How Rare Counterexamples Change What Models Learn from the Majority

Anonymous authors
Paper under double-blind review

Abstract

Models often rely on shortcuts that predict most training examples correctly but fail on rare counterexamples. We study why the same counterexamples can correct shortcut reliance at different rates depending on the remaining training data. In a two-feature model trained with logistic loss, we vary how robust and shortcut feature values are paired within majority examples while holding the counterexamples, feature marginals, and covariance fixed. We identify an indirect mechanism: counterexample updates change majority-example margins, reactivating gradients that both promote robust-feature learning and reinforce the shortcut. The balance between these effects evolves as different majority examples come to dominate learning. Our analysis links these shifting gradient contributions to feature growth and the time to first correct classification, showing that matched marginals and covariance do not determine correction times and that approximations based on the initially hardest examples can miss the subsequent dynamics. A controlled intervention that removes only majority-example updates to the robust feature reduces the ratio of mean correction times between two pairing conditions from 2.17 to 1.02, supporting the role of this indirect pathway. Experiments with small tanh networks on handwritten digits with synthetic backgrounds further show that feature pairing affects worst-group accuracy even when parameter-update norms are matched. Together, these results show how counterexamples reshape subsequent learning from majority examples, revealing a mechanism for shortcut correction that feature marginals and covariance alone do not capture.

Keywords: shortcut learning, learning theory, learning dynamics, spurious correlations, feature learning

1 Introduction

Models can predict most training samples correctly using cues that generalize poorly (Geirhos et al., 2020; 2019). In computer vision, recognizing categories from their usual backgrounds can fail when those backgrounds change (Xiao et al., 2021). Object features may remain underused. Correctly labeled counterexamples break this association. Related approaches use counterfactual examples (Kaushik et al., 2020; Teney et al., 2020) or emphasize poorly served groups and samples (Sagawa et al., 2020a; Nam et al., 2020; Liu et al., 2021). Yet training continues on majority samples containing both object and background information. Will these updates promote object learning or restore background reliance?

Studies of learning bias examine feature simplicity and availability (Rahaman et al., 2019; Shah et al., 2020; Hermann et al., 2024), as well as loss and optimization dynamics (Pezeshki et al., 2021; Puli et al., 2023). Memorizing rare exceptions can coexist with poor generalization (Sagawa et al., 2020b; Bayat et al., 2025), while useful features can remain available in shortcut-reliant classifiers (Izmailov et al., 2022; Kirichenko et al., 2023). In a two-feature model with logistic loss, we hold the counterexamples, feature marginals, and covariance fixed and vary only *feature pairing*: which robust and shortcut strengths occur together in majority samples. Counterexample updates can lower some majority margins and reactivate gradients through changes in model state. These gradients strengthen both features, with the balance shifting as different majority samples come to dominate learning.

Our analysis connects these changing contributions to feature growth and first correction times. Matched marginals and covariance leave these times undetermined, and approximations based on the

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

initially hardest samples can miss later dynamics. Correction order can also reverse for queries with weaker robust signals. We derive counterexample count bounds for reaching targets within a budget and meeting them at the training endpoint, distinguishing slow learning from unreachable targets. Broader counterexample coverage can improve endpoint predictions at a fixed count.

In a controlled stochastic-gradient experiment, feature pairing produces a 2.17-fold difference in mean first correction time. Removing only majority robust updates reduces this ratio to 1.02, supporting the indirect path. Small tanh networks on handwritten digits with synthetic backgrounds retain pairing-dependent worst-group accuracy even with matched per-step parameter-update norms.

Shortcut-to-Knowledge (SK) estimates fixed-budget learning progress from fixed representations before training the diagnostic readout. Its *SK-value* summarizes predicted relative loss reduction under the least favorable candidate context. Across 20 datasets from five domains and 142 label tasks, predicted and trained scores achieve a dataset-level Spearman rank correlation of 0.991.

Our main contributions are:

- We identify and quantify an indirect mechanism through which counterexamples change subsequent majority-sample gradients and robust-feature learning.
- We show that matched feature marginals and covariance do not determine first correction times, and derive counterexample count bounds for specified targets and training budgets.
- We test the mechanism through controlled interventions and neural-network experiments, and introduce and validate the SK assessment of learning progress across datasets.

2 Problem Setup

We use a two-feature binary classification model to isolate the effect of feature pairing in majority data. Controlled models help separate sources of out-of-distribution failure (Nagarajan et al., 2021), while linear networks permit detailed analysis of feature-learning dynamics (Saxe et al., 2014; Gidel et al., 2019; Xu & Ziyin, 2025).

Data distribution. Each sample contains a robust feature x_{r}, a shortcut feature x_{s}, and a label $y \in \{-1, 1\}$. The robust feature has the same sign as the label on every sample. The shortcut feature agrees with the label on majority samples and has the opposite sign on counterexamples, whose labels remain correct. The inputs take the form

$$\boldsymbol{x} = \begin{cases} y(a_{\mathrm{r}}, a_{\mathrm{s}}), & \text{majority sample,} \\ y(b_{\mathrm{r}}, -b_{\mathrm{s}}), & \text{counterexample,} \end{cases} \qquad a_{\mathrm{r}}, a_{\mathrm{s}}, b_{\mathrm{r}}, b_{\mathrm{s}} > 0. \tag{1}$$

The majority feature strengths $\boldsymbol{a} = (a_{\mathrm{r}}, a_{\mathrm{s}})$ follow a distribution P with finite support. The counterexample strengths b_{r} and b_{s} are fixed. A training sample is a counterexample with probability ϵ and a majority sample with probability $1 - \epsilon$, where $0 < \epsilon < 1/2$.

Model and training. The model predicts a weighted sum of the two features:

$$f_\theta(\boldsymbol{x}) = w_{\mathrm{r}} x_{\mathrm{r}} + w_{\mathrm{s}} x_{\mathrm{s}}, \qquad w_j = u_j v_j, \quad j \in \{\mathrm{r}, \mathrm{s}\}. \tag{2}$$

Here $\theta = (\boldsymbol{u}, \boldsymbol{v})$ contains the trainable factors. Training changes the effective weights w_{r} and w_{s} while the input feature strengths remain fixed. This factorization follows diagonal-network analyses of initialization and implicit bias (Woodworth et al., 2020; Moroshko et al., 2020; Pesme et al., 2021).

We use the logistic loss $\ell(m) = \log(1 + e^{-m})$, where $m = y f_\theta(\boldsymbol{x})$ is the classification margin. Positive margins indicate correct predictions, negative margins errors, and zero the classification boundary. The population loss is

$$\begin{aligned} L_P(\theta) = (1 - \epsilon)\, \mathbb{E}_{\boldsymbol{a} \sim P}\big[\ell(w_{\mathrm{r}} a_{\mathrm{r}} + w_{\mathrm{s}} a_{\mathrm{s}})\big] \\ + \epsilon\, \ell(w_{\mathrm{r}} b_{\mathrm{r}} - w_{\mathrm{s}} b_{\mathrm{s}}). \end{aligned} \tag{3}$$

We train on this mixture by parameter gradient flow, $\dot{\theta} = -\nabla_\theta L_P(\theta)$. Each channel starts from equal positive factors, $u_j(0) = v_j(0) > 0$. The initialization scales are specified in the corresponding results.

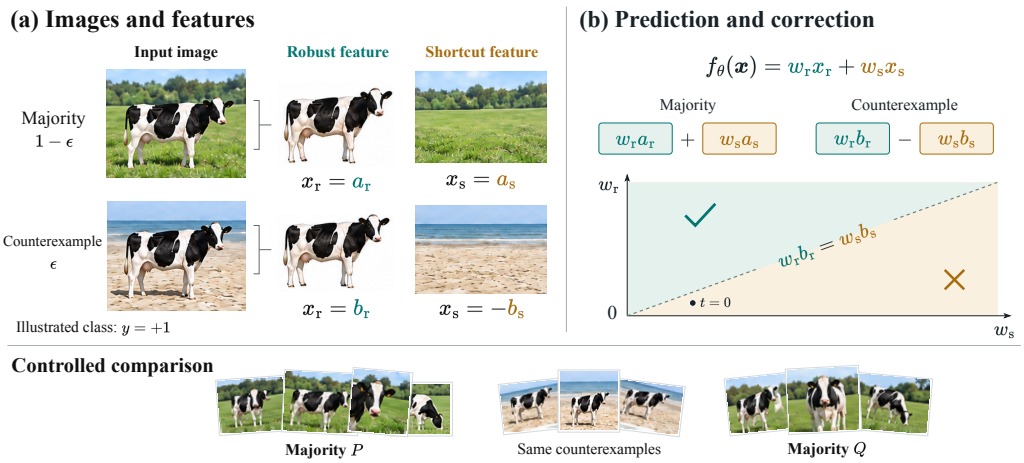


Figure 1: **Problem setup.** (a) Robust and shortcut features. (b) Feature contributions to predictions and the counterexample correction boundary; the check and cross indicate correct and incorrect counterexample predictions. Bottom: P and Q have matched marginals and covariance but different feature pairings. Both settings use the same counterexamples, initialization, and training procedure.

Controlled comparison. We compare majority distributions P and Q with identical marginals for a_r and a_s and the same covariance. Only *feature pairing* changes: which robust and shortcut strengths occur together within a sample. The counterexample strengths b_r, b_s, counterexample probability ϵ, model initialization, and training rule are the same in both settings.

First correction time. Under the positive initialization, all majority margins are positive. The counterexample margin at time t is $m_{ce}(t) = w_r(t)b_r - w_s(t)b_s$. We assume $m_{ce}(0) < 0$: the shortcut contribution initially outweighs the robust contribution, so the counterexamples are misclassified. For a given majority distribution P, we define the first correction time as

$$T_P = \inf\{t \geq 0 : m_{ce}(t) \geq 0\}, \tag{4}$$

the first time the counterexample margin reaches the classification boundary. If the boundary is never reached, we set $T_P = +\infty$.

Figure 1 summarizes the data, prediction rule, and controlled comparison.

3 How Counterexamples Change Majority Learning

Training-data attribution relates samples to predictions (Koh & Liang, 2017; Pruthi et al., 2020; Ilyas et al., 2022; Park et al., 2023). We study how a counterexample changes gradients from other samples. Here, *majority learning* denotes the majority-sample contribution to updates, including during mixed training.

3.1 State-Dependent Majority Gradients

Write $\boldsymbol{w} = (w_r, w_s)$ and recall the classification margin $m = yf_\theta(\boldsymbol{x})$. For logistic loss, the gradient weight is $q(m) = -\ell'(m) = (1 + e^m)^{-1}$. It is small for large positive margins and increases as the margin falls. The majority contributions to the two channels are

$$G_{j,P}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{a}\sim P}\left[a_j q(w_r a_r + w_s a_s)\right], \qquad j \in \{r, s\}. \tag{5}$$

Under the balanced factorization in Section 2, the mixed-training dynamics satisfy

$$\begin{aligned} \dot{w}_r &= 2w_r\left[(1-\epsilon)G_{r,P}(\boldsymbol{w}) + \epsilon b_r q(m_{ce})\right], \\ \dot{w}_s &= 2w_s\left[(1-\epsilon)G_{s,P}(\boldsymbol{w}) - \epsilon b_s q(m_{ce})\right]. \end{aligned} \tag{6}$$

The factors $2w_j$ arise from training u_j and v_j. Majority samples increase both weights, while counterexamples contribute to robust growth and shortcut reduction.

After a counterexample update, a majority sample's margin falls if the decrease in its shortcut contribution exceeds the increase in its robust contribution. The resulting increase in $q(m)$ strengthens both of its gradient contributions, although its features remain unchanged. We call this influence on subsequent majority updates, transmitted through the model state, a *counterexample-induced wave*; Figure 2(c) illustrates this path.

Training on the majority loss $\mathbb{E}_{\boldsymbol{a}\sim P}[\ell(\boldsymbol{w}\cdot\boldsymbol{a})]$ alone gives $\dot{w}_j = 2w_j G_{j,P}(\boldsymbol{w})$, so

$$\frac{\mathrm{d}\log w_{\mathrm{r}}}{\mathrm{d}\log w_{\mathrm{s}}} = h_P(\boldsymbol{w}) := \frac{G_{\mathrm{r},P}(\boldsymbol{w})}{G_{\mathrm{s},P}(\boldsymbol{w})}. \tag{7}$$

This ratio is an average of $a_{\mathrm{r}}/a_{\mathrm{s}}$, weighted by each sample's probability and $a_{\mathrm{s}}q(m)$. Feature pairing determines both the ratios and their changing weights, and thus the balance between robust and shortcut growth.

3.2 From a Local Update to Additional Learning

To measure the induced growth, we compare two continuations from the same model state $\boldsymbol{w}^0$. One branch receives a counterexample gradient step of size $0 < \eta < 1/b_{\mathrm{s}}$ on the balanced factors, yielding $w_{\mathrm{r}}^+ > w_{\mathrm{r}}^0$ and $0 < w_{\mathrm{s}}^+ < w_{\mathrm{s}}^0$. The other receives no such step. Both then follow the normalized majority flow, with trajectories $\boldsymbol{w}(t)$ from $\boldsymbol{w}^+$ and $\bar{\boldsymbol{w}}(t)$ from $\boldsymbol{w}^0$. These majority-only continuations isolate the induced gradient response.

Let $\gamma = w_{\mathrm{s}}^+/w_{\mathrm{s}}^0 \in (0,1)$ and let τ be the first time $w_{\mathrm{s}}(t)$ returns to w_{s}^0. Both branches are evaluated at this same time. This stopping event concerns shortcut recovery, whereas T_P concerns the counterexample classification boundary. Writing $J_{\mathrm{dir}} = \log(w_{\mathrm{r}}^+/w_{\mathrm{r}}^0)$ for the direct robust log gain, we obtain the indirect response

$$\begin{aligned} J_{\mathrm{ind}} &:= \log\frac{w_{\mathrm{r}}(\tau)}{\bar{w}_{\mathrm{r}}(\tau)} - J_{\mathrm{dir}} \\ &= 2\int_0^\tau \left[G_{\mathrm{r},P}(\boldsymbol{w}(t)) - G_{\mathrm{r},P}(\bar{\boldsymbol{w}}(t))\right]\mathrm{d}t. \end{aligned} \tag{8}$$

After removing control growth and J_{dir}, J_{ind} measures additional robust growth from majority training. Figure 2(b) illustrates a positive response at shortcut recovery.

For small w_{r}^+, we integrate $h_P(0,z)$ along shortcut recovery. The following bound controls the approximation error over the full interval.

Proposition 3.1 (Finite majority recovery)**.** *Under the positive finite-support assumptions of Section 2 and the update specified above, $\tau < \infty$ and*

$$\left|\log\frac{w_{\mathrm{r}}(\tau)}{w_{\mathrm{r}}^+} - \int_{\gamma w_{\mathrm{s}}^0}^{w_{\mathrm{s}}^0} h_P(0,z)\,\frac{\mathrm{d}z}{z}\right| \le C_P w_{\mathrm{r}}^+\left(\gamma^{-\kappa_P} - 1\right), \tag{9}$$

where $\kappa_P = \max_{\boldsymbol{a}\in\operatorname{supp} P} a_{\mathrm{r}}/a_{\mathrm{s}}$ and $C_P \ge 0$ depends only on P.

The integral retains the full majority distribution: samples with small initial gradients can gain influence during recovery. The bound controls recovery growth, while (8) isolates the counterexample's indirect effect. Derivations and proofs are in Appendix A.1.

4 Correction Times and Counterexample Requirements

Data construction requires knowing whether a correction target can be reached within a training budget and how many counterexamples it requires. We use the majority-response mechanism to derive bounds on correction time and counterexample counts under mixed training. This finite-time analysis complements work on limiting classifier directions (Soudry et al., 2018; Ji & Telgarsky, 2019; Lyu & Li, 2020).

4.1 Feature Pairing and Correction Times

With the counterexamples fixed, a budget that suffices under one majority pairing may be too short under another. Consider an initially weak robust weight $w_{\mathrm{r}}(0) = e^{-L}$ and a fixed $w_{\mathrm{s}}(0) = s_0 > 0$.

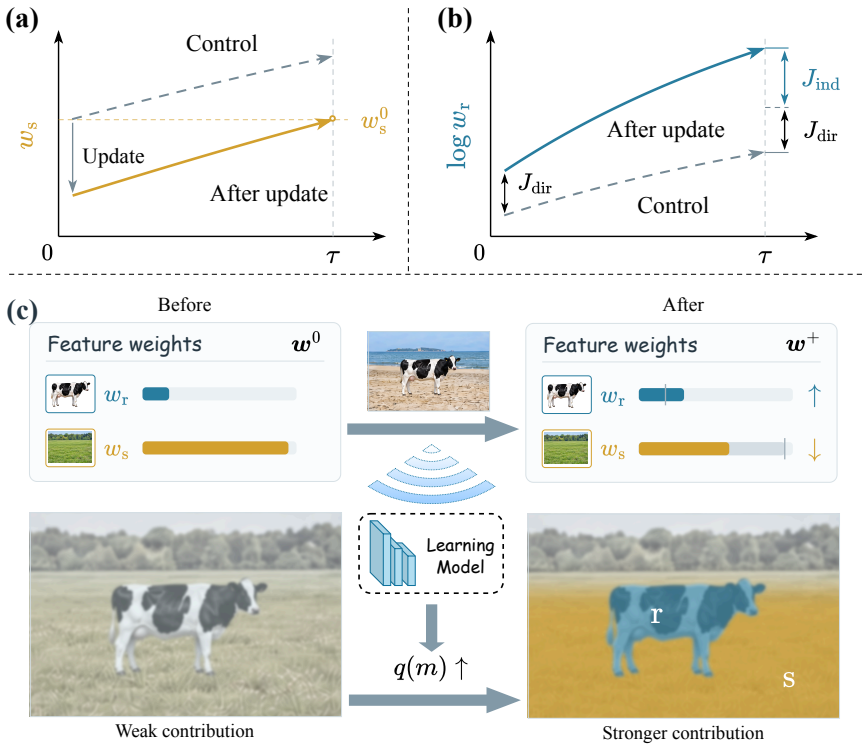


Figure 2: **Majority learning after a counterexample update.** Both branches start from $\boldsymbol{w}^0$; only the solid branch receives the counterexample update. Both then undergo identical majority-only training for the same duration. (a) Shortcut recovery to w_s^0 defines τ. (b) The robust log-weight gap comprises direct gain J_{dir} and indirect response J_{ind} (positive here). (c) Model-state changes alter the gradient weight $q(m)$ of an unchanged majority sample.

Theorem 4.1 (First correction time). *Assume $b_r \le a_r < a_s$ on the majority support, $b_r < b_s$, and $(1-\epsilon)\mathbb{E}_P[a_s] > \epsilon b_s$. Let $S_\epsilon > 0$ be the unique shortcut-axis balance satisfying $(1-\epsilon)G_{s,P}(0, S_\epsilon) = \epsilon b_s q(-b_s S_\epsilon)$. Then, as $L \to \infty$,*

$$T_P(L,\epsilon) = \frac{L}{\lambda_P(\epsilon)} + O(1), \qquad \lambda_P(\epsilon) = 2\epsilon q(-b_s S_\epsilon)\left[b_r + b_s h_P(0, S_\epsilon)\right]. \tag{10}$$

The two terms in brackets are the direct counterexample contribution and the majority contribution. Matched shortcut marginals give the same S_ϵ, while feature pairing changes h_P through $\mathbb{E}_P[a_r \mid a_s]$. Thus, matched marginals and covariance leave the leading correction time undetermined.

To follow learning beyond the training counterexamples, consider queries with signed features $(b_s/\chi, -b_s)$. They are corrected when w_r/w_s reaches χ; larger χ means a weaker robust signal. Figure 3(a,b) uses $N = 3000$ majority records, 200 counterexamples, $L = 8$, $s_0 = 2$, and $(b_r, b_s) = (0.05, 2)$. At the training target $\chi = 40$, Q corrects earlier than P (298 versus 324). At $\chi = 320$, Q takes 3.24 times as long. Panel (b) tracks the weakest robust point's gradient share, showing how the samples driving robust learning change during training.

4.2 Counterexample Counts and Reachability

For a specified target and training budget, we seek the minimum counterexample count. The bounds distinguish insufficient time for robust learning from insufficient counterexamples to reach a weak-signal target at all.

Give N majority records and k counterexamples equal weight, so $\epsilon_k = k/(N+k)$. Keep the initialization fixed across counts. The budget B is gradient-flow time for the loss averaged over all $N+k$ records. Write $T_{P,\chi}(k)$ for the first time $w_{\mathrm{r}}/w_{\mathrm{s}} \geq \chi$, and let $k_{\mathrm{hit},P}(B,\chi)$ be the smallest $1 \leq k < N$ with $T_{P,\chi}(k) \leq B$, or $+\infty$ if none exists.

Two stages constrain this count: growth of the weak robust weight and subsequent progress toward the query. Under the preceding support conditions, suppose the unique weakest majority point has $a_{\mathrm{r},*} = b_{\mathrm{r}}$, shortcut strength $a_{\mathrm{s},*}$, and probability p_*. Set $\alpha = k/N$ and $\Lambda_P(\alpha) = \lambda_P(\alpha/(1+\alpha))$. In the joint limit $B \to \infty$, $L/B \to \beta > 0$, and $\chi/\log B \to \rho > 0$, the count-ratio threshold is

$$\alpha_P^* = \max\left\{ \underbrace{\Lambda_P^{-1}(\beta)}_{\text{robust growth}}, \underbrace{\frac{p_* a_{\mathrm{s},*}}{b_{\mathrm{s}}} \exp\left[-\frac{a_{\mathrm{s},*}+b_{\mathrm{s}}}{b_{\mathrm{r}}\rho}\right]}_{\text{weak-signal coverage}} \right\}. \tag{11}$$

Here Λ_P is assumed strictly increasing on the relevant range, and α_P^* lies inside its positive-balance range with $0 < \alpha_P^* < 1$. Ratios separated below the threshold fail; those separated above it succeed for sufficiently large B. As the target weakens, the coverage constraint can overtake the growth constraint and reverse the preferred pairing (Figure 3(c)).

A separate obstruction arises when $a_{\mathrm{r},*} < b_{\mathrm{r}}$. Retain $a_{\mathrm{r}} < a_{\mathrm{s}}$ and assume that all other majority robust strengths exceed b_{r}. For an initially incorrect query with $\chi > (a_{\mathrm{s},*}+b_{\mathrm{s}})/(b_{\mathrm{r}}-a_{\mathrm{r},*})$, every budget satisfies

$$k_{\mathrm{hit},P}(B,\chi) \geq \left\lceil \frac{N p_* (a_{\mathrm{s},*} - a_{\mathrm{r},*})}{b_{\mathrm{r}} + b_{\mathrm{s}}} \right\rceil. \tag{12}$$

Below this count, the weak query remains unreachable at any budget, even if the training counterexamples are corrected.

4.3 Finite-Budget Guarantees and Signal Coverage

Coverage can deteriorate after first correction. We give necessary and sufficient count bounds for meeting an error target at the chosen training endpoint, and examine whether broader counterexample coverage improves predictions at a fixed count. On queries $(b_{\mathrm{r}}U, -b_{\mathrm{s}})$, with $U \sim \mathrm{Unif}(0,1]$, the error is

$$\mathcal{E}_P(t;k) = \min\left\{1, \frac{b_{\mathrm{s}} w_{\mathrm{s}}(t)}{b_{\mathrm{r}} w_{\mathrm{r}}(t)}\right\}. \tag{13}$$

Let $k_{\mathrm{end},P}(B,\delta)$ be the smallest admissible count with $\mathcal{E}_P(B;k) \leq \delta$, where $0 < \delta < 1$; use $+\infty$ if no count succeeds. The endpoint count can therefore exceed $k_{\mathrm{hit},P}(B, b_{\mathrm{s}}/(\delta b_{\mathrm{r}}))$.

We bound the required counts using state comparisons and two-sided arrival-time bounds computed from the data and initialization. For the matched construction above, decrease every majority robust strength by 0.005 and keep the remaining settings fixed. At $B = 10^4$ and $\delta = 0.01$, the bounds give

$$601 \leq k_{\mathrm{end},P} \leq 700, \qquad 1201 \leq k_{\mathrm{end},Q} \leq 1400. \tag{14}$$

Thus, 700 counterexamples suffice under P, while 1200 still fail under Q: the minimum requirements differ by a factor greater than 1.71. Figure 3(d) shows these guarantees across budgets. The intervals bracket minimum counts, rather than statistical uncertainty. The lower bounds at $B = 10^8$ exceed the upper bounds at $B = 10^4$: longer training can require more counterexamples at the endpoint.

