	$+0.93$	$+0.69\ [-0.12, 1.57]$

Table 3: Bounded paired changes in random-feature regression on Waterbirds and Spawrious. Entries are medians of $2(x_{\mathrm{after}}-x_{\mathrm{before}})/(|x_{\mathrm{after}}|+|x_{\mathrm{before}}|)$ over eight condition pairs; a denominator below 10^{-12} gives zero. The simulation IQR describes variation across condition pairs, not uncertainty over training datasets. A change from negative to positive saturates at $+2$, irrespective of its magnitude. The shortcut entries at $+2.00$ indicate agreement in direction, not in the size of the coefficient change.

Architecture and initialization. Inputs are RGB images bilinearly resized to 64×64, with intensities in $[0, 1]$. Three 3×3 convolutional layers have channel counts $3 \to 32 \to 64 \to 128$, stride one, padding one, and biases. Each is followed by BatchNorm and ReLU. The first two blocks end with 2×2 max-pooling; the third ends with global average pooling, followed by an affine $128 \to 2$ classifier. Runs use seeds $r \in \{0, 1, 2\}$, set before model construction. There is no pretraining or custom initialization: convolutional and affine weights and biases use PyTorch's default uniform initialization $[-f^{-1/2}, f^{-1/2}]$, where f is the layer's fan-in; BatchNorm scales start at one and offsets at zero.

BatchNorm and evaluation. One update accumulates gradients of the summed cross-entropy divided by P over a fixed, non-shuffled sequence of loader batches. BatchNorm uses each loader batch's statistics during training, with numerical stabilizer 10^{-5} and running-statistics momentum 0.1. Evaluation switches the network to evaluation mode and uses the stored running statistics without recalibration, then restores training mode.

Evaluation checkpoints. Metrics are recorded on the sorted, deduplicated grid

$$\mathcal{T} = \{0, T\} \cup \{10j : j = 1, \dots, 20\} \cup \left\{ \mathrm{round}\left[10(T/10)^{j/119}\right] : j = 0, \dots, 119 \right\}. \tag{211}$$

At $T = 1.6 \times 10^6$, this gives 137 checkpoints. Consecutive-checkpoint criteria below refer to this nonuniform grid, not to consecutive updates.

The main experiment varies $P \in \{150, 300, 600, 1500, 3000, 6000\}$ at each $p_{\mathrm{train}} \in \{0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.99\}$, with three paired seeds. At fixed agreement, both aligned and conflicting counts change with P. The reference trajectory uses $P_{\mathrm{ref}} = 6000$; the relative departure is measured using the equally weighted aligned and conflicting risks. The departure time uses a persistent 10% threshold, and the fit pools $p_{\mathrm{train}} \geq 0.60$. The reference self-departure is zero and does not enter the fit. For each paired seed, define the condition-balanced risk $R_P(t) = [R_{\mathrm{aligned}}(t) + R_{\mathrm{conflicting}}(t)]/2$ and the relative departure

$$d_P(t) = \frac{|R_P(t) - R_{P_{\mathrm{ref}}}(t)|}{\max\{|R_{P_{\mathrm{ref}}}(t)|, 0.20\}}.$$

We smooth d_P with a centered three-checkpoint average, replicating the boundary values, and define t_{dep} as the first positive-time checkpoint starting four consecutive smoothed values at or above 0.10. The plotted t_{reuse} denotes this diagnostic, not the theoretical crossover. Non-crossing trajectories supply no measured departure time; the horizon is not a crossing. Fig. 14 shows the diagnostic at $p_{\mathrm{train}} = 0.80$.

Descriptive power-law fit. For each agreement $p = p_{\mathrm{train}}$ and seed r, departure times are measured against the reference with the same (p, r). We fit

$$\log t_{\mathrm{dep}}(P, p, r) = a_{p,r} + b \log(P/P_{\mathrm{ref}}) + \epsilon_{P,p,r} \tag{212}$$

by unweighted least squares, with a separate intercept $a_{p,r}$ for each of the 18 agreement–seed pairs, common slope $b = 0.429$, and regression residual $\epsilon_{P,p,r}$. The fit pools $p \in$

$\{0.60, 0.70, 0.80, 0.90, 0.95, 0.99\}$ and $P \in \{150, 300, 600, 1500, 3000\}$. Non-crossing runs are omitted rather than assigned the training horizon; a pair with fewer than two observed sizes contributes no slope information. The reference is excluded. The fit uses individual departures, not the plotted averages. In Fig. 6(d), points are geometric means across agreement–seed pairs at each P; bars are the empirical 2.5th–97.5th percentiles in log time, transformed back to steps, not confidence intervals. The fitted-slope and square-root lines each have a free amplitude, fitted in log space to these geometric means.

Finite-reference dependence. At fixed agreement, suppose an early correction were $R_P(t) - R_\infty(t) = Ct^2/P$, with online-learning risk R_∞ and constant $C \neq 0$ independent of P. Then $R_P(t) - R_{P_{\mathrm{ref}}}(t) = Ct^2(1/P - 1/P_{\mathrm{ref}})$. With approximately constant risk normalisation, and neglecting checkpoint, smoothing, and persistence effects, the threshold time would scale as

$$t_{\mathrm{dep}} \propto \sqrt{\frac{P}{1 - P/P_{\mathrm{ref}}}}, \qquad P < P_{\mathrm{ref}}. \tag{213}$$

The $\sqrt{P}$ scaling requires $P/P_{\mathrm{ref}} \ll 1$, whereas the fitted ratios span 0.025–0.5. Excluding the self-comparison does not remove this reference effect. This calculation is illustrative: a t^2/P risk expansion has not been established for the CNN.

The pooled exponent 0.43 in Fig. 6(d) is descriptive: agreement-specific scaling and sensitivity to the reference, threshold, denominator floor, smoothing, and persistence rule have not been established. Neither this fit nor the accuracy gap distinguishes reuse from changes in minority counts, BatchNorm statistics, or learned representations.

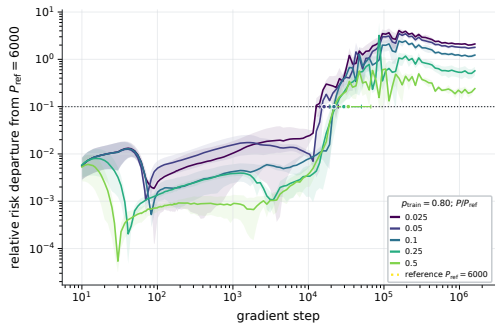

Fig. 14: Relative condition-balanced-risk departure on Colored MNIST at $p_{\mathrm{train}} = 0.80$. Curves are means over paired seeds and bands are standard errors. The dotted line is the 10% threshold; colored markers give the geometric-mean measured departure time t_{dep}, with horizontal bars spanning the 2.5–97.5 percentiles across seeds. These bars describe seed-to-seed spread, not uncertainty in the fitted exponent. Colors identify P/P_{ref} for $P_{\mathrm{ref}} = 6000$. The reference self-departure is zero by construction and is not drawn. *Parameters:* App. B.

Step-size check. At $p_{\mathrm{train}} = 0.80$, we compared $\eta \in \{5, 2.5, 1.25, 0.625, 0.3125\} \times 10^{-3}$ for $P \in \{150, 300, 600, 1500\}$ and three seeds at fixed physical time $\eta T = 250$. Smaller steps reduce the largest loss excursions in the checked trajectories, but the roughness does not decrease monotonically throughout the grid. These checks do not establish convergence over the full $\eta T = 1000$ horizon or all agreement probabilities of the main experiment.