To broaden signal coverage in the same shifted setting, keep $k = 600$ and replace 30 counterexamples with a second type with robust strength 0.025 instead of 0.05, retaining shortcut strength -2. Each run uses its fixed mixture throughout training. At $B = 10^4$, the weak-signal error falls from 3.31% to 0.00048% under P and from 7.90% to 0.70% under Q (Figure 3(e)). The harder counterexamples retain gradient weight after the original type becomes easy, sustaining pressure against the shortcut. Their effect still depends on majority pairing. Appendices A.2–A.4 give proofs, construction details, and numerical protocols.

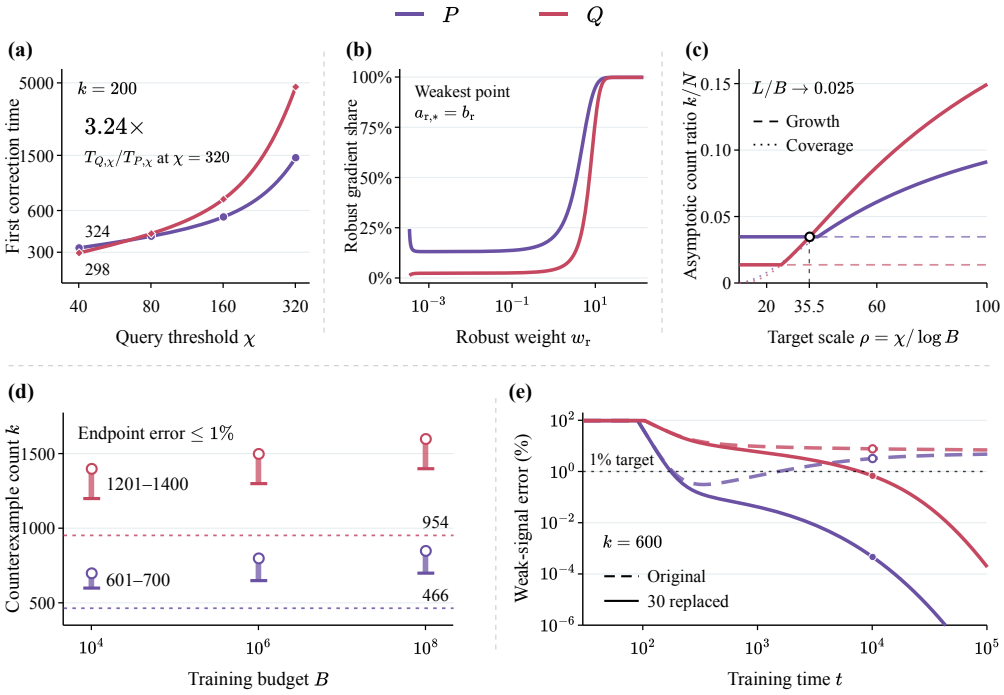


Figure 3: **Feature pairing and counterexample requirements.** (a,b) Correction times and the weakest point's share of the majority robust gradient, with 200 counterexamples. (c) Asymptotic growth and coverage thresholds. (d) Necessary (caps) and sufficient (circles) counts for 1% endpoint error; dotted lines mark first-entry bounds. (e) Replacing 30 of 600 counterexamples (solid) versus no replacement (dashed). Panels (d,e) shift majority robust strengths by -0.005. P/Q comparisons match marginals, covariance, initialization, and counterexamples.

5 Mechanism Tests and Dataset Assessment

5.1 Testing the Indirect Path

We train the factorized model of Section 2 with stochastic gradient descent under the matched feature pairings P and Q. At $\epsilon = 0.001$, 64 paired runs yield a ratio of mean first correction times of 2.17. Removing only majority-sample updates to the robust factors reduces this ratio to 1.02, while both pairings take longer to correct (Figure 4(a)). All counterexample updates and majority shortcut updates remain active. This supports majority robust learning as the source of the timing gap.

We examine the local response in neural networks with separate image and background branches, using Fashion-MNIST images (Xiao et al., 2017). At checkpoints from mixed-training trajectories, we copy the model and apply only the counterexample background update, leaving the image branch fixed. We then measure the majority contribution to growth of the image-channel weight, denoted G_M, on the same majority samples. This contribution increases at all 720 recorded states (Figure 4(b)). A change confined to the background branch can therefore strengthen the learning signal supplied by unchanged majority images.

We also train small tanh networks on the 8×8 handwritten-digit data distributed with scikit-learn (Alpaydin & Kaynak, 1998; Pedregosa et al., 2011), using synthetic backgrounds. The paired datasets retain the same digit identities, labels, counterexamples, feature marginals, and matched input moments. For each of three initializations, both pairings start from the same pretrained state and use the same per-step parameter-update norm. At step 500, their worst-group accuracy gaps range from 3.52 to 19.07 percentage points (Figure 4(c)). Protocols and further interventions are in Appendices B.1–B.3.

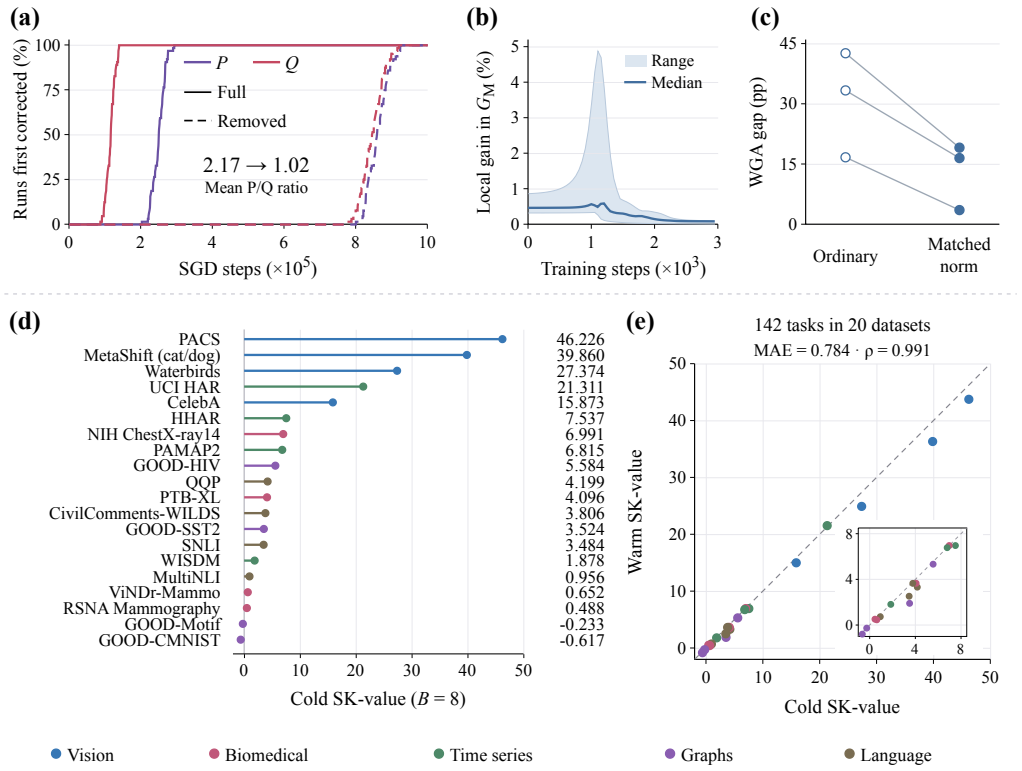


Figure 4: **Mechanism tests and dataset assessment.** (a) First correction under full updates (solid) and removal of majority robust updates (dashed), with 64 paired runs per condition. (b) Relative change in G_M after a counterexample background update with image parameters fixed; median and full range over 12 trajectories. (c) Worst-group accuracy gaps under ordinary and matched-norm updates; connected points share an initialization. (d,e) SK rankings and validation on 20 datasets and 142 tasks at $B = 8$. Cold predicts readout dynamics; warm trains the two coefficients. Both use fixed representations. Colors denote domains; the inset enlarges low scores. MAE and Spearman ρ weight datasets equally.

5.2 Assessing Learning Support Before Training

When a model already relies on a shortcut, how much can the available data improve its predictions within a limited budget? Existing tools assess data through valuation, training dynamics, or early gradients (Ghorbani & Zou, 2019; Toneva et al., 2019; Swayamdipta et al., 2020; Paul et al., 2021). Task embeddings and transferability scores assess tasks or pretrained representations (Achille et al., 2019; Nguyen et al., 2020; You et al., 2021; Huang et al., 2022). SK estimates fixed-budget progress under the least favorable candidate context before readout training.

The assessment uses fixed representations and a readout with two coefficients, one for task information and one for context. Visual representations include CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024); the encoders remain fixed. Starting from a context-only prediction, we use the initial source-risk gradient and Hessian to predict the readout trajectory under a local quadratic approximation, without parameter updates. Holding each query's task component fixed, let $\widehat{L}_i(t)$ be its largest predicted cross-entropy loss over candidate contexts. For dataset D and budget B, set

$$\text{SK-value}(D; B) = \frac{100}{B} \int_0^B \mathbb{E}_{\text{bal}} \left[1 - \frac{\widehat{L}_i(t)}{\widehat{L}_i(0)} \right] \mathrm{d}t, \tag{15}$$

where $\mathbb{E}_{\text{bal}}$ averages equally over query records within each class, classes within each task, and tasks within the dataset. A score of 10 means a predicted 10% relative loss reduction, averaged over the training budget; negative scores indicate deterioration.

5.3 Learning Support Across Datasets

We assess 20 datasets across vision, biomedical data, time series, graphs, and language, covering 142 label tasks. These include PACS (Li et al., 2017), MetaShift (Liang & Zou, 2022), and datasets from WILDS and GOOD (Koh et al., 2021; Gui et al., 2022). At the common budget $B = 8$, measured in normalized readout-gradient-flow time, SK-values range from -0.617 to 46.226 (Figure 4(d)).

We validate SK by training the two readout coefficients on every task and applying the same scoring rule. Representations, initializations, candidate contexts, and budgets match those used for prediction. The predicted and measured dataset scores achieve a Spearman rank correlation of 0.991, with a mean absolute error of 0.784 points (Figure 4(e)).

Appendix C gives the algorithm, protocols, dataset details, full results, and numerical checks.

6 Discussion and Conclusion

Counterexamples should be selected alongside the majority data they will accompany. Data-pruning theory also shows that the value of difficult samples depends on the amount of available data (Sorscher et al., 2022). In our setting, subsequent majority learning can sustain or undo the same local correction. A counterexample's current loss or violation of the shortcut is therefore an incomplete selection criterion. One promising criterion is whether it helps majority samples sustain gradients that support robust prediction. This response could be checked at representative training stages alongside performance on unseen counterexamples.

Counterexample coverage should reflect the intended deployment setting. If task signals are weaker there, correcting easier training counterexamples may leave target predictions unreliable. Adding familiar counterexamples and covering new difficulties may therefore yield different benefits. Selection should consider whether examples continue to supply corrective gradients while the model relies on the shortcut. Continued training can also let majority samples reinforce the shortcut and erode earlier gains. Data sufficiency thus depends on the prediction target and training endpoint. SK assesses this learning support under a specified representation, model, and budget.

For more complex tasks, we can examine how counterexample updates change other samples' gradients: residual-driven updates in regression, and correct-versus-competing class scores in multiclass classification. Deep networks also require separating prediction changes from representation changes, a distinction central to lazy-training analyses (Chizat et al., 2019). These responses should be tested under context shifts. The learning-theoretic question is when existing task information supplies a sustained learning signal, and what conditions allow that signal to support robust prediction.

AI Use Statement

We used generative AI tools throughout manuscript preparation to translate Chinese drafts into English draft and revise text, improve the organization and presentation of theoretical arguments and experimental findings, and maintain consistency in terminology, notation, and figure captions. The authors take full responsibility for the final manuscript, including its mathematical claims, experimental results, figures, code, and references.

Ethics Statement

This study investigates shortcut learning through theoretical analysis and controlled computational experiments. The dataset assessment uses existing public datasets and released representations, including biomedical datasets. It involves no new participant recruitment or clinical interventions. Results are reported at the task and dataset levels.

Reproducibility Statement

The model assumptions and training dynamics are specified in Section 2 and alongside the corresponding results. Appendix A provides proofs, matched constructions, and numerical details for majority learning, correction times, counterexample requirements, and coverage. Appendix B documents the stochastic-gradient tests, neural probes, channel interventions, and matched-norm experiments.

Appendix C specifies the SK algorithm and validation protocol, dataset and task definitions, complete results for all 142 tasks, and budget and numerical checks.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task Embedding for Meta-Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

E. Alpaydin and C. Kaynak. Optical Recognition of Handwritten Digits, 1998. URL https://doi.org/10.24432/c50p49.

Reza Bayat, Mohammad Pezeshki, Elvis Dohmatob, David Lopez-Paz, and Pascal Vincent. The Pitfalls of Memorization: When Memorization Hurts Generalization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d5ae1c5167be330b90f4c0a1eed7f8f0-Paper-Conference.pdf.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500. ACM, 2019. doi: 10.1145/3308560.3317593. URL http://dx.doi.org/10.1145/3308560.3317593.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1811.12231.

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 http://dx.doi.org/10.1038/s42256-020-00257-z.

Amirata Ghorbani and James Zou. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2242–2251. PMLR, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit Regularization of Discrete Gradient Dynamics in Linear Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f39ae9ff3a81f499230c4126e01f421b-Paper.pdf.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A Graph Out-of-Distribution Benchmark. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2059–2073. Curran Associates, Inc., 2022. doi: 10.52202/068431-0150. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0dc91de822b71c66a7f54fa121d8cbb9-Paper-Datasets_and_Benchmarks.pdf.

Katherine Hermann, Hossein Mobahi, Thomas Fel, and Michael Mozer. On the Foundations of Shortcut Learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/bf215fa7fe70a38c5e967e59c44a99d0-Paper-Conference.pdf.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly Easy Transferability Estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 9201–9225. PMLR, 2022. URL https://proceedings.mlr.press/v162/huang22d.html.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding Predictions with Data and Data with Predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 9525–9587. PMLR, 2022. URL https://proceedings.mlr.press/v162/ilyas22a.html.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On Feature Learning in the Presence of Spurious Correlations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38516–38532. Curran Associates, Inc., 2022. doi: 10.52202/068431-2791. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/fb64a552feda3d981dbe43527a80a07e-Paper-Conference.pdf.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.02032.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1909.12434.

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.

Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Weixin Liang and James Zou. MetaShift: A Dataset of Datasets for Evaluating Contextual Distribution Shifts and Training Conflicts. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=MTex8qKavoS.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 6781–6792. PMLR, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Kaifeng Lyu and Jian Li. Gradient Descent Maximizes the Margin of Homogeneous Neural Networks. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1906.05890.

Edward Moroshko, Blake E Woodworth, Suriya Gunasekar, Jason D Lee, Nati Srebro, and Daniel Soudry. Implicit Bias in Deep Linear Classification: Initialization Scale vs Training Accuracy. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22182–22193. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fc2022c89b61c76bbef978f1370660bf-Paper.pdf.

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://arxiv.org/abs/2010.15775.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: De-biasing Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20673–20684. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/eddc3427c5d77843c2253f1e799fe933-Paper.pdf.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A New Measure to Evaluate Transferability of Learned Representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7294–7305. PMLR, 2020. URL https://proceedings.mlr.press/v119/nguyen20b.html.

Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific data*, 10:277, 2023. doi: 10.1038/s41597-023-02100-7. URL https://doi.org/10.1038/s41597-023-02100-7.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing Model Behavior at Scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 27074–27113. PMLR, 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep Learning on a Data Diet: Finding Important Examples Early in Training. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20596–20607. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ac56f8fe9eea3e4a365f29f0f1957c55-Paper.pdf.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29218–29230. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f4661398cb1a3abd3ffe58600bf11322-Paper.pdf.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1256–1272. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0987b8b338d6c90bbedd8631bc499221-Paper.pdf.

Nicolas Pilia, Claudia Nagel, Gustavo Lenis, Silvia Becker, Olaf Dössel, and Axel Loewe. ECGdeli - An open source ECG delineation toolbox for MATLAB. *SoftwareX*, 13:100639, 2021. doi: 10.1016/j.softx.2020.100639. URL http://dx.doi.org/10.1016/j.softx.2020.100639.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf.

Aahlad Puli, Lily 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, pp. 71874–71910. Curran Associates, Inc., 2023. doi: 10.52202/075280-3146. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e35460304fdf6df523f068a59aaf8829-Paper-Conference.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the Spectral Bias of Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5301–5310. PMLR, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Attila Reiss and Didier Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. In *2012 16th International Symposium on Wearable Computers*, pp. 108–109. IEEE, 2012. doi: 10.1109/iswc.2012.13. URL http://dx.doi.org/10.1109/ISWC.2012.13.

Jorge Reyes-Ortiz and Davide Anguita. Human Activity Recognition Using Smartphones, 2013. URL https://doi.org/10.24432/c54s4k.

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://arxiv.org/abs/1911.08731.

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, pp. 8346–8356. PMLR, 2020b. URL https://proceedings.mlr.press/v119/sagawa20a.html.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The Pitfalls of Simplicity Bias in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9573–9585. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6cfe0e6127fa25df2a0ef2ae1067d915-Paper.pdf.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information*

Processing Systems, volume 35, pp. 19523–19536. Curran Associates, Inc., 2022. doi: 10.52202/068431-1419. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7b75da9b61eda40fa35453ee5d077df6-Paper-Conference.pdf.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The Implicit Bias of Gradient Descent on Separable Data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL http://jmlr.org/papers/v19/18-188.html.

Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, pp. 127–140. ACM, 2015. doi: 10.1145/2809695.2809718. URL http://dx.doi.org/10.1145/2809695.2809718.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746/.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision. In *Computer Vision – ECCV 2020*, pp. 580–599. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58607-2_34. URL http://dx.doi.org/10.1007/978-3-030-58607-2_34.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1812.05159.

Hari M. Trivedi, Maryam Vazirabad, Felipe C. Kitamura, Yan Chen, and Helen Frazer. Open-Source Dataset for the RSNA Screening Mammography Cancer Detection Challenge. *Radiology: Artificial Intelligence*, 8(2), 2026. doi: 10.1148/ryai.250375. URL http://dx.doi.org/10.1148/ryai.250375.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), 2020. doi: 10.1038/s41597-020-0495-6. URL http://dx.doi.org/10.1038/s41597-020-0495-6.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1804.07461.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1113–1120. ACM, 2009. doi: 10.1145/1553374.1553516. URL http://dx.doi.org/10.1145/1553374.1553516.

Gary Weiss. WISDM Smartphone and Smartwatch Activity and Biometrics Dataset, 2019. URL https://doi.org/10.24432/c5hk59.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and Rich Regimes in Overparametrized Models. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 3635–3673. PMLR, 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or Signal: The Role of Image Backgrounds in Object Recognition. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.09994.

Yizhou Xu and Liu Ziyin. Three Mechanisms of Feature Learning in a Linear Network. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9941833e8327910ef25daeb9005e4748-Paper-Conference.pdf.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical Assessment of Pre-trained Models for Transfer Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12133–12143. PMLR, 2021. URL https://proceedings.mlr.press/v139/you21b.html.

A Theoretical Analysis and Controlled Numerical Results

A.1 Majority Learning after Counterexample Updates

This subsection gives the derivations for Section 3. Let $\{\boldsymbol{a}_i\}_{i=1}^n$ be the support of P, with $\boldsymbol{a}_i = (a_{\mathrm{r},i}, a_{\mathrm{s},i})$, probabilities $p_i > 0$, and $\sum_i p_i = 1$. All feature strengths are positive. We retain the factorization $w_j = u_j v_j$ and the balanced positive initialization of Section 2.

A.1.1 Factor Dynamics and Finite Updates

For logistic loss, define

$$q(m) = -\ell'(m) = \frac{1}{1+e^m}.$$

Writing $m_i = w_{\mathrm{r}} a_{\mathrm{r},i} + w_{\mathrm{s}} a_{\mathrm{s},i}$ and $m_{\mathrm{ce}} = b_{\mathrm{r}} w_{\mathrm{r}} - b_{\mathrm{s}} w_{\mathrm{s}}$, the majority contributions are

$$G_{j,P}(\boldsymbol{w}) = \sum_{i=1}^n p_i a_{j,i} q(m_i), \qquad j \in \{\mathrm{r}, \mathrm{s}\}. \tag{A.1}$$

The derivatives of the mixed population loss are therefore

$$\begin{aligned} \frac{\partial L_P}{\partial w_{\mathrm{r}}} &= -(1-\epsilon) G_{\mathrm{r},P}(\boldsymbol{w}) - \epsilon b_{\mathrm{r}} q(m_{\mathrm{ce}}), \\ \frac{\partial L_P}{\partial w_{\mathrm{s}}} &= -(1-\epsilon) G_{\mathrm{s},P}(\boldsymbol{w}) + \epsilon b_{\mathrm{s}} q(m_{\mathrm{ce}}). \end{aligned} \tag{A.2}$$

Training the factors gives

$$\dot{u}_j = -v_j \frac{\partial L_P}{\partial w_j}, \qquad \dot{v}_j = -u_j \frac{\partial L_P}{\partial w_j}.$$

Consequently, $\frac{\mathrm{d}}{\mathrm{d}t}(u_j^2 - v_j^2) = 0$. Starting from $u_j = v_j > 0$, the factors remain equal and positive: on this invariant set,

$$u_j(t) = u_j(0) \exp\left[-\int_0^t \frac{\partial L_P}{\partial w_j}(\boldsymbol{w}(s))\, \mathrm{d}s \right].$$

The loss derivatives are bounded because the support is finite and $0 < q(m) < 1$. Thus the factors remain finite at every finite time. Differentiating their product yields

$$\dot{w}_j = -(u_j^2 + v_j^2) \frac{\partial L_P}{\partial w_j} = -2 w_j \frac{\partial L_P}{\partial w_j}. \tag{A.3}$$

Substituting (A.2) gives the mixed-training dynamics in the main text. The factor $2w_j$ follows from the parameterization; it would be absent if w_j itself were trained by ordinary gradient flow.

For the controlled majority continuation, the loss is

$$L_{\mathrm{maj},P}(\boldsymbol{w}) = \sum_{i=1}^n p_i \ell(m_i), \qquad \dot{w}_j = 2 w_j G_{j,P}(\boldsymbol{w}). \tag{A.4}$$

This loss is normalized within the majority distribution, so the mixture factor $1-\epsilon$ is absent. Since $G_{\mathrm{s},P} > 0$,

$$\frac{\mathrm{d} \log w_{\mathrm{r}}}{\mathrm{d} \log w_{\mathrm{s}}} = h_P(\boldsymbol{w}) := \frac{G_{\mathrm{r},P}(\boldsymbol{w})}{G_{\mathrm{s},P}(\boldsymbol{w})} = \sum_i \pi_i(\boldsymbol{w}) \kappa_i, \tag{A.5}$$

where

$$\kappa_i = \frac{a_{\mathrm{r},i}}{a_{\mathrm{s},i}}, \qquad \pi_i(\boldsymbol{w}) = \frac{p_i a_{\mathrm{s},i} q(m_i)}{G_{\mathrm{s},P}(\boldsymbol{w})}. \tag{A.6}$$

The weights π_i sum to one. They describe shares of the majority shortcut gradient and vary with the model state, even though the sample probabilities p_i remain fixed.

A finite counterexample update. Let $\boldsymbol{w}^0$ be the effective weights before the intervention, and set $q_0 = q(b_{\mathrm{r}} w_{\mathrm{r}}^0 - b_{\mathrm{s}} w_{\mathrm{s}}^0)$. Apply one gradient step of size η to the factors using the counterexample loss alone. The exact updates are

$$\begin{aligned} u_{\mathrm{r}}^+ &= u_{\mathrm{r}}^0 + \eta b_{\mathrm{r}} q_0 v_{\mathrm{r}}^0, \quad v_{\mathrm{r}}^+ = v_{\mathrm{r}}^0 + \eta b_{\mathrm{r}} q_0 u_{\mathrm{r}}^0, \\ u_{\mathrm{s}}^+ &= u_{\mathrm{s}}^0 - \eta b_{\mathrm{s}} q_0 v_{\mathrm{s}}^0, \quad v_{\mathrm{s}}^+ = v_{\mathrm{s}}^0 - \eta b_{\mathrm{s}} q_0 u_{\mathrm{s}}^0. \end{aligned} \tag{A.7}$$

For $0 < \eta < 1/b_{\mathrm{s}}$, balance and positivity are preserved. Multiplying the updated factors gives

$$\begin{aligned} w_{\mathrm{r}}^+ &= w_{\mathrm{r}}^0 (1 + \eta b_{\mathrm{r}} q_0)^2, \\ w_{\mathrm{s}}^+ &= \gamma w_{\mathrm{s}}^0, \qquad \gamma = (1 - \eta b_{\mathrm{s}} q_0)^2 \in (0, 1). \end{aligned} \tag{A.8}$$

The direct robust log gain is

$$J_{\mathrm{dir}} = \log \frac{w_{\mathrm{r}}^+}{w_{\mathrm{r}}^0} = 2 \log(1 + \eta b_{\mathrm{r}} q_0). \tag{A.9}$$

These expressions retain the full finite step; no expansion in η is used.

For majority sample i, the margin changes by

$$m_i^+ - m_i^0 = a_{\mathrm{r},i}(w_{\mathrm{r}}^+ - w_{\mathrm{r}}^0) - a_{\mathrm{s},i}(w_{\mathrm{s}}^0 - w_{\mathrm{s}}^+).$$

Because q is strictly decreasing,

$$q(m_i^+) > q(m_i^0) \quad \Longleftrightarrow \quad a_{\mathrm{s},i}(w_{\mathrm{s}}^0 - w_{\mathrm{s}}^+) > a_{\mathrm{r},i}(w_{\mathrm{r}}^+ - w_{\mathrm{r}}^0). \tag{A.10}$$

When this condition holds, both of that sample's contributions $p_i a_{j,i} q(m_i)$ increase. Different feature pairings can satisfy the condition differently; the update need not increase every majority sample's gradient weight.

A.1.2 Recovery and the Same-Time Response

Starting from $\boldsymbol{w}(0) = \boldsymbol{w}^+$, follow the majority flow (A.4) and define

$$\tau = \inf\{t \geq 0 : w_{\mathrm{s}}(t) = w_{\mathrm{s}}^0\}.$$

This clock starts after the counterexample step. Its stopping event is shortcut-weight recovery, distinct from the first correction time T_P under mixed training. Finiteness of τ is established in the proof below.

Let $\bar{\boldsymbol{w}}(t)$ follow the same majority flow, starting from $\bar{\boldsymbol{w}}(0) = \boldsymbol{w}^0$ without the counterexample step. Evaluate both branches at the intervention branch's recovery time τ. The control is run for that same duration; it has no separate recovery stopping rule.

Integrating the robust log-growth rate on each branch gives

$$\begin{aligned} \log \frac{w_{\mathrm{r}}(\tau)}{w_{\mathrm{r}}^0} &= J_{\mathrm{dir}} + 2 \int_0^\tau G_{\mathrm{r},P}(\boldsymbol{w}(t))\, \mathrm{d}t, \\ \log \frac{\bar{w}_{\mathrm{r}}(\tau)}{w_{\mathrm{r}}^0} &= 2 \int_0^\tau G_{\mathrm{r},P}(\bar{\boldsymbol{w}}(t))\, \mathrm{d}t. \end{aligned}$$

Subtracting the second identity and the direct gain from the first proves

$$\begin{aligned} J_{\mathrm{ind}} &:= \log \frac{w_{\mathrm{r}}(\tau)}{\bar{w}_{\mathrm{r}}(\tau)} - J_{\mathrm{dir}} \\ &= 2 \int_0^\tau [G_{\mathrm{r},P}(\boldsymbol{w}(t)) - G_{\mathrm{r},P}(\bar{\boldsymbol{w}}(t))]\, \mathrm{d}t. \end{aligned} \tag{A.11}$$

Thus J_{ind} measures the difference in robust log growth supplied by majority training after the two initial states diverge. A positive value indicates additional robust growth on the intervention branch. The finite-recovery bound alone does not establish this sign: its recovery integral includes growth that would also occur without the counterexample step.

This controlled comparison isolates the response transmitted through the model state. It does not assume that counterexamples are introduced partway through the mixed training process, or that majority-only recovery occurs between successive counterexamples.

Finite recovery under majority training. Set

$$\kappa_{\min} = \min_i \kappa_i, \qquad \kappa_P = \max_i \kappa_i, \qquad a_{\mathrm{r},\max} = \max_i a_{\mathrm{r},i}.$$

We prove Proposition 3.1. Using $J_{\mathrm{dir}} = \log(w_{\mathrm{r}}^+/w_{\mathrm{r}}^0)$, its bound can equivalently be written as

$$\left| \log \frac{w_{\mathrm{r}}(\tau)}{w_{\mathrm{r}}^0} - J_{\mathrm{dir}} - \int_{\gamma w_{\mathrm{s}}^0}^{w_{\mathrm{s}}^0} h_P(0,z)\,\frac{\mathrm{d}z}{z} \right| \le C_P w_{\mathrm{r}}^+ \left(\gamma^{-\kappa_P} - 1\right), \tag{A.12}$$

where

$$C_P = \frac{(\kappa_P - \kappa_{\min}) a_{\mathrm{r},\max}}{4\kappa_P}. \tag{A.13}$$

Proof. Use the increasing coordinate $\xi = \log(w_{\mathrm{s}}/w_{\mathrm{s}}^+)$, and let $X = \log(1/\gamma)$. By (A.5),

$$\frac{\mathrm{d}\log w_{\mathrm{r}}}{\mathrm{d}\xi} = h_P(\boldsymbol{w}) \in [\kappa_{\min}, \kappa_P].$$

Before recovery, $0 \le \xi \le X$, so

$$w_{\mathrm{r}}^+ \le w_{\mathrm{r}} \le w_{\mathrm{r}}^+ e^{\kappa_P \xi} \le w_{\mathrm{r}}^+ \gamma^{-\kappa_P}, \qquad w_{\mathrm{s}}^+ \le w_{\mathrm{s}} \le w_{\mathrm{s}}^0.$$

On this compact rectangle, $G_{\mathrm{s},P}$ has a strictly positive lower bound. Hence $\dot{\xi} = 2G_{\mathrm{s},P}$ is bounded away from zero until recovery. It follows that ξ reaches X in finite time.

We next bound the error from evaluating h_P at zero robust weight. Differentiating the weighted average in (A.5) gives

$$\frac{\partial h_P}{\partial w_{\mathrm{r}}} = -\operatorname{Cov}_{\pi(\boldsymbol{w})}\left(\kappa_i,\, a_{\mathrm{r},i}[1 - q(m_i)]\right). \tag{A.14}$$

The two arguments lie in intervals of lengths $\kappa_P - \kappa_{\min}$ and at most $a_{\mathrm{r},\max}$. For any bounded random variable Z with range length d, $\operatorname{Var}(Z) \le d^2/4$. Applying this bound and Cauchy–Schwarz to the covariance yields

$$\left|\frac{\partial h_P}{\partial w_{\mathrm{r}}}\right| \le \frac{(\kappa_P - \kappa_{\min}) a_{\mathrm{r},\max}}{4} = C_P \kappa_P.$$

This estimate also holds between zero and the actual robust weight. Therefore,

$$|h_P(w_{\mathrm{r}}, w_{\mathrm{s}}) - h_P(0, w_{\mathrm{s}})| \le C_P \kappa_P w_{\mathrm{r}}.$$

Integrating along the recovery trajectory gives

$$\begin{aligned} &\left| \log \frac{w_{\mathrm{r}}(\tau)}{w_{\mathrm{r}}^+} - \int_{w_{\mathrm{s}}^+}^{w_{\mathrm{s}}^0} h_P(0,z)\,\frac{\mathrm{d}z}{z} \right| \\ &\qquad \le C_P \kappa_P \int_0^X w_{\mathrm{r}}^+ e^{\kappa_P \xi}\,\mathrm{d}\xi = C_P w_{\mathrm{r}}^+ \left(\gamma^{-\kappa_P} - 1\right). \end{aligned}$$

Finally, $\log(w_{\mathrm{r}}(\tau)/w_{\mathrm{r}}^0) = J_{\mathrm{dir}} + \log(w_{\mathrm{r}}(\tau)/w_{\mathrm{r}}^+)$, which proves (A.12). □

The integral retains all majority samples and their changing gradient weights throughout recovery. Evaluating it at $w_{\mathrm{r}} = 0$ is an approximation controlled by the bound; the actual robust weight remains positive.

Numerical examples of the majority response. We illustrate the controlled comparison with

$$P = \mathrm{Unif}\{(0.05, 1), (0.95, 2), (0.5, 3)\}.$$

Table A.1 distinguishes the parameters used to construct the curves in Figure 2 from the numerical example below. Each case applies one exact counterexample step to the balanced factors, followed by normalized majority training on both branches. The stopping time is the intervention branch's shortcut recovery time τ.

In the numerical example, the majority point $(0.05, 1)$ has the smallest margin before the update and supplies 96.34% of the majority shortcut contribution. Immediately after the update, its share falls

Table A.1: Settings and responses for two controlled majority continuations. All gains are natural-log weight differences.

Quantity	Figure 2 construction	Numerical example
$(w_{\mathrm{r}}^0, w_{\mathrm{s}}^0)$	$(0.04, 1)$	$(0.01, 4)$
$(b_{\mathrm{r}}, b_{\mathrm{s}})$	$(0.1, 1)$	$(0.05, 2)$
η	0.3	0.25
$(w_{\mathrm{r}}^+, w_{\mathrm{s}}^+)$	$(0.041772, 0.609834)$	$(0.010251, 1.000671)$
τ	0.842556	24.170097
J_{dir}	0.043343	0.024837
J_{ind}	0.057253	0.193723
$\log[w_{\mathrm{r}}(\tau)/\bar{w}_{\mathrm{r}}(\tau)]$	0.100596	0.218560

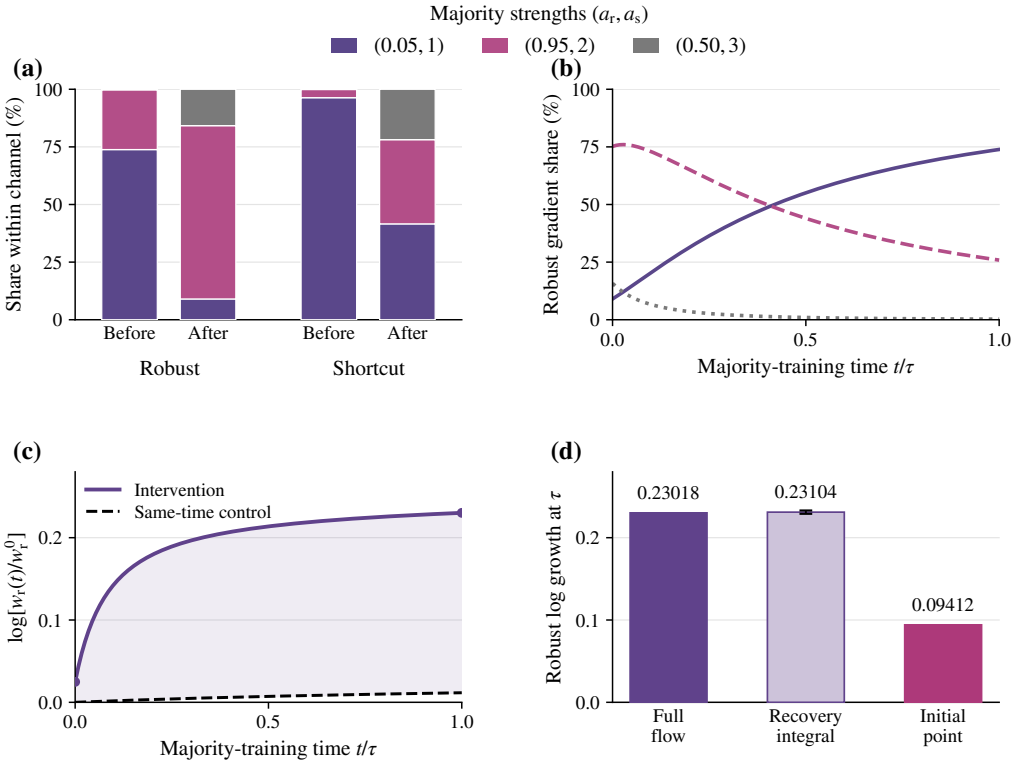


Figure A.1: **A finite update changes subsequent majority learning.** (a) Within-channel gradient shares before and after the update. (b) Robust gradient shares during majority continuation. (c) Robust log growth in the intervention and same-time control. (d) Full-flow growth at τ, the recovery-integral prediction, and the initial-point approximation. The error bar is the deterministic recovery bound. Colors in (a,b) identify samples; all panels use the numerical example in Table A.1.

to 41.59%. Both aggregate channel contributions increase: $G_{\mathrm{r},P}$ grows by a factor of 122.49, and $G_{\mathrm{s},P}$ by a factor of 34.63. Thus a change in relative share can accompany an increase in absolute contribution. Panel (b) follows all three robust shares through the subsequent recovery interval.

The intervention branch gains $\log[w_{\mathrm{r}}(\tau)/w_{\mathrm{r}}^{0}] = 0.230182$. The same-time control gains 0.011622. Subtracting the latter gives a response of 0.218560, which separates into $J_{\mathrm{dir}} = 0.024837$ and $J_{\mathrm{ind}} = 0.193723$. These quantities distinguish the robust growth observed after the update from the additional growth attributable to the controlled intervention.

To evaluate an approximation based on the initially hardest point, replace h_P throughout recovery by $a_{\mathrm{r},1}/a_{\mathrm{s},1} = 0.05$. Its predicted total robust log growth is

$$J_{\mathrm{dir}} + 0.05 \log(1/\gamma) = 0.094118. \tag{A.15}$$

The full recovery-integral prediction is

$$J_{\mathrm{dir}} + \int_{\gamma w_{\mathrm{s}}^{0}}^{w_{\mathrm{s}}^{0}} h_P(0, z)\, \frac{\mathrm{d}z}{z} = 0.231038. \tag{A.16}$$

Relative to the full-flow value 0.230182, these predictions have errors of 59.11% and 0.37%, respectively. The latter has absolute error 0.000857, below the bound 0.002029 in (A.12). The initial-point approximation misses contributions that become substantial after the counterexample changes the margins.

For numerical evaluation, we integrate in $\xi = \log(w_{\mathrm{s}}/w_{\mathrm{s}}^{+})$ from 0 to $\log(1/\gamma)$, so recovery occurs at the fixed endpoint. The control evolves under the same physical clock, obtained from $\mathrm{d}t/\mathrm{d}\xi = 1/(2G_{\mathrm{s},P})$ along the intervention branch. Fourth-order Runge–Kutta integration with 4000 steps gives changes below 10^{-8} in τ and the reported gains when compared with 2000 steps. The integral in (A.11) is also accumulated independently and agrees with the difference of the endpoint log weights to within 10^{-11}.

Correction without direct robust updates from counterexamples. The complementary channel experiment has a continuous-model counterpart. Multiply only the counterexample gradients to the robust factors by $\beta \in [0, 1]$, retaining every other update. Balanced factors then obey

$$\begin{aligned} \dot{w}_{\mathrm{r}} &= 2w_{\mathrm{r}}[(1-\epsilon)G_{\mathrm{r},P} + \beta\epsilon b_{\mathrm{r}} q(m_{\mathrm{ce}})], \\ \dot{w}_{\mathrm{s}} &= 2w_{\mathrm{s}}[(1-\epsilon)G_{\mathrm{s},P} - \epsilon b_{\mathrm{s}} q(m_{\mathrm{ce}})]. \end{aligned} \tag{A.17}$$

This is a specified intervention on the vector field; for $\beta \neq 1$ we do not assume it is gradient flow of a common scalar loss. Assume $0 < a_{\mathrm{r},i} < a_{\mathrm{s},i}$ for every majority point, $0 < b_{\mathrm{r}} < b_{\mathrm{s}}$, and $0 < \epsilon < 1$. From any positive initialization with negative counterexample margin, the flow first corrects that counterexample in finite time and subsequently retains the correct sign, including at $\beta = 0$.

To prove finite arrival, note that before correction $q(m_{\mathrm{ce}}) \geq 1/2$. For sufficiently large w_{s}, the majority shortcut gradient is below $\epsilon b_{\mathrm{s}}/4$, so w_{s} is bounded above by a constant S_b on the incorrect part of the trajectory. There, $w_{\mathrm{r}} < (b_{\mathrm{s}}/b_{\mathrm{r}})S_b$ as well. On this compact rectangle, the majority robust gradient has a strictly positive lower bound. Hence w_{r} grows at least exponentially until arrival, which rules out infinite waiting.

For transversality and persistence, write $\chi = b_{\mathrm{s}}/b_{\mathrm{r}}$ and parameterize the boundary by $(w_{\mathrm{r}}, w_{\mathrm{s}}) = (\chi s, s)$. The difference of the two log-growth brackets there is

$$B_\beta(s) = -(1-\epsilon) \sum_i p_i (a_{\mathrm{s},i} - a_{\mathrm{r},i}) q((\chi a_{\mathrm{r},i} + a_{\mathrm{s},i})s) + \frac{\epsilon(b_{\mathrm{s}} + \beta b_{\mathrm{r}})}{2}.$$

Its derivative is strictly positive. At a tangential first arrival, the vector field moves along the boundary toward increasing s, so the second derivative of the margin is positive. That would make the margin positive just before arrival, a contradiction. The crossing is therefore strict. Since w_{r} continues to increase, any later encounter with the boundary has larger s and larger $B_\beta(s)$, and cannot cross back into the incorrect region. This persistence concerns the training counterexample. It does not contradict the weak-query return analyzed below.

A.2 Feature Pairing and First Correction Times

We first prove the first correction time result in Section 4.1, then verify the matched construction and the finite-stage comparisons in Figure 3(a,b). Throughout this subsection, training uses the full mixture loss of Section 2. The definitions of q, $G_{j,P}$, and h_P are those in Appendix A.1.

A.2.1 A Uniform First-Correction Estimate

Fix $w_{\mathrm{r}}(0) = e^{-L}$ and $w_{\mathrm{s}}(0) = s_0 > 0$. For a fixed query with input $\boldsymbol{x} = y(b_{\mathrm{s}}/\chi, -b_{\mathrm{s}})$, define

$$T_{P,\chi}(L,\epsilon) = \inf\left\{t \geq 0 : \frac{w_{\mathrm{r}}(t)}{w_{\mathrm{s}}(t)} \geq \chi\right\}. \tag{A.18}$$

Taking $\chi = b_{\mathrm{s}}/b_{\mathrm{r}}$ recovers the training counterexample and its first correction time T_P. The following statement also covers any fixed, weaker robust signal.

Proposition A.1 (Uniform first correction time)**.** *Let P have finite positive support satisfying $b_{\mathrm{r}} \leq a_{\mathrm{r}} < a_{\mathrm{s}}$, and suppose $0 < b_{\mathrm{r}} < b_{\mathrm{s}}$. Define*

$$\epsilon_{\mathrm{ax}} = \frac{\mathbb{E}_P[a_{\mathrm{s}}]}{\mathbb{E}_P[a_{\mathrm{s}}] + b_{\mathrm{s}}}.$$

Fix a compact interval $I = [\epsilon_-, \epsilon_+] \subset (0, \min\{1/2, \epsilon_{\mathrm{ax}}\})$, $s_0 > 0$, and a finite $\chi \geq b_{\mathrm{s}}/b_{\mathrm{r}}$. For $\epsilon \in I$, let $S_\epsilon > 0$ solve

$$(1-\epsilon)G_{\mathrm{s},P}(0, S_\epsilon) = \epsilon b_{\mathrm{s}} q(-b_{\mathrm{s}} S_\epsilon). \tag{A.19}$$

This solution is unique. With

$$\begin{aligned}\lambda_P(\epsilon) &= 2\left[(1-\epsilon)G_{\mathrm{r},P}(0,S_\epsilon) + \epsilon b_{\mathrm{r}} q(-b_{\mathrm{s}}S_\epsilon)\right] \\ &= 2\epsilon q(-b_{\mathrm{s}}S_\epsilon)\left[b_{\mathrm{r}} + b_{\mathrm{s}} h_P(0, S_\epsilon)\right],\end{aligned} \tag{A.20}$$

there are constants $C < \infty$ and L_0 such that

$$\sup_{\epsilon \in I}\left|T_{P,\chi}(L,\epsilon) - \frac{L}{\lambda_P(\epsilon)}\right| \leq C, \qquad L \geq L_0. \tag{A.21}$$

The constants may depend on P, b_{r}, b_{s}, s_0, χ, and I.

Proof. For this proof, write $r = w_{\mathrm{r}}$ and $s = w_{\mathrm{s}}$. Suppressing the fixed distribution P, define the mixed contributions

$$\begin{aligned}\mathcal{G}_\epsilon(r,s) &= (1-\epsilon)G_{\mathrm{r},P}(r,s) + \epsilon b_{\mathrm{r}} q(b_{\mathrm{r}} r - b_{\mathrm{s}} s), \\ \mathcal{D}_\epsilon(r,s) &= (1-\epsilon)G_{\mathrm{s},P}(r,s) - \epsilon b_{\mathrm{s}} q(b_{\mathrm{r}} r - b_{\mathrm{s}} s).\end{aligned} \tag{A.22}$$

Thus $\dot{r} = 2r\mathcal{G}_\epsilon$ and $\dot{s} = 2s\mathcal{D}_\epsilon$. Set $g_\epsilon(s) = \mathcal{G}_\epsilon(0,s)$ and $d_\epsilon(s) = \mathcal{D}_\epsilon(0,s)$. All bounds below are uniform over $\epsilon \in I$.

The shortcut-axis balance. The function d_ϵ is strictly decreasing, with

$$d_\epsilon(0) = \frac{(1-\epsilon)\mathbb{E}_P[a_{\mathrm{s}}] - \epsilon b_{\mathrm{s}}}{2} > 0, \qquad \lim_{s\to\infty} d_\epsilon(s) = -\epsilon b_{\mathrm{s}} < 0.$$

It therefore has a unique positive root S_ϵ. Since $d'_\epsilon(S_\epsilon) < 0$, the root varies continuously with ϵ. Both S_ϵ and $\lambda_P(\epsilon) = 2g_\epsilon(S_\epsilon)$ are bounded away from zero on I. Substitution of (A.19) gives the second expression in (A.20).

A bounded region before first correction. Let $a_{\mathrm{s,min}}$ and $a_{\mathrm{s,max}}$ be the smallest and largest majority shortcut strengths. When $m_{\mathrm{ce}} = b_{\mathrm{r}} r - b_{\mathrm{s}} s \leq 0$,

$$\mathcal{D}_\epsilon(r,s) \leq a_{\mathrm{s,max}} e^{-a_{\mathrm{s,min}} s} - \frac{\epsilon_- b_{\mathrm{s}}}{2}.$$

When $m_{\mathrm{ce}} \geq 0$, the support condition $a_{\mathrm{r}} \geq b_{\mathrm{r}}$ implies

$$\frac{q(a_{\mathrm{r}} r + a_{\mathrm{s}} s)}{q(m_{\mathrm{ce}})} \leq 2e^{-(a_{\mathrm{s}}+b_{\mathrm{s}})s}.$$

Hence

$$\mathcal{D}_\epsilon(r,s) \le q(m_{\mathrm{ce}}) \left[2a_{\mathrm{s,max}} e^{-(a_{\mathrm{s,min}}+b_{\mathrm{s}})s} - \epsilon_- b_{\mathrm{s}} \right].$$

Choose $K \ge s_0$ large enough that both upper bounds are negative at $s \ge K$. Then $s(t) \le K$ for all t. The robust weight increases because $\mathcal{G}_\epsilon > 0$. If it remained bounded, $\mathcal{G}_\epsilon$ would have a positive lower bound on the resulting compact region. This would force r to grow exponentially, a contradiction. Thus $r(t) \to \infty$, and every fixed finite query threshold is reached.

Before the first query event, $r < \chi s$. The quantity $\mathcal{D}_\epsilon(0,0)$ has a uniformly positive lower bound on I. Continuity therefore gives $\mathcal{D}_\epsilon(r,s) > 0$ whenever s is sufficiently small and $0 \le r \le \chi s$. There is consequently a constant $s_{\min} > 0$, independent of L, such that $s(t) \ge s_{\min}$ before the event. On this part of the trajectory, $r \le \chi K$ and $\mathcal{G}_\epsilon \ge g_* > 0$. Writing $T = T_{P,\chi}(L,\epsilon)$, we obtain

$$\int_0^T r(t)\,\mathrm{d}t = \int_{e^{-L}}^{r(T)} \frac{\mathrm{d}r}{2\mathcal{G}_\epsilon(r,s(r))} \le \frac{\chi K}{2g_*}. \tag{A.23}$$

The inverse trajectory $s(r)$ is well defined because r increases.

Controlling the accumulated error. Define

$$\begin{aligned} \zeta_\epsilon(s) &= \frac{g_\epsilon(s) - g_\epsilon(S_\epsilon)}{d_\epsilon(s)}, \\ \Psi_\epsilon(s) &= \int_{S_\epsilon}^{s} \zeta_\epsilon(v)\,\frac{\mathrm{d}v}{v}, \qquad F_\epsilon(r,s) = \log r - \Psi_\epsilon(s). \end{aligned} \tag{A.24}$$

At $s = S_\epsilon$, the apparent singularity is removable: $\zeta_\epsilon(S_\epsilon) = g'_\epsilon(S_\epsilon)/d'_\epsilon(S_\epsilon)$. The functions in (A.24) are uniformly bounded on the relevant positive compact intervals, except for the explicit $\log r$ term in F_ϵ. Differentiating along the full mixed flow gives

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F_\epsilon(r,s) &= \lambda_P(\epsilon) + 2\big[\mathcal{G}_\epsilon(r,s) - g_\epsilon(s)\big] \\ &\quad - 2\zeta_\epsilon(s)\big[\mathcal{D}_\epsilon(r,s) - d_\epsilon(s)\big] = \lambda_P(\epsilon) + O(r). \end{aligned} \tag{A.25}$$

Smoothness on the bounded region makes the $O(r)$ estimate uniform. Its time integral is bounded by (A.23). Initially, $F_\epsilon(e^{-L}, s_0) = -L - \Psi_\epsilon(s_0)$. At first correction, $r(T) = \chi s(T)$ and $s(T) \in [s_{\min}, K]$, so $F_\epsilon(r(T), s(T)) = O(1)$. Integration of (A.25) yields $\lambda_P(\epsilon)T = L + O(1)$. The uniform positive lower bound on λ_P proves (A.21). □

The proof follows the full trajectory up to the query event. The coordinate Ψ_ϵ accounts for changes in the shortcut weight, while (A.23) controls the accumulated departure from the axis calculation. The remainder is bounded as $L \to \infty$ for fixed χ and I; it need not be small at a finite initialization, and the estimate does not assert uniformity as $\epsilon \to 0$ or $\chi \to \infty$.

A.2.2 Matched Pairings and Conditional Structure

Consider the two majority distributions

$$\begin{aligned} P &= \mathrm{Unif}\{(0.05,1),(0.95,2),(0.5,3)\}, \\ Q &= \mathrm{Unif}\{(0.5,1),(0.05,2),(0.95,3)\}, \end{aligned} \qquad (b_{\mathrm{r}}, b_{\mathrm{s}}) = (0.05, 2). \tag{A.26}$$

Both robust marginals are uniform on $\{0.05, 0.5, 0.95\}$, and both shortcut marginals are uniform on $\{1,2,3\}$. Their common mean, second-moment matrix, and covariance matrix are

$$\begin{aligned} \boldsymbol{\mu} = \begin{pmatrix} 0.5 \\ 2 \end{pmatrix}, \quad \mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] = \begin{pmatrix} 0.385 & 1.15 \\ 1.15 & 14/3 \end{pmatrix}, \\ \mathrm{Cov}(\boldsymbol{a}) = \begin{pmatrix} 0.135 & 0.15 \\ 0.15 & 2/3 \end{pmatrix}. \end{aligned} \tag{A.27}$$

The shared shortcut marginal gives the same S_ϵ. For this construction, $\epsilon_{\mathrm{ax}} = 1/2$, so every fixed $0 < \epsilon < 1/2$ satisfies the positive-balance condition.

Proposition A.2 (Different times under matched statistics)**.** *For* (A.26)*, fix* $s_0 > 0$*,* $0 < \epsilon < 1/2$*, and any finite* $\chi \geq 40$*. Then* $\lambda_Q(\epsilon) > \lambda_P(\epsilon)$ *and*

$$\lim_{L\to\infty} \frac{T_{P,\chi}(L,\epsilon)}{T_{Q,\chi}(L,\epsilon)} = \frac{\lambda_Q(\epsilon)}{\lambda_P(\epsilon)} > 1. \tag{A.28}$$

In particular, $T_{Q,\chi} < T_{P,\chi}$ *for all sufficiently large* L*.*

Proof. The counterexample contributions to the two axis growth rates are identical. Subtracting the majority contributions gives

$$\lambda_Q - \lambda_P = 0.3(1-\epsilon)\left[q(S_\epsilon) - 2q(2S_\epsilon) + q(3S_\epsilon)\right] > 0. \tag{A.29}$$

The inequality follows from $S_\epsilon > 0$ and

$$q''(z) = \frac{e^z(e^z-1)}{(1+e^z)^3} > 0 \qquad (z > 0).$$

Applying Proposition A.1 to both distributions proves (A.28). □

More generally, let ν denote the shortcut marginal. At the axis,

$$G_{\mathrm{r},P}(0,s) = \sum_{v\in\mathrm{supp}(\nu)} \nu(v)\,\mathbb{E}_P[a_\mathrm{r} \mid a_\mathrm{s} = v]\,q(vs). \tag{A.30}$$

Matching means and covariance constrains the averages of the conditional robust strengths with weights $\nu(v)$ and $v\nu(v)$. The gradient uses the different weights $\nu(v)q(vs)$. The construction above preserves the complete marginals and covariance while changing this gradient-weighted contribution.

Conditional feature means and the initial growth rate. Let μ_s be the common marginal law of the majority shortcut strength. At a robust-axis state $(0, s)$,

$$G_{\mathrm{r},P}(0,s) = \int \mathbb{E}_P[a_\mathrm{r} \mid a_\mathrm{s} = v] q(sv)\,\mathrm{d}\mu_\mathrm{s}(v). \tag{A.31}$$

Thus the common marginal and the conditional robust mean determine the axis growth rate. Covariance fixes only one weighted average of that conditional mean and generally leaves (A.31) undetermined. Even identical conditional means do not fix the later majority field: for finite support, expansion about $w_\mathrm{r} = 0$ gives

$$G_{\mathrm{r},P}(w_\mathrm{r},s) = G_{\mathrm{r},P}(0,s) + w_\mathrm{r}\mathbb{E}_P[a_\mathrm{r}^2 q'(sa_\mathrm{s})] + O(w_\mathrm{r}^2).$$

The next term already involves conditional second moments.

For the three-by-three grid underlying the matched construction, the robust strengths are $(0.05, 0.5, 0.95)$ and shortcut strengths $(1, 2, 3)$, each with uniform marginals. Covariance 0.15 means $\mathbb{E}[a_\mathrm{r}a_\mathrm{s}] = 1.15$. Every feasible pairing has conditional robust means

$$(m_1, m_2, m_3) = (x, 1.05 - 2x, x + 0.45), \qquad 0.05 \leq x \leq 0.5.$$

This follows by solving the mean and cross-moment constraints and using $0.05 \leq m_j \leq 0.95$. The endpoint $x = 0.05$ is P and $x = 0.5$ is Q. For $s > 0$,

$$G_{\mathrm{r},Q}(0,s) - G_{\mathrm{r},P}(0,s) = 0.15\left[q(s) - 2q(2s) + q(3s)\right] > 0, \tag{A.32}$$

because q is strictly convex on the positive half-line. The axis growth for any feasible intermediate pairing lies between these endpoints. At the common shortcut-axis equilibrium, the same ordering holds for the growth coefficient in the first-correction estimate. This orders its leading coefficient; it does not order every finite-initialization correction time.

Table A.2: First correction times for the query thresholds in Figure 3(a). Values are numerical solutions of the full mixed dynamics, rounded to three decimals.

χ	$T_{P,\chi}$	$T_{Q,\chi}$
40	324.345	298.151
80	393.792	411.347
160	542.694	728.321
320	1451.214	4704.802

Finite-stage queries and the majority gradient share. Figure 3(a,b) uses (A.26), with $N = 3000$ majority records, $k = 200$ counterexamples, $\epsilon = k/(N + k) = 1/16$, $w_{\mathrm{r}}(0) = e^{-8}$, and $w_{\mathrm{s}}(0) = 2$. Each majority support point has 1000 records. All $N + k$ records have equal weight in the loss. The mixture and initialization remain fixed as the query threshold changes; different queries are evaluated along the same trajectory for each pairing.

For the training counterexamples, $\chi = 40$ and Q corrects earlier. At $\chi = 320$, the time under Q is approximately 3.24 times that under P. This finite-stage reversal is compatible with (A.28): the latter takes $L \to \infty$ at each fixed query threshold, whereas the table fixes $L = 8$.

For panel (b), let $*$ index the majority support point with the smallest robust strength. The plotted quantity is

$$\frac{p_* a_{\mathrm{r},*} q(a_{\mathrm{r},*} w_{\mathrm{r}} + a_{\mathrm{s},*} w_{\mathrm{s}})}{G_{\mathrm{r},P}(\boldsymbol{w})}. \tag{A.33}$$

It is the fraction of the majority robust contribution supplied by that point. Counterexample contributions are excluded from the denominator. The factors $2w_{\mathrm{r}}(1 - \epsilon)$ cancel when forming the corresponding fraction of the robust weight update. The point is $(0.05, 1)$ under P and $(0.05, 2)$ under Q. Panel (b) uses w_{r} as its horizontal coordinate; equal horizontal positions represent equal robust weights, which the two trajectories may reach at different times.

The numerical integration uses fourth-order Runge–Kutta in log effective weights and $\log(1 + t)$, with a maximum step of 2.5×10^{-4} in the latter coordinate. First crossings are interpolated in the log weight ratio. The plotted query thresholds are $40 + 0.5j$, $j = 0, \ldots, 560$. Halving the step from 5×10^{-4} changes their correction times by less than 4.81×10^{-5} in absolute value and 6.86×10^{-8} in relative value. These checks assess numerical convergence of the full-flow results; the time comparison in Proposition A.2 follows from the preceding proof.

A squared-loss control. The role of the loss can be checked without another training experiment. Let $\boldsymbol{z} = y\boldsymbol{x}$ under the full mixture and retain the same factorization $\boldsymbol{w}(\theta)$. For squared loss,

$$L_P^{\mathrm{sq}}(\theta) = \tfrac{1}{2} - \boldsymbol{w}(\theta)^\top \mathbb{E}_P[\boldsymbol{z}] + \tfrac{1}{2}\boldsymbol{w}(\theta)^\top \mathbb{E}_P[\boldsymbol{z}\boldsymbol{z}^\top]\boldsymbol{w}(\theta).$$

Matching feature marginals, covariance, and counterexamples matches both moments, so the two risks are identical at every parameter value. Deterministic gradient flow and full-gradient descent therefore have identical trajectories from the same initialization. Single-sample SGD need not have the same noise law. This exact control concerns the specified linear prediction rule and does not extend the logistic correction-time theorem to arbitrary regression models.

Figure A.2 relates the full set of majority gradient shares to the evolving weight states.

A.3 Counterexample Requirements under a Training Budget

A.3.1 Growth, Coverage, and First Entry

For N majority records and k counterexamples with equal record weights, set

$$\alpha = \frac{k}{N}, \qquad \epsilon_k = \frac{\alpha}{1 + \alpha}, \qquad \Lambda_P(\alpha) = \lambda_P\left(\frac{\alpha}{1 + \alpha}\right).$$

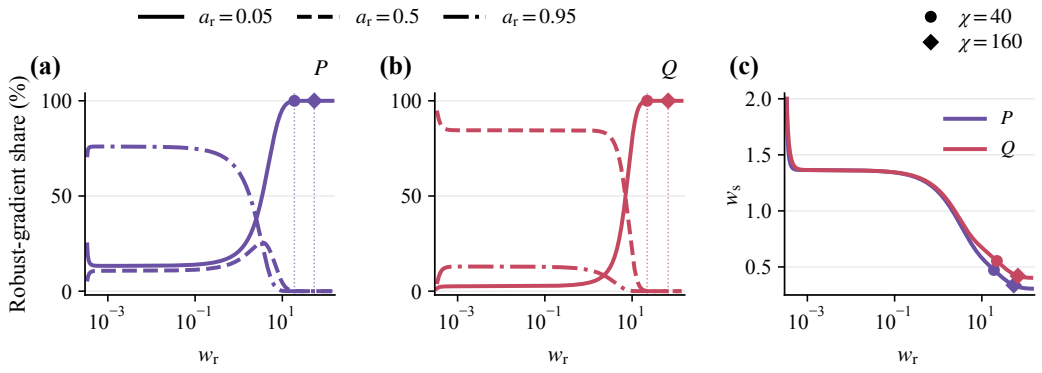


Figure A.2: **Changing majority contributions.** (a,b) All three majority robust-gradient shares under P and Q. (c) The corresponding weight states. Circles and diamonds mark first arrival at $\chi = 40$ and 160 along the same full-flow trajectories. The horizontal coordinate matches robust weights, not training times.

The loss is averaged over all $N + k$ records. Writing $r = w_r$ and $s = w_s$ in the calculations below, the resulting dynamics are

$$\begin{aligned} \dot{r} &= \frac{2r}{1+\alpha}\left[G_{r,P}(r,s) + \alpha b_r q(b_r r - b_s s)\right], \\ \dot{s} &= \frac{2s}{1+\alpha}\left[G_{s,P}(r,s) - \alpha b_s q(b_r r - b_s s)\right]. \end{aligned} \tag{A.34}$$

We retain $r(0) = e^{-L}$ and $s(0) = s_0$ for every count. Thus changing k changes both mixture contributions and their common normalization, while leaving the initialization fixed.

The result below combines the time needed to grow an initially weak robust weight with the time needed to reach a weak-signal query. We first establish the latter time scale. Assume that the finite majority support satisfies $b_r \le a_{r,i} < a_{s,i}$. There is a unique point $*$ with $a_{r,*} = b_r$, probability $p_* > 0$, and shortcut strength $a_{s,*}$; every other point has $a_{r,i} > b_r$.

The late shortcut weight and the query time scale. For fixed $\alpha > 0$ in the model's admissible range, define

$$S_\infty(\alpha) = \left[\frac{1}{a_{s,*} + b_s} \log \frac{p_* a_{s,*}}{\alpha b_s}\right]_+, \qquad [z]_+ = \max\{z, 0\}. \tag{A.35}$$

This quantity is the late shortcut weight, distinct from the early shortcut-axis balance S_ϵ.

Proposition A.3 (Weak-query time scale)**.** *Under the support conditions above, every positive initial state satisfies $r(t) \to \infty$ and $s(t) \to S_\infty(\alpha)$. If $S_\infty(\alpha) > 0$, then*

$$s(t) \ge \min\{s(0), S_\infty(\alpha)\} \qquad (t \ge 0). \tag{A.36}$$

Let $R_{P,\chi}(\alpha; \boldsymbol{w}_{\text{init}})$ be the first time $r/s \ge \chi$ for a trajectory started from a positive effective weight vector $\boldsymbol{w}_{\text{init}}$. Then

$$\frac{\log R_{P,\chi}(\alpha; \boldsymbol{w}_{\text{init}})}{\chi} \longrightarrow b_r S_\infty(\alpha) \qquad \text{as } \chi \to \infty. \tag{A.37}$$

The convergence is uniform when α and $\boldsymbol{w}_{\text{init}}$ range over fixed compact subsets of the positive admissible parameter range and $(0, \infty)^2$, respectively.

Proof. The shortcut upper barrier in the proof of Proposition A.1 uses $a_{r,i} \ge b_r$ and $\epsilon > 0$. It therefore gives $s \le K$ here as well, uniformly on the stated compact sets. The same argument gives $r \to \infty$. Since r increases, use $\log r$ as a coordinate. Dividing the equations in (A.34),

$$\frac{\mathrm{d}\log s}{\mathrm{d}\log r} = \Phi_\alpha(r,s) := \frac{G_{s,P}(r,s) - \alpha b_s q(b_r r - b_s s)}{G_{r,P}(r,s) + \alpha b_r q(b_r r - b_s s)}. \tag{A.38}$$

Divide numerator and denominator by $e^{-b_r r}$. The points with $a_{r,i} > b_r$ vanish in the limit, uniformly for $0 \le s \le K$. Thus

$$\Phi_\alpha(r,s) \longrightarrow \phi_\alpha(s) := \frac{p_* a_{s,*} - \alpha b_s e^{(a_{s,*}+b_s)s}}{b_r\left[p_* + \alpha e^{(a_{s,*}+b_s)s}\right]}. \tag{A.39}$$

If $S_\infty > 0$, this limiting field is positive below S_∞ and negative above it. If $S_\infty = 0$, it is negative for every $s > 0$.

For the finite-time lower barrier, use the exact logistic inequality $q(x+v)/q(x) > e^{-v}$ for $v > 0$. At $0 < s \le S_\infty$,

$$p_* a_{s,*} \frac{q(b_r r + a_{s,*} s)}{q(b_r r - b_s s)} > p_* a_{s,*} e^{-(a_{s,*}+b_s)s} \ge \alpha b_s.$$

The distinguished majority point alone gives $\dot{s} > 0$ there; the other majority contributions are positive. This proves (A.36) and excludes convergence to zero when $S_\infty > 0$. On compact intervals away from S_∞, the limiting field has a strict sign. The uniform convergence in (A.39) then traps s in every fixed neighborhood of S_∞, proving the asserted limit.

These trapping arguments are uniform on the stated parameter and initial-state compact sets. To see the lower-side control when S_∞ is positive, the finite-time barrier bounds s away from zero whenever S_∞ is separated from zero. When S_∞ approaches zero, it suffices to trap s below any fixed positive tolerance. Above that tolerance and the limiting equilibrium, the field has a uniformly negative sign. Consequently, for every $\delta > 0$, one can choose a common robust weight beyond which $|s - S_\infty| < \delta$.

For all sufficiently large r, the robust contribution obeys

$$C_- e^{-b_r r} \le G_{r,P}(r,s) + \alpha b_r q(b_r r - b_s s) \le C_+ e^{-b_r r}$$

for positive constants C_-, C_+ uniform on these compact sets. Here C_-, C_+ are bounds used only in this proof. Integrating the reciprocal robust velocity gives

$$\log t(r) = b_r r + O(\log r). \tag{A.40}$$

The accumulated time to a fixed large robust weight is uniformly bounded and does not affect this estimate.

At the first query event, write $r_\chi = \chi s(r_\chi)$. These event weights tend to infinity uniformly as χ grows. Indeed, on any bounded positive interval of r, the bound $\Phi_\alpha \ge -b_s/b_r$ keeps s uniformly away from zero, so r/s is bounded there. Combining (A.40) with $r_\chi \le K\chi$ yields

$$\frac{\log R_{P,\chi}}{\chi} = b_r s(r_\chi) + o(1) \longrightarrow b_r S_\infty(\alpha).$$

All estimates used above are uniform on the prescribed compact sets. □

The joint count threshold. The positive shortcut-axis balance exists for

$$0 < \alpha < \alpha_{\max}, \qquad \alpha_{\max} = \min\left\{1, \frac{\mathbb{E}_P[a_s]}{b_s}\right\}.$$

The upper restriction $\alpha < 1$ retains $\epsilon < 1/2$. For a fixed budget B, let $k_{\mathrm{hit},P}(B,\chi)$ be the smallest integer $1 \le k < N$ for which the query is first corrected by B, with value $+\infty$ if no such count exists.

Proposition A.4 (Joint growth and coverage threshold). *Assume the support conditions of Proposition A.3 and that Λ_P is strictly increasing on $(0, \alpha_{\max})$. Consider*

$$B \to \infty, \qquad L/B \to \beta > 0, \qquad \chi/\log B \to \rho > 0.$$

Suppose $\Lambda_P^{-1}(\beta)$ exists and

$$\alpha_P^* = \max\left\{\Lambda_P^{-1}(\beta), \frac{p_* a_{s,*}}{b_s} \exp\left[-\frac{a_{s,*} + b_s}{b_r \rho}\right]\right\} \in (0, \alpha_{\max}). \tag{A.41}$$

For every fixed positive separation from α_P^*, all count ratios below the threshold by that separation eventually fail. Every fixed ratio above it and below $\alpha_{\max}$ eventually succeeds. If $N \to \infty$ along record counts that preserve P, then

$$\frac{k_{\mathrm{hit},P}(B,\chi)}{N} \longrightarrow \alpha_P^*. \tag{A.42}$$

Preserving P with equal-weight records means that Np_i is an integer for every support point; for the matched construction, N may grow through multiples of three.

Proof. Set $\chi_0 = b_{\mathrm{s}}/b_{\mathrm{r}}$ and let $\boldsymbol{w}_{\mathrm{hit}}$ be the effective weights at the first training-counterexample correction. For all sufficiently large B, $\chi > \chi_0$, and the exact first query time decomposes as

$$T_{P,\chi}(L,\epsilon) = T_{P,\chi_0}(L,\epsilon) + R_{P,\chi}(\alpha;\boldsymbol{w}_{\mathrm{hit}}). \tag{A.43}$$

On each compact subinterval of $(0,\alpha_{\max})$, Proposition A.1 gives $T_{P,\chi_0} = L/\Lambda_P(\alpha) + O(1)$. Its proof also places $\boldsymbol{w}_{\mathrm{hit}}$ in a common positive compact set, since $r = \chi_0 s$ and s has positive lower and upper bounds at that event. Proposition A.3 therefore applies uniformly to the continuation from these states and gives

$$\frac{T_{P,\chi_0}}{B} \longrightarrow \frac{\beta}{\Lambda_P(\alpha)}, \qquad R_{P,\chi} = B^{\rho b_{\mathrm{r}} S_\infty(\alpha)+o(1)}.$$

If $\Lambda_P(\alpha) < \beta$, the first stage alone exceeds the budget. If $\rho b_{\mathrm{r}} S_\infty(\alpha) > 1$, the continuation exceeds the budget. The two reverse strict inequalities imply that the first stage uses a fraction of B bounded below one and the continuation takes $o(B)$ time. Their sum is then less than B. Solving the two strict conditions for α gives (A.41).

To rule out successful count ratios approaching zero, subtract the log-weight dynamics and use $a_{\mathrm{r},i} < a_{\mathrm{s},i}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\log\frac{r}{s} < 2\epsilon(b_{\mathrm{r}} + b_{\mathrm{s}}).$$

Correction by B consequently requires

$$2\epsilon(b_{\mathrm{r}} + b_{\mathrm{s}})B > L + \log(\chi s_0).$$

Since $L/B \to \beta > 0$, successful candidates stay a positive distance from zero. The remaining candidates below α_P^* by a fixed separation lie in a compact interval, where the preceding estimates are uniform. This proves their eventual failure. For an upper bound, select a ratio strictly between α_P^* and $\alpha_{\max}$ and round $N\alpha$ upward. The ratio changes by at most $1/N$, so the strict success conditions remain valid for sufficiently large N and B. Letting the separation decrease proves (A.42). □

The maximum in (A.41) reflects two requirements on the same count. Above it, the early stage leaves a fixed fraction of the budget and the later stage uses $o(B)$. The result leaves the equality case open. For a fixed finite N, integer rounding remains relevant; the limiting expression does not certify an exact finite-budget minimum count.

Evaluating the threshold for the matched construction. We verify the monotonicity assumption for (A.26) before evaluating the threshold. Write $x = e^{S_\epsilon} > 1$ and

$$D_0(x) = 6x^5 + x^4 + x^3 + 11x^2 - x + 6.$$

Solving the common axis-balance equation and substituting into the growth rates gives

$$\begin{aligned}
\epsilon(x) &= \frac{x^4 + x^3 + 5x^2 - x + 6}{D_0(x)},\\
\lambda_P(x) &= \frac{x^2(3x^4 + 37x^3 + 29x^2 - 3x + 66)}{10(x^2+1)D_0(x)},\\
\lambda_Q(x) &= \frac{x^2(21x^4 - 17x^3 + 83x^2 - 21x + 66)}{10(x^2+1)D_0(x)}.
\end{aligned} \tag{A.44}$$

The function $\epsilon(x)$ is strictly decreasing: the majority shortcut contribution decreases with S_ϵ, while the counterexample shortcut contribution increases. For $X \in \{P, Q\}$, define the polynomial

$$H_X(x) = -5(x^2+1)^2 D_0(x)^2 \frac{\mathrm{d}\lambda_X(x)}{\mathrm{d}x}.$$

The coefficients of $H_X(1+z)$, in ascending powers of z, are

$$\begin{aligned} P : (&216, 8592, 43136, 111708, 186846, 219444, 187780, \\ &118049, 53668, 16837, 3305, 330, 9), \\ Q : (&216, 8592, 41840, 103572, 164382, 182940, 149152, \\ &90959, 41842, 14479, 3701, 654, 63). \end{aligned} \tag{A.45}$$

All are positive, so λ_P and λ_Q decrease with $x > 1$ and increase with $\epsilon \in (0, 1/2)$. Both range from 0 to $11/40$. Thus the inverse growth thresholds used below are well defined.

At $\beta = 0.025$, they are $\Lambda_P^{-1}(\beta) = 0.0350081381$ and $\Lambda_Q^{-1}(\beta) = 0.0139509766$. The weakest robust point is paired with shortcut strength 1 under P and 2 under Q, so

$$\begin{aligned} \alpha_P^*(\rho) &= \max\{0.0350081381, \tfrac{1}{6}e^{-60/\rho}\}, \\ \alpha_Q^*(\rho) &= \max\{0.0139509766, \tfrac{1}{3}e^{-80/\rho}\}. \end{aligned} \tag{A.46}$$

These are the curves in Figure 3(c). Their governing constraints switch at $\rho_P = 38.4513$ and $\rho_Q = 25.2080$. Between these values, P is growth-limited and Q is coverage-limited. Their intersection is

$$\rho_{PQ} = \frac{80}{\log[(1/3)/0.0350081381]} \approx 35.4993.$$

Below ρ_Q, the ratio α_Q^*/α_P^* is a constant less than one. Between the two switches it increases strictly; above ρ_P, it equals $2e^{-20/\rho}$ and continues to increase. Thus the ordering crosses exactly once.

Table A.3: Joint asymptotic count thresholds at $\beta = 0.025$. The last two columns scale the limiting ratios by 3000 for comparison; they are not certified finite-budget integer counts.

ρ	α_P^*	α_Q^*	$3000\alpha_P^*$	$3000\alpha_Q^*$
20	0.035008	0.013951	105.02	41.85
30	0.035008	0.023161	105.02	69.48
40	0.037188	0.045112	111.57	135.34
60	0.061313	0.087866	183.94	263.60

The values are obtained from the axis-balance equation and the closed-form coverage term. We solve the scalar balance and growth threshold equations by bisection; no training trajectories are used to fit these thresholds.

A budget-independent lower bound for reaching weak queries. We now change the support assumption, as in the second part of Section 4.2. Suppose $a_{\mathrm{r},i} < a_{\mathrm{s},i}$ for every majority point, one distinguished point satisfies $a_{\mathrm{r},*} < b_{\mathrm{r}}$, and the remaining robust strengths exceed b_{r}. The late-limit formula (A.35) was derived under different support conditions and is not used here.

Proposition A.5 (Necessary count for first entry)**.** *Assume the initial query is incorrect and*

$$\chi > \frac{a_{\mathrm{s},*} + b_{\mathrm{s}}}{b_{\mathrm{r}} - a_{\mathrm{r},*}}. \tag{A.47}$$

For every training budget B,

$$k_{\mathrm{hit},P}(B, \chi) \geq \left\lceil \frac{Np_*(a_{\mathrm{s},*} - a_{\mathrm{r},*})}{b_{\mathrm{r}} + b_{\mathrm{s}}} \right\rceil. \tag{A.48}$$

Every count below this bound fails to reach the query boundary, regardless of the duration of training.

Proof. At the first arrival at $s = r/\chi$ from $s/r > 1/\chi$, the derivative of $\log(r/s)$ must be nonnegative. Let $q_i = q(a_{\mathrm{r},i}r + a_{\mathrm{s},i}s)$ and $q_{\mathrm{ce}} = q(b_{\mathrm{r}}r - b_{\mathrm{s}}s)$. Subtracting the two log-weight equations gives the necessary condition

$$k(b_{\mathrm{r}} + b_{\mathrm{s}}) \geq N \sum_i p_i (a_{\mathrm{s},i} - a_{\mathrm{r},i}) \frac{q_i}{q_{\mathrm{ce}}}. \tag{A.49}$$

On the boundary, (A.47) implies

$$b_{\mathrm{r}} - b_{\mathrm{s}}/\chi > a_{\mathrm{r},*} + a_{\mathrm{s},*}/\chi > 0.$$

Since q decreases strictly and $r > 0$,

$$\frac{q_*}{q_{\mathrm{ce}}} = \frac{1 + e^{(b_{\mathrm{r}} - b_{\mathrm{s}}/\chi)r}}{1 + e^{(a_{\mathrm{r},*} + a_{\mathrm{s},*}/\chi)r}} > 1.$$

All remaining summands in (A.49) are positive. Hence a first arrival requires $k > Np_*(a_{\mathrm{s},*} - a_{\mathrm{r},*})/(b_{\mathrm{r}} + b_{\mathrm{s}})$, which implies (A.48). The argument contains no budget term, so it applies to every finite first-arrival time. □

For the shifted construction in Figure 3(d,e), $a_{\mathrm{r},*} = 0.045$, $p_* = 1/3$, $N = 3000$, and $(b_{\mathrm{r}}, b_{\mathrm{s}}) = (0.05, 2)$. The distinguished shortcut strengths remain 1 under P and 2 under Q. At $\chi = 4000$, condition (A.47) holds for both, and (A.48) gives 466 and 954 counterexamples, respectively. These are the first-entry reference bounds in panel (d). They do not establish the sufficient count or the count required to retain the target at the chosen endpoint.

A.3.2 Endpoint Guarantees

The first arrival at a query boundary and success at a prescribed training endpoint impose different requirements. We now give the comparison arguments behind the endpoint bounds in Figure 3(d). For this subsection and the coverage analysis below, write $r = w_{\mathrm{r}}$, $s = w_{\mathrm{s}}$, $c_i = a_{\mathrm{r},i}$, $a_i = a_{\mathrm{s},i}$, $d = b_{\mathrm{r}}$, and $b = b_{\mathrm{s}}$. These are local abbreviations for the features and weights already defined in Section 2. With $\alpha = k/N$, set

$$\begin{aligned} C(r,s) &= \sum_i p_i c_i q(c_i r + a_i s), & A(r,s) &= \sum_i p_i a_i q(c_i r + a_i s), \\ q_{\mathrm{c}}(r,s) &= q(dr - bs), & \Phi_\alpha(r,s) &= \frac{A(r,s) - \alpha b q_{\mathrm{c}}(r,s)}{C(r,s) + \alpha d q_{\mathrm{c}}(r,s)}. \end{aligned} \tag{A.50}$$

The normalized population flow satisfies

$$\frac{\dot r}{r} = \frac{2(C + \alpha d q_{\mathrm{c}})}{1 + \alpha}, \qquad \frac{\dot s}{s} = \frac{2(A - \alpha b q_{\mathrm{c}})}{1 + \alpha}, \qquad \frac{\mathrm{d}\log s}{\mathrm{d}\log r} = \Phi_\alpha(r,s). \tag{A.51}$$

In particular, r increases strictly and can serve as the independent variable. All time bounds below retain the factor $1 + \alpha$.

For uniform weak-signal queries $y(dU, -b)$ with U uniform on $(0,1)$, the error is $E_\alpha(t) = \min\{1, bs(t)/(dr(t))\}$. Given an error target $0 < \delta < 1$, put $\vartheta = \delta d/b$; the target is $s/r \leq \vartheta$. We assume $s_0/r_0 > \vartheta$, $a_i > c_i > 0$, and one majority point (c, a) with mass p satisfies $c < d$, while $c_i > d$ for all other points. Finally, require $0 < \vartheta < \gamma_{\mathrm{cov}} = (d - c)/(a + b)$. The first-entry and endpoint requirements are

$$\begin{aligned} k_{\mathrm{hit}}(B) &= \min\{k : \text{some } t \leq B \text{ has } s_k(t)/r_k(t) \leq \vartheta\}, \\ k_{\mathrm{end}}(B) &= \min\{k : s_k(B)/r_k(B) \leq \vartheta\}. \end{aligned} \tag{A.52}$$

An empty feasible set has minimum $+\infty$. The first quantity is nonincreasing in B and never exceeds the second. No corresponding monotonicity in B is assumed for the endpoint requirement.

An exclusion bound over the whole trajectory. The preceding first-entry result gives $k_0 = \lceil Np(a - c)/(b + d) \rceil$ as a necessary count. Fix a candidate upper ratio $\bar\alpha \geq \alpha_0 = k_0/N$. For $0 \leq s \leq \vartheta r$, define

$$Q_i^\vartheta(r) = \frac{q((c_i + a_i\vartheta)r)}{q((d - b\vartheta)r)}, \qquad Q_i^0(r) = \frac{q(c_i r)}{q(dr)}.$$

The ratio $q(c_i r + a_i s)/q(dr - bs)$ decreases with s, so its values lie between these two endpoints. Let

$$U = \sum_i p_i(a_i - c_i)Q_i^{\vartheta} - \bar{\alpha}(b+d), \qquad D_{\vartheta} = \sum_i p_i c_i Q_i^{\vartheta} + \bar{\alpha}d,$$
$$D_0 = \sum_i p_i c_i Q_i^0 + \bar{\alpha}d, \qquad U_* = p(a-c)Q_*^{\vartheta} - \bar{\alpha}(b+d),$$
$$D_* = pcQ_*^{\vartheta} + \bar{\alpha}d, \qquad D_{\mathrm{o}} = \sum_{i \neq *} p_i c_i Q_i^0.$$

Two valid lower bounds are

$$L_1 = \begin{cases} U/D_{\vartheta}, & U < 0, \\ U/D_0, & U \geq 0, \end{cases} \quad L_2 = \begin{cases} U_*/D_*, & U_* < 0, \\ U_*/(D_* + D_{\mathrm{o}}), & U_* \geq 0, \end{cases} \quad L = \max\{L_1, L_2\}. \tag{A.53}$$

For every $0 < \alpha \leq \bar{\alpha}$ in the target region,

$$\frac{\mathrm{d}\log(s/r)}{\mathrm{d}\log r} \geq L(r;\bar{\alpha}). \tag{A.54}$$

For L_1, bound numerator and denominator separately, choosing the denominator endpoint according to the numerator's sign. For L_2, discard the other positive numerator terms. The remaining fraction is increasing in Q_*, and the same sign distinction gives the stated bound. Uniformity over smaller counts follows from

$$\partial_\alpha \Phi_\alpha = -\frac{q_{\mathrm{c}}(bC + dA)}{(C + \alpha d q_{\mathrm{c}})^2} < 0. \tag{A.55}$$

Let r_+ solve $p(a-c)Q_*^{\vartheta}(r_+) = \bar{\alpha}(b+d)$. The left-hand side is strictly increasing and unbounded. If $r_+ \leq r_0$, the boundary points outward from initialization onward. Otherwise define

$$\mathcal{D} = \int_{r_0}^{r_+} [-L(v)]_+ \frac{\mathrm{d}v}{v}, \qquad \mathcal{P}(R) = \int_{r_+}^{R} L(v)\frac{\mathrm{d}v}{v},$$

and choose $R_{\mathrm{out}} \geq r_+$ with $\mathcal{P}(R_{\mathrm{out}}) = \mathcal{D}$. Such a radius exists because $L_2 \to (a-c)/c > 0$. The possible decrease of $\log(s/r)$ during a visit to the target region is at most $\mathcal{D}$. Beyond r_+, its increase is at least $\mathcal{P}$, and the boundary prevents reentry. Thus every count $k \leq N\bar{\alpha}$ fails the target whenever $r > R_{\mathrm{out}}$. This argument allows earlier entries and exits.

For any count capable of first entry, $\dot{r}/r \geq 2\alpha_0 dq(dr)/(1+\alpha_0)$. Consequently an explicit upper bound on the arrival time at this exclusion radius is

$$\mathcal{U}(\bar{\alpha}) = \frac{1+\alpha_0}{2\alpha_0 d}\int_{r_0}^{R_{\mathrm{out}}} \frac{1+e^{dv}}{v}\,\mathrm{d}v. \tag{A.56}$$

If $B > \mathcal{U}(\bar{\alpha})$, then $k_{\mathrm{end}}(B) > N\bar{\alpha}$.

Keeping the state in the comparison. The previous bound allows each radius to use a different favorable shortcut state. The stronger bound used in the main figure restricts this freedom. Normalize the robust-gradient contributions by

$$\omega_i = \frac{p_i c_i q(c_i r + a_i s)}{C + \bar{\alpha} d q_{\mathrm{c}}}, \quad \omega_{\mathrm{c}} = \frac{\bar{\alpha} d q_{\mathrm{c}}}{C + \bar{\alpha} d q_{\mathrm{c}}}, \quad \eta_i = \frac{a_i}{c_i}, \quad \eta_{\mathrm{c}} = -\frac{b}{d}.$$

Here η_i denotes a feature ratio, not the SGD step size. Let $\sigma(v) = 1 - q(v)$ and $\ell_i = -a_i\sigma(c_i r + a_i s)$, $\ell_{\mathrm{c}} = b\sigma(dr - bs)$. Differentiating the normalized weights gives the identity

$$\partial_s \Phi_{\bar{\alpha}} = \sum_{i<j} \omega_i \omega_j (\eta_i - \eta_j)(\ell_i - \ell_j), \tag{A.57}$$

where the sum includes the counterexample index.

On a strip $0 \leq s \leq \bar{s}(r)$, set $t_i = p_i c_i q(c_i r + a_i s)/(\bar{\alpha} d q_{\mathrm{c}})$ and $t_{\mathrm{c}} = 1$. Each t_i decreases with s, so endpoint evaluation bounds both t_i and ℓ_i. For each pair, let D_{ij} bound the negative part of

$(\eta_i - \eta_j)(\ell_i - \ell_j)$ on these intervals and let P_{ij} bound $\omega_i\omega_j$. The latter is obtained by maximizing $xy/(o + x + y)^2$ over the endpoint rectangle for t_i, t_j, with o the sum of the lower bounds for the other t's. The four edges suffice: for fixed x, the maximizing y is $o + x$ clipped to its interval, and conversely. If $o = 0$, the maximum on the possible ridge is also attained on an edge. It follows that

$$K(r) = \sum_{i<j} D_{ij}P_{ij}, \qquad \Phi_\alpha(r, s) \geq \phi_0(r) - K(r)s, \quad \phi_0(r) = \Phi_{\bar\alpha}(r, 0), \tag{A.58}$$

uniformly over $0 < \alpha \leq \bar\alpha$ and the entire strip. This follows by integrating (A.57) from zero and using (A.55).

In the target strip, put $x = \log r$, $z = s/r$, and $y = \vartheta/z$. Equation (A.58) implies, while $y \geq 1$,

$$\frac{\mathrm{d}y}{\mathrm{d}x} \leq -[\phi_0(e^x) - 1]y + \vartheta e^x K(e^x).$$

The following reflected linear comparison handles repeated entries. For coefficients $a(x), h(x)$, define

$$J(x) = \int_{x_e}^{x} a(v)\,\mathrm{d}v, \qquad H(x) = \int_{x_e}^{x} e^{J(v)}h(v)\,\mathrm{d}v,$$
$$Y(x) = e^{-J(x)}\left[H(x) + \max\left\{Y_e, \sup_{x_e \leq v \leq x}\left(e^{J(v)} - H(v)\right)\right\}\right]. \tag{A.59}$$

For $Y_e \geq 1$, this is the solution of $Y' = -aY + h$ reflected at the lower boundary one. Multiplication by e^J gives the formula: the running maximum is the smallest nondecreasing correction that keeps $Y \geq 1$. Scalar comparison gives $\max\{1, y(x)\} \leq Y(x)$ when it holds initially, with $a(x) = \phi_0(e^x) - 1$ and $h(x) = \vartheta e^x K(e^x)$. The symbols J, H in this formula are integration coefficients, unrelated to the response terms $J_{\mathrm{dir}}, J_{\mathrm{ind}}$.

The initial comparison state is also bounded from the features. Assume $r_0 \leq 1$. A constant $S \leq s_0$ satisfying

$$\sum_i p_i a_i q(c_i + a_i S) \geq \bar\alpha b q(-bS) \tag{A.60}$$

gives $s(1) \geq S$: the vector field at $s = S$ points upward throughout $r_0 \leq r \leq 1$. Apply (A.58) on $0 \leq s \leq S$ to S/s. Its reflected comparison has coefficients $a(x) = \phi_0(e^x)$, $h(x) = SK(e^x)$ and initial value one. Thus $s(r_e) \geq S/Y_S(\log r_e)$ for any fixed $r_e \geq 1$. Taking the largest lower bound among valid strips remains valid. The reported calculation uses $r_e = 20$ and $S \in \{0.01, 0.025, 0.05, 0.1, 0.2\}$. If this yields $\underline{s}_e$, initialize the target-strip comparison at $Y_e = \max\{1, \vartheta r_e/\underline{s}_e\}$. Let R_* be a radius beyond r_+ where this comparison reaches one. The true trajectory is then on or outside the target boundary, whose field points outward thereafter. For every $r > R_*$ it fails the target.

A tighter clock. Retaining one majority term improves the time bound. Define

$$\mu = \frac{bc + ad}{a + b}, \qquad H_* = (a + b)\left[\frac{(pc)^b(\alpha_0 d)^a}{a^a b^b}\right]^{1/(a+b)}.$$

Minimizing two exponentials over s gives $pce^{-cr-as} + \alpha_0 de^{-dr+bs} \geq H_* e^{-\mu r}$. For $0 \leq \zeta < d$, separate the cases $s \leq (d - \zeta)r/b$ and $s > (d - \zeta)r/b$ to obtain

$$g_\zeta(r) = \min\left\{\frac{H_* e^{-\mu r}}{1 + e^{-\min\{c,\zeta\}r}}, \alpha_0 dq(\zeta r)\right\}.$$

The first case uses positive margins and the identity $q(v) = e^{-v}/(1 + e^{-v})$; the second retains only the counterexample. For any chosen set of allowed ζ values,

$$\widetilde{\mathcal{U}}(\bar\alpha) = \int_{r_0}^{R_*} \frac{\mathrm{d}r}{r}\left[\max\left\{\frac{2\alpha_0 dq(dr)}{1 + \alpha_0}, \frac{2\max_\zeta g_\zeta(r)}{1 + \bar\alpha}\right\}\right]^{-1} \tag{A.61}$$

is an upper bound on the time to the exclusion radius, uniformly for $\alpha_0 \leq \alpha \leq \bar\alpha$. Therefore $B > \widetilde{\mathcal{U}}(\bar\alpha)$ excludes every $k \leq N\bar\alpha$. The calculation uses $\zeta = 0, c, \mu$ and $\mu + r^{-1}\log(\alpha_0 d/H_*)$ clipped to $[0, c]$. Each choice is independently valid; their maximum requires no assumption of global optimization. Candidate counts are checked separately, without assuming monotonicity of the computed bound.

Sufficient counts and a guaranteed time window. For a fixed α, choose $S \geq s_0$ and $R_{\max} > r_0$ with

$$\sum_i p_i a_i \left[e^{-a_i S} + e^{\max\{(d-c_i)r_0, (d-c_i)R_{\max}\} - (a_i+b)S} \right] \leq \alpha b. \tag{A.62}$$

The ratio $q(c_i r + a_i S)/q(dr - bS)$ is bounded by the bracketed quantity, so the field at $s = S$ points downward throughout the stated range. Thus $s(r) \leq S$ there. Since $0 < q(c_i r + a_i s)/q_c \leq Q_i^0(r)$, an upper bound on the phase field is

$$\overline{\Phi}_\alpha(r) = \max_{I \subseteq \{1,\dots,n\}} \frac{\sum_{i \in I} p_i a_i Q_i^0(r) - \alpha b}{\sum_{i \in I} p_i c_i Q_i^0(r) + \alpha d}. \tag{A.63}$$

A linear-fractional function with positive denominator attains its maximum over a box at a vertex, which proves the bound. Equivalently, $v = \overline{\Phi}_\alpha(r)$ is the unique solution of

$$\sum_i p_i (a_i - v c_i)_+ Q_i^0(r) = \alpha(b + vd).$$

This form avoids enumerating subsets when the support is large. For a fixed anchor $r_e \in [r_0, R_{\max}]$, integration gives

$$\overline{s}_\alpha(r) = \begin{cases} S, & r \leq r_e, \\ S \min\left\{1, \exp\left[\displaystyle\int_{r_e}^{r} \overline{\Phi}_\alpha(v) \frac{\mathrm{d}v}{v}\right]\right\}, & r > r_e, \end{cases} \qquad s(r) \leq \overline{s}_\alpha(r). \tag{A.64}$$

Any anchor is valid. The reported calculation chooses the first root of $\overline{\Phi}_\alpha = 0$ and checks all eight vertices for the three-point majority support.

Suppose this envelope is at most ϑr on an interval $[R_{\mathrm{in}}, R_{\mathrm{out}}] \subset [r_0, R_{\max}]$. Define the two speed bounds

$$\begin{aligned} V_-(r) &= \frac{2}{1+\alpha} \left[\sum_i p_i c_i q(c_i r + a_i \overline{s}_\alpha(r)) + \alpha d q(dr) \right], \\ V_+(r) &= \frac{2}{1+\alpha} \left[\sum_i p_i c_i q(c_i r) + \alpha d q(dr - b\overline{s}_\alpha(r)) \right], \\ \mathcal{T}_-(R) &= \int_{r_0}^{R} \frac{\mathrm{d}v}{v V_+(v)}, \qquad \mathcal{T}_+(R) = \int_{r_0}^{R} \frac{\mathrm{d}v}{v V_-(v)}. \end{aligned} \tag{A.65}$$

Monotonicity of q gives $V_- \leq \dot{r}/r \leq V_+$, so the true arrival time at R lies between $\mathcal{T}_-(R)$ and $\mathcal{T}_+(R)$.

Proposition A.6 (Endpoint guarantee)**.** *Under the upper-barrier and envelope conditions above,*

$$\mathcal{T}_+(R_{\mathrm{in}}) \leq B \leq \mathcal{T}_-(R_{\mathrm{out}}) \quad \Longrightarrow \quad E_\alpha(B) \leq \delta. \tag{A.66}$$

Proof. The first inequality ensures that the trajectory has reached R_{in}; the second ensures that it has not passed R_{out}. Since r increases, the state is within the interval where $s(r)/r \leq \vartheta$. □

If the lower time endpoint exceeds the upper one, this comparison gives no guarantee. It does not prove failure of the actual trajectory. Likewise, monotonicity of $s_\alpha(r)$ in α, which follows from (A.55), does not establish monotonicity at a fixed time: changing the count also changes the clock.

Numerical evaluation. The shifted construction subtracts 0.005 from every majority robust strength, with $N = 3000$, $(r_0, s_0) = (e^{-8}, 2)$, $(d, b) = (0.05, 2)$, and target $\delta = 0.01$. Thus $c = 0.045$, $\vartheta = 0.00025$, and k_0 equals 466 for P and 954 for Q. The sufficient comparisons use $S = 2$ and $R_{\max} = 1000$. Table A.4 combines exclusions of all smaller counts with individually sufficient counts. These are bounds on the unknown minimum, not exact thresholds or uncertainty intervals.

At $B = 10^4$, the bounds imply $k_{\mathrm{end},Q}/k_{\mathrm{end},P} \geq 1201/700 > 1.715$. The corresponding lower ratios at 10^6 and 10^8 exceed 1.626 and 1.648. For P with 700 counterexamples, the guaranteed

Table A.4: Bounds on the minimum counterexample count for 1% endpoint error in the shifted construction.

Budget B	$k_{\mathrm{end},P}(B)$	$k_{\mathrm{end},Q}(B)$
10^4	$[601, 700]$	$[1201, 1400]$
10^6	$[651, 800]$	$[1301, 1500]$
10^8	$[701, 850]$	$[1401, 1600]$

state interval is $[33.1585, 175.0009]$, giving a time window $[670.493, 11828.499]$. For Q with 1400, these intervals are $[15.7825, 229.9132]$ and $[307.623, 104615.178]$. The sufficient-bound candidates $P, 650$ and $Q, 1200$ yield no certified time window, although the existing full-flow trajectories do reach the target temporarily.

Table A.5: Exclusion-time bounds and the return above 1% error in previously computed full-flow trajectories. Bounds exclude all counts up to the stated candidate; the last column concerns that candidate alone.

Pairing, count	Basic bound	State bound	Tighter clock	Actual return
$P, 600$	431573.73	10676.72	4015.50	1453.96
$P, 650$	68554731.25	208310.18	42762.04	26179.38
$Q, 1100$	5001.11	1420.18	1168.02	289.81
$Q, 1200$	371351.77	7439.45	3672.46	2214.89

Table A.6: Previously computed full-flow checks at six guaranteed endpoints. All errors are percentages; the target is 1%.

Pairing, count	Budget	Error (%)
$P, 700$	10^4	0.018134
$P, 800$	10^6	0.001084
$P, 850$	10^8	0.001635
$Q, 1400$	10^4	0.000514
$Q, 1500$	10^6	0.002633
$Q, 1600$	10^8	0.013500

The coefficients and bounds are evaluated by scalar quadrature and root finding, without using future training states as inputs. Recorded precision checks change the $P, 600$ exclusion radius by 2.03×10^{-11} in relative terms. For the $P, 700$ guarantee, the time endpoints change by 1.55×10^{-9} and 1.91×10^{-12}. These checks assess floating-point evaluation; the comparison inequalities provide the mathematical guarantees.

Figure A.3 places the count calculations beside their finite-flow checks and guaranteed time windows.

A.4 Coverage and Persistence of Correction

This subsection explains why changing the range of counterexamples can have a different effect from repeating the same type. We retain the local notation of (A.50). The query distribution remains fixed throughout each comparison.

A positive limiting error despite correct training counterexamples. Assume $a_i > c_i > 0$, with one distinguished point (c, a) of mass p satisfying $c < d$. Set

$$\gamma_{\mathrm{cov}} = \frac{d - c}{a + b}, \qquad \mu = \frac{bc + ad}{a + b}, \qquad \Delta_i = (c_i - c) + (a_i - a)\gamma_{\mathrm{cov}}.$$

Require $\Delta_i > 0$ for the other majority points. This is the strict separation of their margins along the proposed limiting direction.

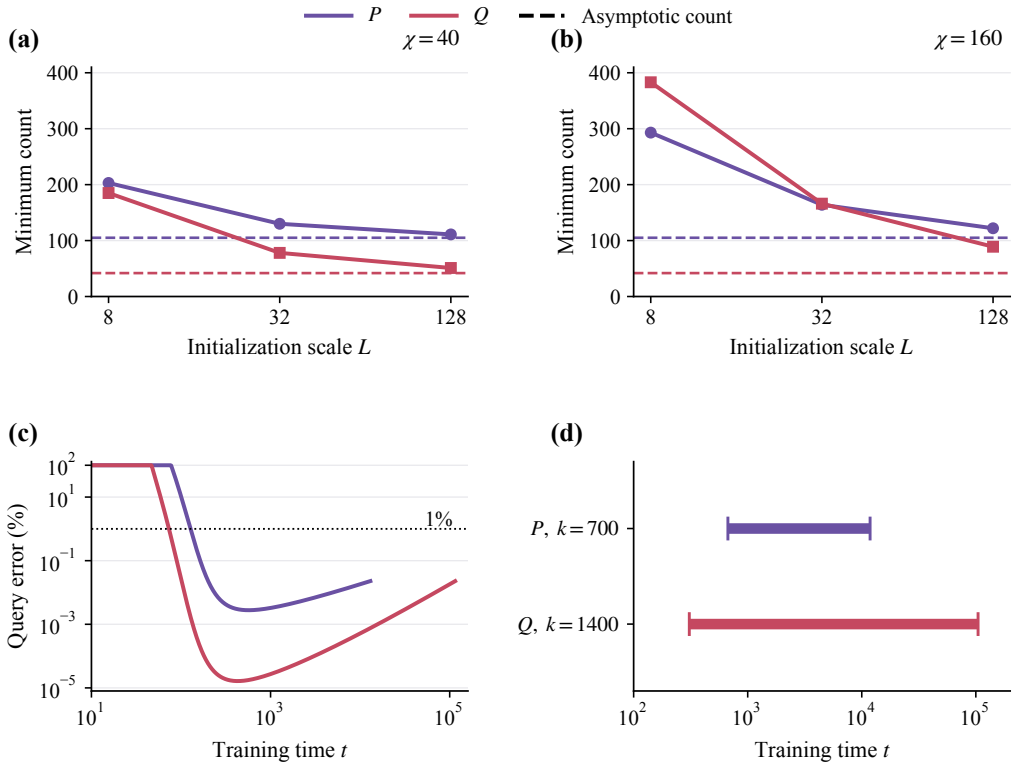


Figure A.3: **Count bounds and valid training windows.** (a,b) Minimum integer counts for $\chi = 40$ and 160, with $w_{\mathrm{r}}(0) = e^{-L}$ and budget $40L$; dashed lines are asymptotic counts. (c,d) Full-flow query errors and sufficient 1% time windows for $P, k = 700$ and $Q, k = 1400$. These panels instead use the shifted majority support and fixed $L = 8$. Window endpoints are conservative guarantees, not observed first-entry or exit times.

Proposition A.7 (Persistent error on weaker queries)**.** *For every fixed $\alpha > 0$ and positive initialization,*

$$s(r) = \gamma_{\mathrm{cov}} r + z_\alpha + o(1), \qquad z_\alpha = \frac{1}{a+b} \log \frac{p(a-c)}{\alpha(b+d)}, \qquad r(t) \sim \frac{\log t}{\mu}. \tag{A.67}$$

For queries $y(dU, -b)$ with CDF G continuous at $b\gamma_{\mathrm{cov}}/d$, the limiting error is $G(b\gamma_{\mathrm{cov}}/d)$, while the training-counterexample margin tends to $+\infty$.

Proof. First, r cannot remain bounded. If it did, the shortcut field would be negative for all sufficiently large s, making s bounded too. On the resulting compact set, $\dot{r}/r$ has a positive lower bound, a contradiction. Hence $r \to \infty$.

Put $z = s/r$. If $z \geq \gamma_{\mathrm{cov}} + \xi$ for a fixed $\xi > 0$, every majority margin exceeds the counterexample margin by a positive multiple of r. The exact logistic ratios then give $\Phi_\alpha \to -b/d$ uniformly in this region. If $0 < z \leq \gamma_{\mathrm{cov}} - \xi$, the counterexample gradient is exponentially smaller than that of the distinguished majority point. Since all $a_i/c_i > 1$, the phase field $\Phi_\alpha - 1$ has a positive lower bound there for large r. These signs force entry into, and prevent exit from, the cone $\gamma_{\mathrm{cov}} - \xi < z < \gamma_{\mathrm{cov}} + \xi$. Letting ξ decrease gives $s/r \to \gamma_{\mathrm{cov}}$.

Now write $h = s - \gamma_{\mathrm{cov}} r = o(r)$. The two relevant margins are $\mu r + ah$ and $\mu r - bh$. The remaining margin gaps are $\Delta_i r + (a_i - a)h$ and therefore diverge. Logistic tail expansion yields

$$\Phi_\alpha(r, \gamma_{\mathrm{cov}} r + h) = F(h) + o(1), \qquad F(h) = \frac{pa - \alpha b e^{(a+b)h}}{pc + \alpha d e^{(a+b)h}}.$$

The error is uniform on compact h intervals. Along the two tails with $h = o(r)$, the limits are $a/c > 1$ and $-b/d < 0$. Since $\mathrm{d}h/\mathrm{d}r = \gamma_{\mathrm{cov}}(F(h) - 1) + (h/r)F(h) + o(1)$, fixed sufficiently negative and positive boundaries point inward. Scalar comparison first bounds h and then gives convergence to the unique root of $F(h) = 1$, which is z_α. Substitution into the robust equation gives $\dot{r} = Kre^{-\mu r}(1 + o(1))$ for a constant $K > 0$; integration proves the time asymptotic. Finally, $dr - bs = \mu r - bz_\alpha + o(1) \to \infty$ and $bs/(dr) \to b\gamma_{\mathrm{cov}}/d$. □

The limiting direction does not depend on a fixed positive count, although the offset z_α and finite-time behavior do. In the shifted construction, the uniform-query limits are 6.67% for P and 5% for Q. An error target below these limits cannot hold at every sufficiently late time for any fixed count of the original counterexample type. This does not preclude success at a specified finite budget.

For example, with 600 counterexamples under P, the recorded query error reaches about 0.31930% near time 335.135, then rises to 3.3141% at 10^4 and 5.3692% at 10^6. The training-counterexample margin is already positive during this return. Correction of that training type and coverage of weaker queries are therefore distinct events.

Figure A.4 follows query error, weight states, and the balance between majority reinforcement and counterexample suppression. Its gradient-balance diagnostic is

$$\frac{\sum_i p_i(a_i - c_i)q(c_i r + a_i s)}{\alpha(b+d)q(dr - bs)}.$$

A value above one makes $\log(s/r)$ increase. The top panels hold $k = 600$ fixed and shift all majority robust strengths by -0.005, 0, or $+0.005$. The lower panels use the negative shift with $k = 600$ under P and $k = 1200$ under Q.

Replacing counterexamples at a fixed count. Fix a total of K counterexamples. Replace M examples of type $(d_1, -b)$ by type $(d_2, -b)$, where $0 < d_2 < d_1$. The replacement count is denoted M to keep m available for margins. Let $\alpha_1 = (K - M)/N$, $\alpha_2 = M/N$, and $q_j = q(d_j r - bs)$. The mixture is fixed from initialization; no examples are inserted partway through training. Its flow is

$$\begin{aligned} \frac{\dot{r}}{r} &= \frac{2}{1 + K/N}\big[C + \alpha_1 d_1 q_1 + \alpha_2 d_2 q_2\big], \\ \frac{\dot{s}}{s} &= \frac{2}{1 + K/N}\big[A - b(\alpha_1 q_1 + \alpha_2 q_2)\big]. \end{aligned} \tag{A.68}$$

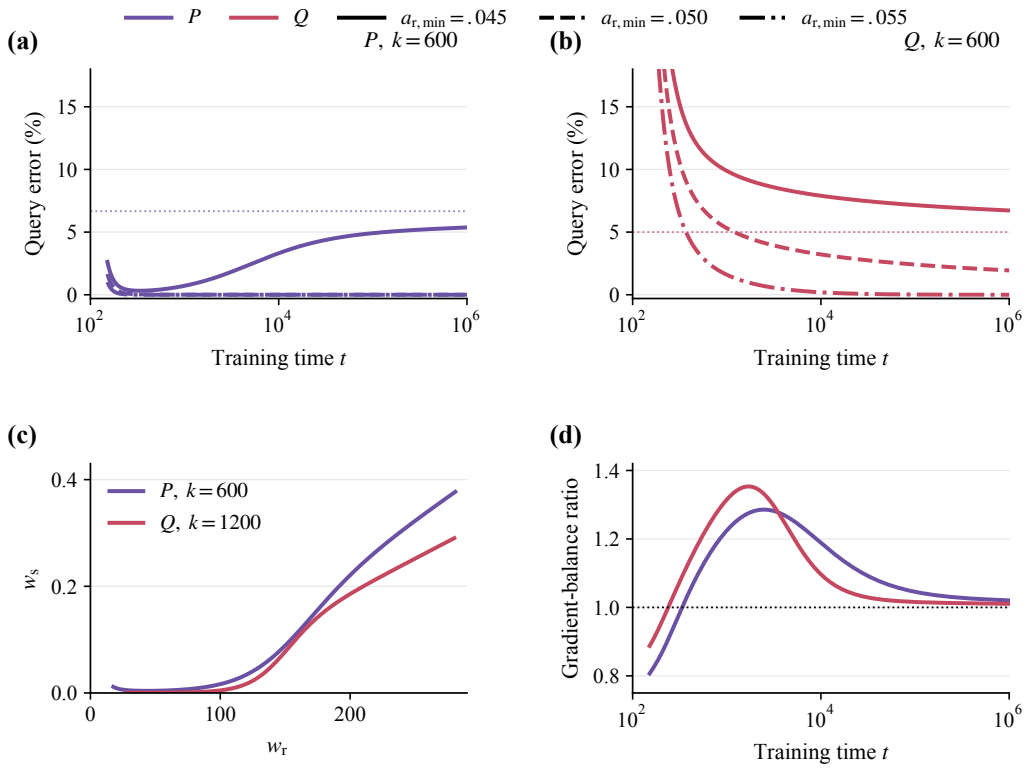


Figure A.4: **Correction and its later persistence.** (a,b) Weak-query errors under three majority-strength shifts, with 600 counterexamples; dotted lines give the negative-shift limits. (c,d) Weight states and the gradient-balance ratio for the negative shift, using $P, k = 600$ and $Q, k = 1200$. Values above one in (d) increase the shortcut-to-robust weight ratio. Counterexamples remain in the mixture throughout.

Proposition A.8 (Coverage from a weaker counterexample type)**.** *If* $0 < d_2 < \min\{d_1, \min_i c_i\}$ *and* $\alpha_2 > 0$*, then*

$$s(r) = K_2 r^{-b/d_2}(1 + o(1)), \qquad r(t) \sim \frac{\log t}{d_2}, \qquad K_2 > 0. \tag{A.69}$$

Every fixed query $y(d_0 u, -b_0)$ *with* $d_0, u, b_0 > 0$ *is eventually classified correctly. Uniform weak-query error tends to zero.*

Proof. The boundedness contradiction used above again gives $r \to \infty$. Uniformly for $s \geq 0$,

$$\frac{q(c_i r + a_i s)}{q(d_2 r - bs)} \leq e^{-c_i r - a_i s} + e^{-(c_i - d_2)r - (a_i + b)s} \to 0.$$

Thus the second counterexample type eventually dominates every positive majority shortcut term, and s decreases and remains bounded. The first counterexample's gradient divided by the second then also vanishes exponentially. Dividing the two flow equations yields $\mathrm{d}\log s / \mathrm{d}\log r = -b/d_2 + O(e^{-\nu r})$ for some $\nu > 0$. The remainder is integrable, giving a positive finite K_2. The robust equation is asymptotically $\dot{r} = 2\alpha_2 d_2 r e^{-d_2 r}(1 + o(1))/(1 + K/N)$, which proves the time relation. The query conclusion follows from $s/r \to 0$. □

A single replacement satisfies this asymptotic condition in the deterministic finite-support model. Its influence at a finite budget can still be small. At a common state, replacement changes the derivative of $\log(s/r)$ by

$$-\frac{2M}{N + K}\mathcal{H}(r, s), \qquad \mathcal{H}(r, s) = (b + d_2)q_2 - (b + d_1)q_1. \tag{A.70}$$

For each fixed $s \geq 0$, $\mathcal{H}$ has one zero as r increases. Indeed its sign is that of

$$-(d_1 - d_2) + e^{-bs}\big[(b + d_2)e^{d_1 r} - (b + d_1)e^{d_2 r}\big].$$

This expression is negative at zero, strictly increasing, and tends to infinity. The replacement can therefore initially slow the decrease in s/r and later strengthen it. This is a statement about the field at a common state, not an ordering of two complete trajectories.

The weaker type's share of counterexample shortcut suppression is

$$\pi_2(r, s) = \frac{M q_2}{(K - M)q_1 + M q_2}. \tag{A.71}$$

For $0 < M < K/2$, the inequality $q_2/q_1 < e^{(d_1 - d_2)r}$ gives

$$\pi_2 \geq \tfrac{1}{2} \quad \Longrightarrow \quad r > \frac{\log((K - M)/M)}{d_1 - d_2}. \tag{A.72}$$

With $K = 600$, $d_1 = 0.05$, and $d_2 = 0.025$, the necessary radius is 117.78 for $M = 30$, compared with observed radii 119.64 and 120.15 under P and Q. For $M = 1$, it is 255.81; the observed crossing occurs near 255.9 and only after about 3.15×10^5 and 3.70×10^5 time units. All other settings match the shifted construction above.

Table A.7: Fixed-count replacement with $K = 600$ and $d_2 = 0.025$. Errors are percentages on the unchanged uniform weak-query distribution. Below precision denotes a reported rounded value, not exact zero.

Pairing	Replacements M	Error at 10^4	Error at 10^6
P	0	3.3141	5.3692
P	1	2.7043	1.7572
P	30	0.000476	Below precision
Q	0	7.8981	6.7249
Q	1	7.3778	3.4315
Q	30	0.6996	Below precision

Current gradient shares do not determine the learned state. In a separate, smaller-support-change comparison, take $d_2 = 0.04$ and $M = 120$, keeping $K = 600$. At the same robust weight $r = 170$, the recorded shortcut weights are 0.00152965 for P and 0.182046 for Q. Their current weaker-type gradient shares are 57.7567% and 57.7470%, respectively, but the query errors are 0.0360% and 4.2834%. Similar present shares coexist with different accumulated states.

This history dependence has an exact expression. Set $x = \log r$, $z_P(x) = \log s_P(e^x)$, and define $f_P(x, z)$ by dividing the two equations in (A.68); define f_Q likewise. For $D = z_P - z_Q$, put

$$\mathcal{F}(x) = f_P(x, z_Q(x)) - f_Q(x, z_Q(x)), \qquad \mathcal{K}(x) = \int_0^1 \partial_z f_P(x, z_Q(x) + vD(x))\,\mathrm{d}v.$$

The fundamental theorem of calculus gives $D' = \mathcal{F} + \mathcal{K}D$. Common initialization then implies

$$D(x) = \int_{x_0}^{x} \mathcal{F}(u) \exp\left[\int_u^x \mathcal{K}(v)\,\mathrm{d}v\right] \mathrm{d}u. \tag{A.73}$$

The forcing compares the pairings at the same state; the exponential factor retains subsequent state feedback. No small-perturbation approximation is required.

Figure A.5 compares the replacement trajectories, gradient takeover, and resulting state differences.

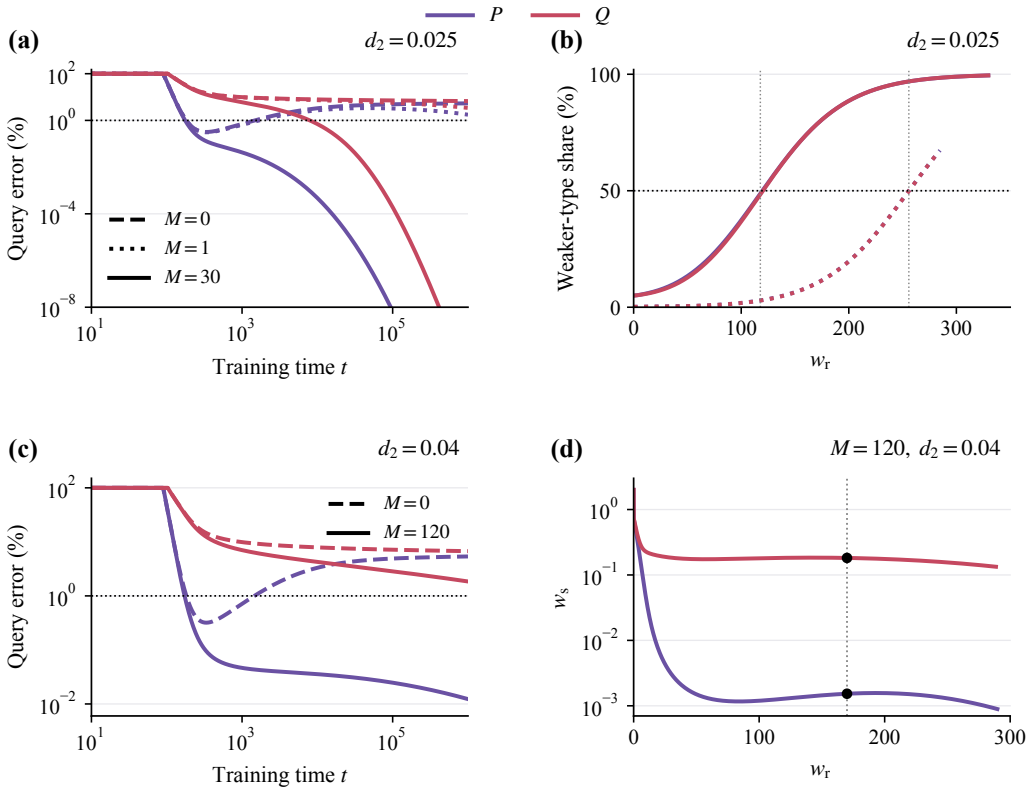


Figure A.5: **Coverage changes at a fixed counterexample count.** All panels retain $K = 600$ counterexamples. With $d_2 = 0.025$, (a) compares $M = 0, 1, 30$ replacements and (b) shows the weaker type's share of the counterexample shortcut gradient; vertical lines give the half-share lower bounds. With $d_2 = 0.04$, (c) compares $M = 0, 120$ and (d) marks the two states at $w_r = 170$. Query errors are percentages.

B MECHANISM TESTS: PROTOCOLS AND FULL RESULTS

B.1 STOCHASTIC-GRADIENT TESTS

Paired updates and initialization. The experiments use the matched distributions in Equation (A.26) and train the four factors with stochastic gradient descent. For each counterexample probability, the two pairings share initialization, learning rate, batch size, and paired random streams. Across counterexample probabilities, the shortcut initialization follows $w_{\mathrm{s}}(0) = \log(1/\epsilon)/a_{\mathrm{s,min}}$, so comparisons across probabilities do not share one fixed initial shortcut weight. The robust initialization is held fixed across these conditions. Table B.1 uses batch size one, learning rate $\eta = 0.1$, and 64 paired runs per condition. The removal condition masks only majority-sample updates to the two robust factors. All counterexample updates and majority shortcut updates remain active.

For the distributions in Figure 4(a), $\epsilon = 0.001$. The common scaled clock is $t_{\mathrm{SGD}} = \epsilon\eta K_{\mathrm{step}}$, where K_{step} counts SGD steps. The stopping event is the first nonnegative counterexample margin. No run in the displayed conditions is censored.

First correction times. Table B.1 reports the ratio of mean first correction times, rather than the mean of within-pair ratios. The reported intervals resample paired random streams with 2000 bootstrap samples. They describe this controlled stochastic experiment.

Table B.1: Ratios of mean first correction times, P/Q. Each condition has 64 paired runs with batch size one. Intervals concern first correction.

ϵ	Majority robust updates	First correction	95% interval
0.01	Retained	1.585	[1.543, 1.632]
0.001	Retained	2.168	[2.125, 2.212]
0.0001	Retained	2.397	[2.354, 2.443]
0.001	Removed	1.018	[1.014, 1.022]

At $\epsilon = 0.001$, removing majority robust updates slows both pairings (Table B.2). The ratio of means falls from 2.168 to 1.018. Pairing P is slower in all 64 full update pairs and in 60 of the 64 removal pairs. The intervention therefore greatly reduces the timing separation while preserving small differences between individual paired runs.

Table B.2: Mean first correction times on the common clock $\epsilon\eta K_{\mathrm{step}}$, with $\epsilon = 0.001$ and $\eta = 0.1$.

Update condition	P	Q	Ratio P/Q
Full updates	24.973928	11.520244	2.168
No majority robust updates	86.114106	84.609952	1.018

In Q, the minimum-margin majority point changes from $(0.5, 1)$ to $(0.05, 2)$ once $w_{\mathrm{r}}/w_{\mathrm{s}} > 20/9$. The counterexample boundary is reached later, at ratio 40. In P, $(0.05, 1)$ remains the minimum-margin point. The gradient weights change smoothly through this exchange; there is no discontinuous switch in the actual majority gradient. At $\epsilon = 10^{-4}$ under Q, first correction occurs after a mean of 117.516 nonempty counterexample batches, about 70.3% of which arrive after entry into the later state region. This percentage counts arrivals, not the share of elapsed time or gradient.

Figure B.1 connects the condition-level timing results with the prespecified actual parameter paths.

Which dynamic information affects finite-time prediction? A separate comparison uses twelve configurations with 48 paired random streams each. The full predictor retains all majority gradients, the sampled counterexample intervals, and the unsaturated logistic weight of each counterexample update. It averages majority-sample index noise while continuing to evolve the state. Predictions are saved before the corresponding reference SGD runs. The full, minimum-margin, and fixed-weight predictors share the reference counterexample schedule. The mean-interval predictor replaces it by deterministic intervals. The first three comparisons are therefore conditional on the saved schedule.

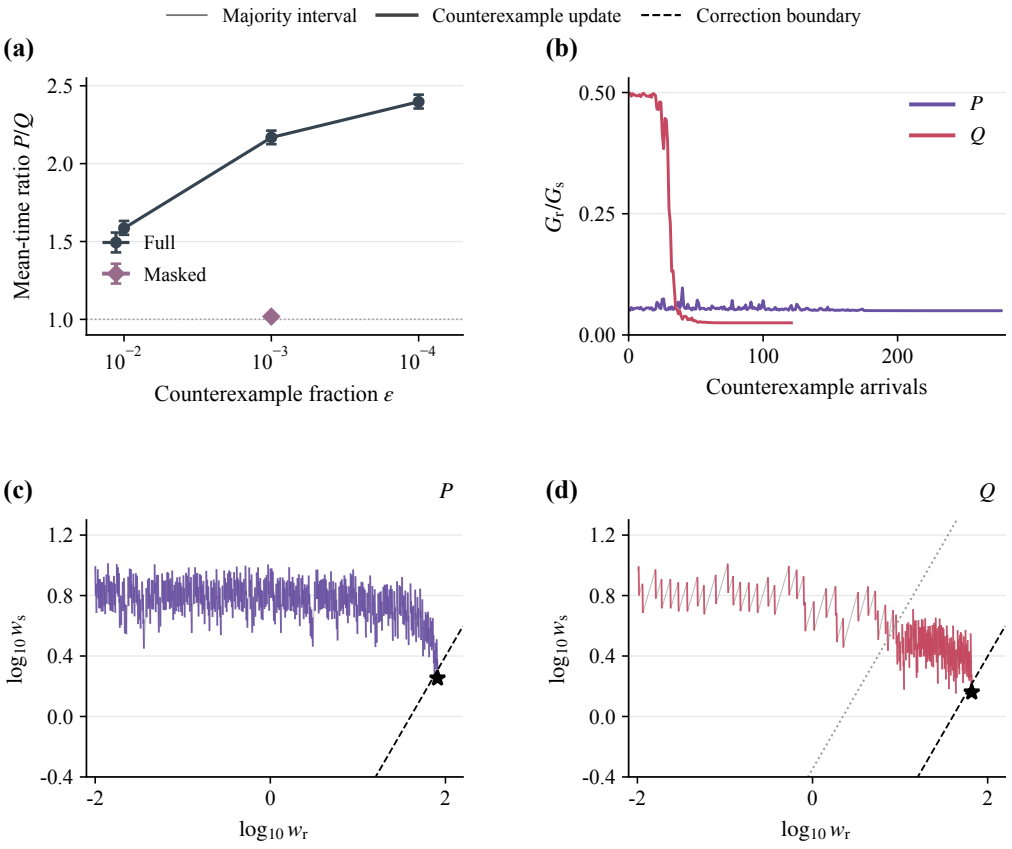


Figure B.1: **Stochastic correction times and actual paths.** (a) Ratios of mean first correction times with 64 paired runs and 95% paired-bootstrap intervals; "Masked" removes only majority robust updates. (b–d) Prespecified repeat-zero paths at $\epsilon = 10^{-4}$ and batch size one: the majority gradient ratio, followed by weight states for P and Q. Gray segments are majority intervals and colored segments are counterexample updates. Dashed lines mark correction; the dotted line in (d) marks the minimum-margin exchange.

Table B.3 specifies all twelve configurations. Archive sources A and B are the paper's P and Q. The additional uniform pairings are

$$P' = \mathrm{Unif}\{(0.15, 0.8), (0.85, 1.6), (0.5, 2.4)\},$$
$$Q' = \mathrm{Unif}\{(0.5, 0.8), (0.15, 1.6), (0.85, 2.4)\},$$

called C and D in the archive. They also match both feature marginals and covariance. Every configuration uses counterexample $(0.05, -2)$, batch size one, positive balanced factors, and $w_{\mathrm{s}}(0) = \log(1/\epsilon)/a_{\mathrm{s,min}}$.

Table B.3: All finite-time prediction configurations, with 48 paired streams each. The labels match the prediction figure.

ID	Pairing	ϵ	η	$w_{\mathrm{r}}(0)$	$w_{\mathrm{s}}(0)$
C01	P	0.003	0.10	0.01	5.809143
C02	P	0.003	0.05	0.01	5.809143
C03	P	0.0003	0.10	0.01	8.111728
C04	P	0.0003	0.05	0.01	8.111728
C05	Q	0.003	0.10	0.01	5.809143
C06	Q	0.003	0.05	0.01	5.809143
C07	Q	0.0003	0.10	0.01	8.111728
C08	Q	0.0003	0.05	0.01	8.111728
C09	P'	0.003	0.10	0.02	7.261429
C10	P'	0.0003	0.10	0.02	10.139660
C11	Q'	0.003	0.10	0.02	7.261429
C12	Q'	0.0003	0.10	0.02	10.139660

For counterexample intervals $W_j \sim \mathrm{Geom}(\epsilon)$ on $\{1, 2, \ldots\}$, the full predictor integrates the majority-only factor flow for $\eta(W_j - 1)$ time units and then applies the exact factor update on the counterexample. This preserves its state-dependent logistic weight and is distinct from replacing it by a gradient-flow step. The minimum-margin variant changes only the majority field; the fixed-weight variant replaces the counterexample's $q(m)$ by one. The mean-interval variant uses $W_j = 1/\epsilon$. Integration uses step-doubled RK4 in log weights, with local absolute tolerance 10^{-9}. Repeat zero is also evaluated at 10^{-10}, checking the event index and common-arrival log states.

Geometric intervals and majority indices use independent pseudo-random streams. For repeat index $j = 0, \ldots, 47$, their seeds are respectively `0x32c0ffee` and `0x6c8e9cf5`, each plus $(j + 1)$ `0x9e3779b9` modulo 2^{32}. The archived generator implementation fixes the resulting sequences. Each run has a presaved cap of 4096 counterexample intervals; no reference run or prediction in this comparison is censored. Reference SGD updates the four factors synchronously and samples majority indices; those indices are not provided to the predictor.

Table B.4 reports all-condition summaries. The full predictor has per-configuration mean bias from -0.325% to $+0.599\%$; 566 of 576 paired trajectory errors are at most 5%, while the largest is 17.17%. For the additional P'/Q' pairings, the reference mean-time ratio is 1.5517 and the predicted ratio is 1.5532. The bias range is a range across configurations, not a confidence interval.

Table B.4: Prediction ablations across twelve configurations. Mean bias is $100(\text{predicted mean}/\text{reference mean} - 1)$. The last column counts configurations whose absolute mean bias is at most 5%.

Predictor	Mean-bias range (%)	Conditions within 5%
Full dynamic predictor	$[-0.33, +0.60]$	12/12
Only current minimum-margin majority point	$[+1.64, +21.70]$	5/12
Counterexample gradient weight fixed at one	$[-13.57, -0.67]$	6/12
Counterexample intervals fixed at their mean	$[+0.67, +10.73]$	7/12

The minimum-margin predictor retains the current hardest point, with its original mass $1/3$; it does not freeze the initial identity. Its errors show why even a dynamically selected single point can miss the contribution of the remaining majority samples. These empirical prediction errors do not certify

a bound for each original SGD trajectory. The accompanying data contain all 576 paired reference records, 2304 prediction endpoints, and 48 configuration–predictor summaries. Full-predictor and reference event traces are archived for all streams; the ablated predictors retain complete traces only for repeat zero. The experiment ran on CPU with Node.js v24.19.0; its configuration, protocol, code snapshot, and prediction seal accompany the records.

Figure B.2 displays all twelve configurations, retaining both mean bias and individual-trajectory accuracy.

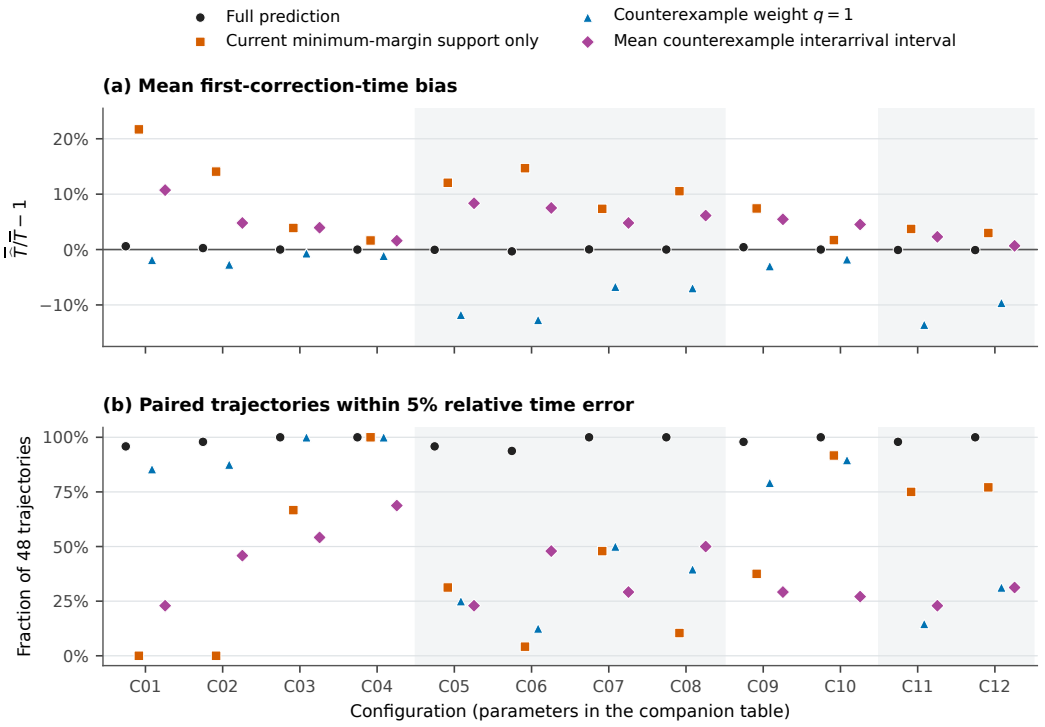


Figure B.2: **Finite-time prediction across configurations.** Signed mean-time bias and the fraction of paired trajectories within 5% error for the four predictors. Each configuration contains 48 reference trajectories; only the mean-interval variant replaces their saved counterexample schedules.

B.2 Neural Probes and Channel Interventions

Majority contribution to the image channel. The channel experiments use an image network h_{ϕ} and a separate scalar background input x_{s}:

$$f(\boldsymbol{x}, x_{\mathrm{s}}) = u_{\mathrm{r}} v_{\mathrm{r}} h_{\phi}(\boldsymbol{x}) + u_{\mathrm{s}} v_{\mathrm{s}} x_{\mathrm{s}}. \tag{B.1}$$

Here h_{ϕ} includes the archived fixed output normalization. The image pathway consists of $\phi, u_{\mathrm{r}}, v_{\mathrm{r}}$; the background pathway consists of $u_{\mathrm{s}}, v_{\mathrm{s}}$. For a majority training view i, write $h_i = y_i h_{\phi}(\boldsymbol{x}_i)$, $a_{\mathrm{s},i} = y_i x_{\mathrm{s},i}$, and $m_i = w_{\mathrm{r}} h_i + w_{\mathrm{s}} a_{\mathrm{s},i}$. Let ρ_i be its original risk weight, including its identity, view, amplitude, and majority mixture weight. The recorded drive is

$$G_M = \sum_{i \in \mathrm{majority}} \rho_i h_i q(m_i), \qquad \left.\frac{\mathrm{d} \log w_{\mathrm{r}}}{\mathrm{d}t}\right|_{\mathrm{majority}} = 2 G_M. \tag{B.2}$$

The derivative identity refers to balanced-factor gradient flow; the probe evaluates its gradient quantity at the actual discrete network states. For fixed positive image features and normalized majority weights, it reduces to $(1 - \epsilon) G_{\mathrm{r},P}$. It is distinct from the image-network gradient norm, whose vector also depends on $\nabla_{\phi} h_{\phi}$. Unlike the theoretical robust feature, learned signed image outputs h_i are not assumed positive on every input.

Virtual updates at actual training states. The four settings combine two Fashion-MNIST tasks, T-shirt/top versus pullover and sandal versus ankle boot, with a tanh CNN and a SiLU residual CNN. Each task uses 512 majority identities, 32 counterexample identities, and 512 held-out confirmation identities. Counterexamples carry risk mass 0.05. For each setting, three majority-source trajectories provide 60 predetermined checkpoints at steps $0, 50, \ldots, 2950$, giving 12 trajectories and 720 states. The source archive calls these sources M/A/B; the checkpoint protocol retains all three and does not average them into independent repeats.

At each checkpoint, a virtual copy receives only the current counterexample background-factor update, with the image network and image factors fixed. The update uses that checkpoint's archived effective step size from the full-parameter norm rule. We evaluate G_M before and after on the same weighted majority views. The copy is discarded after this local measurement; the original mixed-training trajectory supplies the next checkpoint. The plotted relative response is $100(G_M^+ - G_M^0)/G_M^0$. All 720 recorded before/after values are positive, and every recorded response is positive. Checkpoints within a trajectory are dependent.

Read-only forward replay also evaluates each majority margin and $q(m)$ using the original weights over identities, views, and background amplitudes. Means and quantiles use these weights; the mean gradient weight is the mean of $q(m_i)$, not q applied to the mean margin. The weighted mean margin decreases and the weighted mean $q(m_i)$ increases in every probe. Replayed and archived G_M values agree to a maximum relative discrepancy of 4.15×10^{-8}. These comparisons measure a local response at a fixed state rather than a cumulative intervention effect. The paired 720-state table, weighted margin/q summaries, replay code, and comparison report accompany the figure. The replay uses float64 models and margins with the archived float32 image-parameter displacements, without a training update. The tanh and SiLU residual image networks contain 46177 and 71265 parameters, respectively. The archived data seed is 1787.

Figure B.3 shows all three local measurements. Tops denotes T-shirt/top versus pullover; footwear denotes sandal versus ankle boot. Each line retains one majority source and network setting.

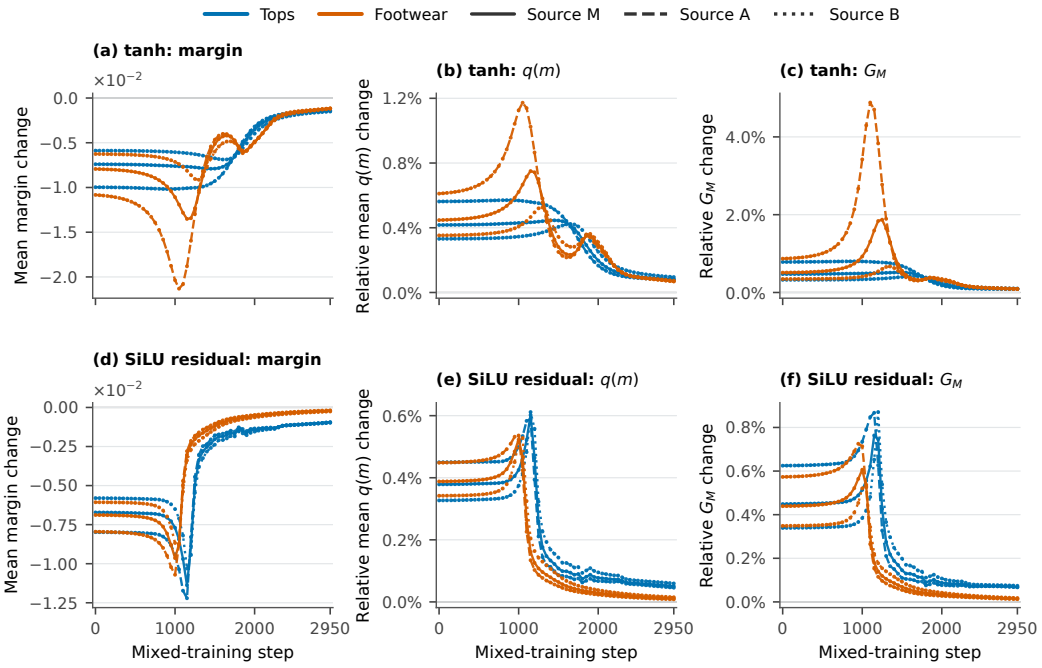


Figure B.3: **Local responses in the image/background models.** Weighted majority-margin changes, relative changes in mean $q(m)$, and relative changes in G_M after the same virtual background update. The twelve checkpoint trajectories retain their task, architecture, and majority-source identities.

Complementary channel interventions. We test whether counterexamples must directly update the image pathway for correction to occur. The experiments use two Fashion-MNIST tasks, T-shirt/top versus pullover and sandal versus ankle boot, each with a `tanh` CNN and a SiLU residual CNN. Within each setting, the interventions start from the same pretrained image network and initial predictor, run for $3{,}000$ steps, and match the full parameter-update norm at each step. The image factor r and background coefficient s in Figure B.4 parameterize the image and background pathways of these networks.

In the *CE-background-only* condition, counterexamples update only the background factors. Their updates to both the image factor and the image network are removed; majority samples retain all update channels. We compare this condition with removing all counterexample updates and with restricting both sample groups to background updates. The full-update trajectory provides an additional reference.

The CE-background-only condition reaches endpoint worst-group accuracies of 62.11%–86.33% across the four settings (Table B.5). Both restricted controls remain at 0% at the recorded checkpoints. Since counterexamples cannot update the image pathway in the reference intervention, its image learning comes entirely from majority samples. The result supports the complementary role of counterexample background updates and majority image updates. Zero worst-group accuracy does not mean that every test image is misclassified.

The middle column partitions the squared norm of the complete update into three disjoint parameter blocks: the image factor, the image network, and the background factors. In this intervention, the first two blocks receive only majority updates. The background block includes the net update from both sample groups. These shares describe where parameter updates occur; they are not an additive decomposition of accuracy gains. The source figure records these shares through step $2{,}950$, whereas accuracy trajectories include step $3{,}000$.

This intervention has two relevant limits. First, the pretrained image network already contains useful information, so the result does not establish that the network discovers robust features from scratch. Second, matching the total update norm allows the background gradient to affect the scale of image updates in the same step. The accuracy differences therefore do not isolate cross-step state feedback alone. These channel interventions also address a different comparison from changing feature pairing; their large accuracy differences should not be reported as pairing effects.

Table B.5: Endpoint worst-group accuracy (%) after $3{,}000$ steps. Values are the reported outcomes of the four displayed settings.

Task	Network	CE background only	No CE updates	All background only
T-shirt/top–pullover	`tanh`	62.11	0.00	0.00
T-shirt/top–pullover	SiLU residual	86.33	0.00	0.00
Sandal–ankle boot	`tanh`	67.19	0.00	0.00
Sandal–ankle boot	SiLU residual	85.16	0.00	0.00

B.3 Matched-Norm Tests in Neural Networks

Data and training. We use 8×8 handwritten digits 3 and 8 with synthetic backgrounds. Background signs agree with labels on majority samples and disagree on counterexamples. Within a split, training and test digit identities are disjoint. The pairings retain the same digit identities, labels, counterexamples, contrast and background-amplitude marginals, and matched first, second, and label-signed first input moments. Only the contrast/background-strength pairing changes.

After common majority pretraining, we train a tanh network with widths 96–32–16–1 on the full weighted risk. Counterexamples have loss weight 0.05; this is a risk weight, rather than a statement about the number of generated inputs. For matched-norm training, both pairings use the same per-step Euclidean parameter-update norm, taken from the common intermediate pairing's baseline trajectory. Writing this common length as ℓ_j and a pairing's full gradient as g_j, the update is $\theta^{j+1} = \theta^j - \ell_j g_j/\|g_j\|_2$ whenever $g_j \neq 0$. The norm includes all trained parameters, and the same sequence ℓ_j is applied to both pairings.

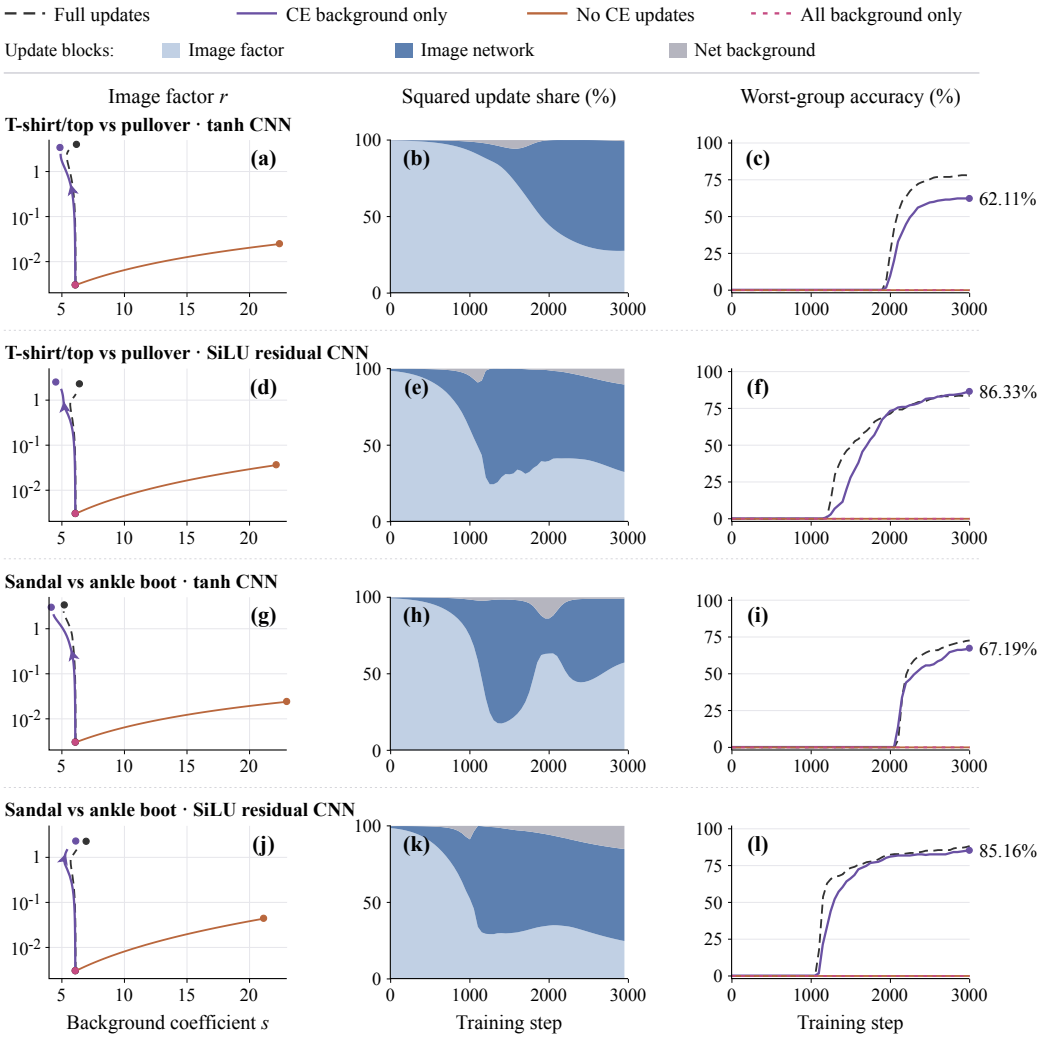


Figure B.4: **Complementary channel interventions.** Each row uses the indicated Fashion-MNIST task and architecture, with matched initialization and per-step parameter-update norms. Left: image/background coefficient trajectories. Middle: squared update-norm shares for the CE-background-only condition. Right: worst-group accuracy; the two zero-accuracy controls overlap. CE denotes counterexamples; "all background only" restricts both sample groups to background updates. Numbers mark reported endpoint accuracies. Full-update references are dashed.

Group accuracies and update-norm matching. Table B.6 gives the step-500 worst-group accuracy differences. The residual gaps under matched norms support a role for the training path beyond unequal update lengths. Because normalization also changes subsequent states, the difference between the two columns is not an additive decomposition of the ordinary accuracy gap.

Table B.6: Worst-group accuracy difference $Q - P$ at step 500, in percentage points. Rows identify the source run's initialization seed.

Split	Seed	Ordinary updates	Matched update norms
Current	109	33.33	16.48
Current	127	42.59	19.07
Current	151	16.67	3.52
Earlier	71	18.70	—
Earlier	83	4.07	—
Earlier	97	0.74	—

The earlier split gives weaker ordinary-update effects and is retained as a separate comparison. A dash indicates that this table has no reported matched-norm value for that earlier run. The two splits are not paired controls for estimating an update-length contribution.

Response predictions on the same saved baseline. The archived digit-network checks also predict the pairing gap from the common intermediate distribution, obtained by interpolating the majority pairings with coefficient $\alpha = 1/2$ while keeping the counterexamples fixed. For ordinary gradient descent, the parameter sensitivity $\boldsymbol{\xi}_j = \partial_\alpha \theta^j$ obeys

$$\boldsymbol{\xi}_{j+1} = (I - \eta H_j)\boldsymbol{\xi}_j - \eta \boldsymbol{b}_j, \qquad H_j = \nabla_\theta^2 L_\alpha(\theta^j), \quad \boldsymbol{b}_j = \partial_\alpha \nabla_\theta L_\alpha(\theta^j). \tag{B.3}$$

For logistic loss, replacing H_j by $G_j^{\mathrm{GGN}} = \sum_i \rho_i q(m_i)(1 - q(m_i))\nabla_\theta m_i \nabla_\theta m_i^\top$ keeps only the generalized Gauss–Newton contribution; ρ_i are the training-risk weights. Both calculations use the same saved baseline states. Matched-norm predictions differentiate the actual normalized update map with its fixed common step lengths. These are predictions across pairing conditions using the complete baseline trajectory, rather than forecasts from the initial state. The current-split predictions were stored before their corresponding branch runs; the earlier split is a retrospective check.

Figure B.5 presents the step-500 comparisons and every saved evaluation time in the current-split ordinary-update check. The response comparisons use the original reported outputs; no new network runs are added.

A separate ordinary-CNN check. The four Fashion-MNIST task/architecture settings also provide a control without separate output-factor branches. After common majority warm-up and 6000 subsequent steps, the matched-norm time-average worst-group accuracy gaps $Q - P$ are $+0.231$, $+2.298$, $+0.503$, and -1.379 percentage points, in the task/architecture order of Table B.5. Their mean is $+0.413$ points. Each setting uses one initialization; the architectures share task data. These results limit the size and consistency of the pairing effect in this additional architecture control, while the channel interventions above address a different question.

Reproduction records. For each paired experiment, reproduction requires the identity split, feature-pairing construction, initial parameters, optimizer settings, and paired sample schedule. The norm-matched experiments additionally require the common update-length sequence; the local probes require the checkpoint and effective step size. These quantities must be preserved together because changing a schedule or normalization changes the controlled comparison.

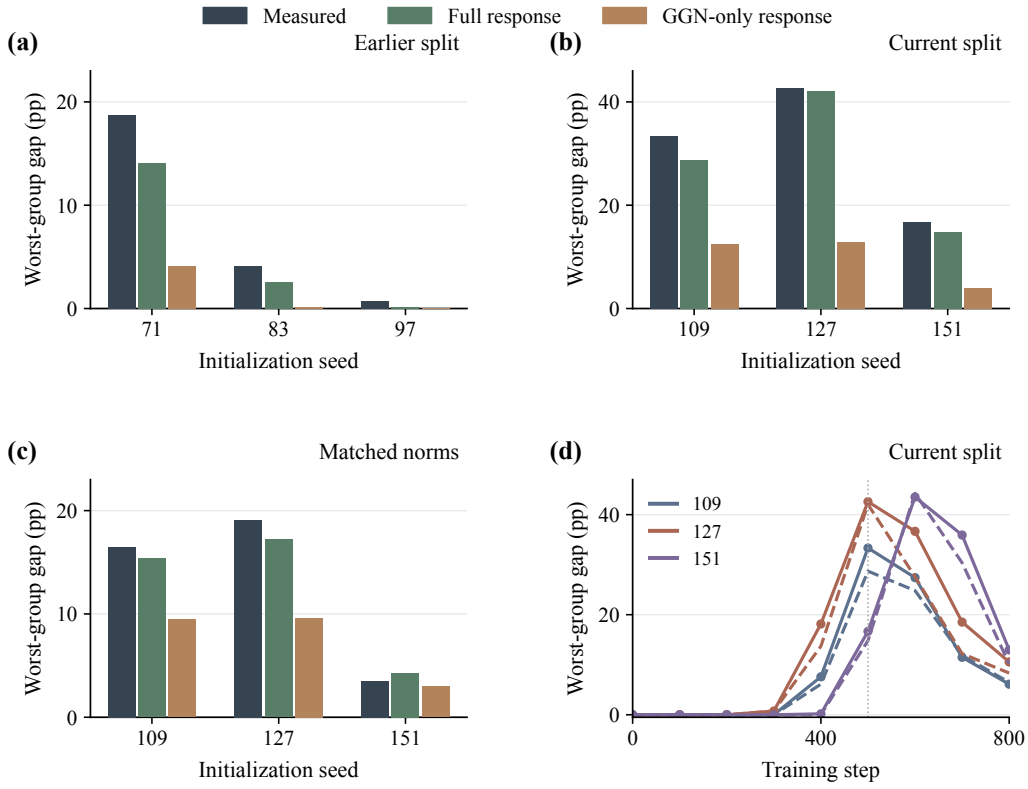


Figure B.5: **Digit-network results across initializations.** Worst-group accuracy gaps $Q - P$, in percentage points. (a–c) Step-500 measurements and response predictions for the earlier split, current split, and current split with matched update norms. GGN-only retains the generalized Gauss–Newton contribution. (d) All saved times for the current ordinary-update runs: solid curves are measured gaps and dashed curves are full-response predictions.

C SK Assessment: Method and Full Validation

This appendix specifies the assessment in Section 5.2 and gives the complete paired results. The assessment predicts the progress of a two-coefficient readout on fixed representations. The validation trains that same readout using the current source-risk gradient. Data, initialization, candidate contexts, and the scoring rule are shared between the two procedures. The figures use *cold* for the predicted path and *warm* for actual readout training. These names do not denote a train/test split or training of the full representation network.

C.1 SK Computation and Validation Protocol

C.1.1 Fixed Readouts and Predicted Dynamics

For one label task, let $\mathcal{Y}$ be its class set and $\boldsymbol{\phi}_i$ its fixed input representation. Three folds rotate through the query role. Each rotation constructs the assessment from the other two folds, called the source set. The source set determines context prototypes, centering, scales, and task directions. Query records are then evaluated with these fixed quantities. The fold rules and representations are specified in Tables C.2 and C.3.

Context logits. Let $\mathcal{Z}$ be the context categories present in the source set, and let $n_{z,y}$ count source records with context z and label y. We use one pseudocount per cell and form

$$\widetilde{C}_z(y) = \log \frac{n_{z,y}+1}{\sum_{z'} n_{z',y} + |\mathcal{Z}|} - \frac{1}{|\mathcal{Y}|}\sum_{y'\in\mathcal{Y}} \log \frac{n_{z,y'}+1}{\sum_{z'} n_{z',y'} + |\mathcal{Z}|}. \tag{C.1}$$

Thus each prototype is a centered vector of context log-likelihoods across classes. For continuous graph size, the source deciles define the categories; repeated quantiles are merged.

Task logits. We impute nonfinite feature entries with source-coordinate medians, using zero when an entire source coordinate is missing. Let $\boldsymbol{\mu}_z$ be the source feature mean within context z, and let $\boldsymbol{\sigma}$ contain the source standard deviations. Coordinates with standard deviation below 10^{-8} use scale one. Define $\boldsymbol{h}_i = (\boldsymbol{\phi}_i - \boldsymbol{\mu}_{z_i}) \oslash \boldsymbol{\sigma}$, where $\oslash$ is elementwise division. A query context absent from the source uses the overall source mean. Writing n for the source size, set

$$\begin{aligned} p_i^0 &= \mathrm{softmax}(\widetilde{C}_{z_i}), \\ \boldsymbol{d}_y &= \frac{1}{n}\sum_{i\in\text{source}} \big(\mathbf{1}\{y_i = y\} - p_i^0(y)\big)\boldsymbol{h}_i, \qquad \widetilde{R}_i(y) = \boldsymbol{h}_i^\top \boldsymbol{d}_y. \end{aligned} \tag{C.2}$$

The same directions $\boldsymbol{d}_y$ are applied to queries. They use source labels only.

Both channels are normalized by their source logit root-mean-square norms, denoted σ_R and σ_C:

$$\sigma_R^2 = \frac{1}{n}\sum_{i\in\text{source}} \|\widetilde{R}_i\|_2^2, \qquad \sigma_C^2 = \frac{1}{n}\sum_{i\in\text{source}} \|\widetilde{C}_{z_i}\|_2^2, \qquad R_i = \frac{\widetilde{R}_i}{\sigma_R}, \quad C_z = \frac{\widetilde{C}_z}{\sigma_C}. \tag{C.3}$$

A channel with norm at most 10^{-14} is set to zero. For a degenerate context channel, we also set $\sigma_C = 0$. The readout logits and initialization are

$$f_{\boldsymbol{c}}(i,z) = c_{\mathrm{r}} R_i + c_{\mathrm{s}} C_z, \qquad \boldsymbol{c}_0 = (0, \sigma_C). \tag{C.4}$$

Initially, these logits reproduce the unnormalized context-only prediction. The coefficients $\boldsymbol{c}$ are trained directly in the validation; they are distinct from Section 2's four trainable factors. The fixed task channel provides a specified reference representation for this diagnostic.

The predicted trajectory. The source risk is the mean cross entropy at native record frequencies:

$$\mathcal{R}(\boldsymbol{c}) = \frac{1}{n}\sum_{i\in\text{source}} \mathrm{CE}\big(f_{\boldsymbol{c}}(i,z_i), y_i\big). \tag{C.5}$$

Let $g_0 = \nabla\mathcal{R}(\boldsymbol{c}_0)$ and $H_0 = \nabla^2\mathcal{R}(\boldsymbol{c}_0)$. The local quadratic risk gives the predicted coefficient path

$$\widehat{\boldsymbol{c}}(t) = \boldsymbol{c}_0 - \int_0^t e^{-uH_0} g_0 \,\mathrm{d}u. \tag{C.6}$$

This solves $\dot{\widehat{\boldsymbol{c}}} = -g_0 - H_0(\widehat{\boldsymbol{c}} - \boldsymbol{c}_0)$. The implementation diagonalizes the two-by-two positive semidefinite Hessian. For each eigenvalue λ, the integration multiplier is $(1 - e^{-t\lambda})/\lambda$, with limit t at zero. This formulation does not require an invertible Hessian. No iterative parameter updates are used to construct the predicted path.

C.1.2 Scoring, Aggregation, and Readout Training

For each query record, hold R_i fixed and evaluate every source context prototype:

$$\widehat{L}_i(t) = \max_{z \in \mathcal{Z}} \mathrm{CE}\big(\widehat{c}_{\mathrm{r}}(t) R_i + \widehat{c}_{\mathrm{s}}(t) C_z, y_i\big), \qquad g_i(t) = 1 - \frac{\widehat{L}_i(t)}{\widehat{L}_i(0)}. \tag{C.7}$$

The denominator belongs to the individual record. We average these ratios after they are formed, retaining negative values. For a task d, query fold f, and class y, let $\mathcal{S}_{d,f,y}(B)$ be $100/B$ times the integral of the class mean of $g_i(t)$. Let $n_{d,f,y}$ be its query-record count, and let J_D denote the number of tasks in dataset D. The complete aggregation is

$$\text{SK-value}(D; B) = \frac{1}{J_D} \sum_{d=1}^{J_D} \frac{1}{|\mathcal{Y}_d|} \sum_{y \in \mathcal{Y}_d} \sum_{f=1}^{3} \frac{n_{d,f,y}}{\sum_{g=1}^{3} n_{d,g,y}} \mathcal{S}_{d,f,y}(B). \tag{C.8}$$

Thus folds are pooled within class by their query counts, classes receive equal weight within each task, and tasks receive equal weight within each dataset. This implements the balanced expectation in Equation (15). The source-risk gradient itself retains the original record frequencies.

Algorithm: SK assessment with fixed representations

Input: Declared dataset and tasks, representations, context rule, three folds, and budget B.
Output: Dataset SK-value and the underlying task scores.

1. For each task and query fold, use the other two folds to construct R_i, C_z, and $\boldsymbol{c}_0$ using Equations (C.1)–(C.4).
2. Evaluate g_0 and H_0 once on the source risk.
3. Evaluate (C.6) at the scoring times. For each query, compute its worst-context loss and relative gain.
4. Integrate the class-mean gains over $[0, B]$. Pool folds, classes, and tasks using (C.8).
5. For validation, replace the predicted path with actual source-gradient updates and repeat the identical query scoring.

Table C.1: Assessment and paired validation. Query records are used for evaluation; the source set constructs the readout and drives training.

Assessment algorithm. The two-by-two Hessian makes trajectory evaluation inexpensive once the fixed logits are available. Representation extraction and source construction remain part of the assessment cost. Query evaluation uses the exact maximum cross entropy over the declared prototypes, including when either readout coefficient changes sign.

Readout training and numerical integration. For each of the 426 instances, the validation starts at the same $\boldsymbol{c}_0$ and updates the two coefficients by

$$\boldsymbol{c}^{j+1} = \boldsymbol{c}^{j} - 0.025\, \nabla \mathcal{R}(\boldsymbol{c}^{j}), \qquad j = 0, \ldots, 319. \tag{C.9}$$

The gradient is recomputed at every step, and all representations remain fixed. The final time is $B = 8$; $B = 0.5$ and $B = 2$ are prefixes of the same trajectory. The completed package contains 399 newly trained instances and 27 reused trajectories checked against every stored source-gradient recurrence, with no failed instances.

Predicted and trained scores use the same original scoring grid: the union of 65 equally spaced points on each interval $[0, b]$, $b \in \{0.5, 2, 8, 32\}$, restricted to the evaluated budget. Coefficient interpolation between Euler states is linear and is used only to evaluate the readout at these grid points. Composite trapezoidal integration is then applied to both paths. The $B = 32$ prediction is not used as a validated training result.

C.2 Datasets, Tasks, and Full Validation Results

The evaluation contains 20 declared dataset scopes and 142 label tasks: 17 single-task datasets, 40 CelebA attributes, 14 NIH ChestX-ray14 findings, and 71 original PTB-XL SCP statements. The three query rotations give 426 task–fold instances. PTB-XL superclass labels are not added as duplicate tasks. Record counts in Table C.2 use each dataset's evaluation unit: an image, ECG record, breast, activity window, graph, sentence pair, or comment.

The dataset sources include CelebA (Liu et al., 2015), PTB-XL (Wagner et al., 2020), NIH ChestX-ray14 (Wang et al., 2017), RSNA (Trivedi et al., 2026), and VinDr-Mammo (Nguyen et al., 2023). The activity datasets are UCI HAR (Reyes-Ortiz & Anguita, 2013), HHAR (Stisen et al., 2015), WISDM (Weiss, 2019), and PAMAP2 (Reiss & Stricker, 2012). The language scopes come from SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), QQP as included in GLUE (Wang et al., 2019), and CivilComments (Borkan et al., 2019). These sources describe the original datasets; the record units, task selection, contexts, and folds below specify our assessment.

Table C.2: Declared dataset scopes and effective record counts. The unit column defines the count; multilabel tasks share records.

Dataset	Tasks	Records	Unit
Waterbirds	1	11,745	image
CelebA	40	202,585	image
MetaShift	1	3,491	image
PACS	1	9,991	image
PTB-XL	71	21,799	ECG record
NIH ChestX-ray14	14	112,120	image
RSNA	1	23,826	breast
ViNDr-Mammo	1	10,001	breast
UCI HAR	1	10,299	activity window
HHAR	1	109,813	activity window
WISDM	1	120,149	activity window
PAMAP2	1	10,588	activity window
GOOD-CMNIST	1	29,400	graph
GOOD-Motif	1	12,600	graph
GOOD-HIV	1	14,454	graph
GOOD-SST2	1	27,270	graph
SNLI	1	559,209	sentence pair
MultiNLI	1	402,517	sentence pair
QQP	1	404,276	sentence pair
CivilComments	1	312,563	comment

Table C.3: Fixed representations, candidate contexts, and fold grouping. Each row uses the scope in Table C.2.

Dataset	Representation	Candidate context	Fold grouping
Waterbirds	CLIP ViT-B/32	Water/land background	Release-row folds
CelebA	CLIP ViT-B/32	Male/Young attributes; target excluded	Identity grouped
MetaShift	CLIP ViT-B/32	Indoor/outdoor	Release-row folds
PACS	DINOv2 ViT-S/14	Art/cartoon/photo/sketch	Image-hash grouped
PTB-XL	ECGDeli physiological features	Acquisition device	Patient grouped
NIH ChestX-ray14	CXR-CLIP	AP/PA view	Patient grouped
RSNA	Released EfficientNet-derived features	Any supplied marker, mole, scar, or clip	Patient grouped; breast unit

Continued on next page

Table C.3 (continued)

Dataset	Representation	Candidate context	Fold grouping
ViNDr-Mammo	Released EfficientNet-derived features	Any supplied marker, mole, scar, or clip	Patient grouped; breast unit
UCI HAR	Original 561-dimensional features	Subject identity	Within-subject temporal blocks
HHAR	Fixed statistical/spectral features	Physical device and sensor	Contiguous temporal thirds
WISDM	Fixed statistical/spectral features	Subject, device type, and sensor	Contiguous temporal thirds
PAMAP2	Fixed statistical/spectral features	Subject identity	Within-segment temporal thirds
GOOD-CMNIST	Fixed graph summaries	Synthetic digit color	Record-level folds
GOOD-Motif	Fixed graph summaries	Base-graph type	Record-level folds
GOOD-HIV	Fixed graph summaries	Source node-count deciles	Record-level folds
GOOD-SST2	Fixed graph summaries	Source node-count deciles	Record-level folds
SNLI	Fixed 512-dimensional signed feature hashing	Hypothesis-only negation; overlap; other	Exact-pair grouped
MultiNLI	Fixed 512-dimensional signed feature hashing	Hypothesis-only negation; overlap; other	Exact-pair grouped
QQP	Fixed 512-dimensional signed feature hashing	Token Jaccard at least 0.5	Exact-pair grouped
CivilComments	Fixed 512-dimensional signed feature hashing	Supplied identity indicator at least 0.5	Exact-text grouped

The language representations use signed feature hashing (Weinberger et al., 2009); the ECG features use ECGdeli (Pilia et al., 2021).

Waterbirds and CelebA retain the available aligned feature rows, with 43 and 14 missing feature records, respectively. MetaShift uses the released cat/dog scope. The mammography tasks average cached views within each breast; ViNDr evaluates the released binary abnormality label. The four GOOD entries use their declared concept/train releases. Language tasks use valid labeled records from the declared public training and validation files.

Fold assignments retain the source protocol. CelebA identities and biomedical patients are grouped; PACS groups byte-identical images. Waterbirds and MetaShift use release-row folds. Language folds keep normalized exact-pair or exact-text duplicates together without deleting their original record frequency. Where hash-based grouping is used, we compute a deterministic hash of `SK-v1:` followed by the group identifier. We convert its first eight hexadecimal digits to an integer and take the remainder modulo three to assign the fold. Time-series queries use stored temporal blocks within observed subjects or devices. The graph scopes use record-level cross-fitting. These splits define the reported diagnostic; they do not impose a common out-of-domain split across the twenty datasets.

Complete dataset and task results. Table C.4 gives all twenty paired dataset scores at $B = 8$. The dataset-equal Spearman correlation is 0.990977, and the mean absolute difference is 0.784060 score points. Figure C.1(a) expands this comparison to all 142 tasks. With each task weighted equally, the corresponding values are 0.997737 and 0.607016. This distinction keeps a dataset with many labels from silently receiving more weight in the main dataset-level result.

Figure C.2 resolves task variation within the three multilabel datasets. These contain 125 tasks; the other 17 datasets each contribute one task to the complete table.

Table C.4: Complete paired dataset scores at $B = 8$. Δ is trained minus predicted, in score points. Tasks receive equal weight within a dataset.

Dataset	Tasks	Predicted	Trained	Δ
PACS	1	46.226	43.735	-2.491
MetaShift	1	39.860	36.309	-3.551
Waterbirds	1	27.374	24.901	-2.473
UCI HAR	1	21.311	21.518	+0.207
CelebA	40	15.873	14.981	-0.892
HHAR	1	7.537	6.932	-0.604
NIH ChestX-ray14	14	6.991	6.934	-0.057
PAMAP2	1	6.815	6.748	-0.067
GOOD-HIV	1	5.584	5.300	-0.285
QQP	1	4.199	3.289	-0.909
PTB-XL	71	4.096	3.631	-0.465
CivilComments	1	3.806	3.622	-0.184
GOOD-SST2	1	3.524	1.871	-1.652
SNLI	1	3.484	2.495	-0.989
WISDM	1	1.878	1.782	-0.096
MultiNLI	1	0.956	0.700	-0.256
ViNDr-Mammo	1	0.652	0.428	-0.224
RSNA	1	0.488	0.485	-0.003
GOOD-Motif	1	-0.233	-0.299	-0.066
GOOD-CMNIST	1	-0.617	-0.826	-0.209

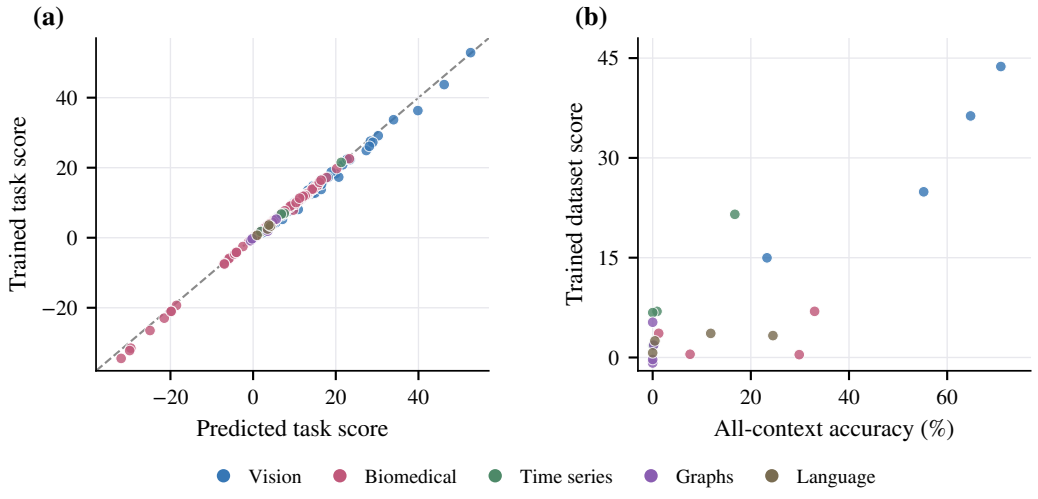


Figure C.1: **Task-level validation and classification coverage.** (a) All 142 paired task scores at $B = 8$; the dashed line is equality. (b) Trained dataset scores versus endpoint all-context accuracy, with one point per dataset. Colors identify domains in both panels.

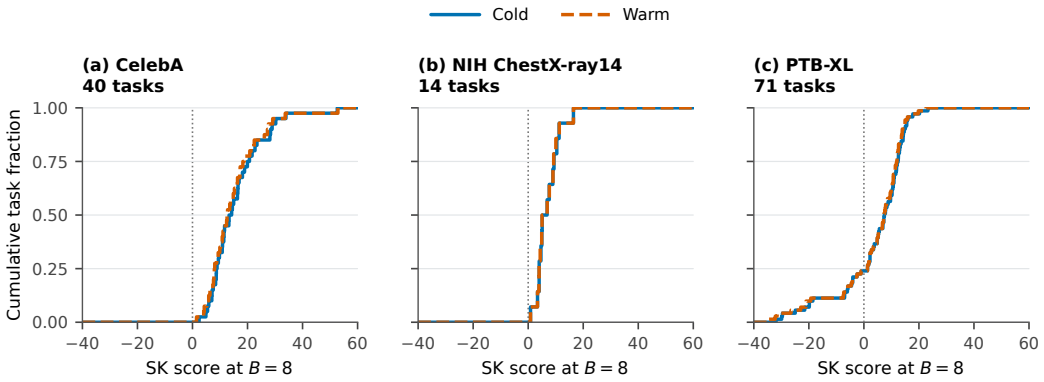


Figure C.2: **Within-dataset task variation.** Predicted and trained score distributions at $B = 8$ for all 40 CelebA attributes, 14 NIH findings, and 71 PTB-XL statements. Each distribution includes every declared task and retains negative scores.

Table C.5 lists every task, including negative scores, in the original dataset/task order. Each row combines its three query folds using (C.8). The column $n_{\min}$ is the smallest class count across all query records for that task. The signed difference is trained minus predicted score. These are task results, not 426 independent repeats.

Table C.5: All 142 original label tasks at $B = 8$. Each row pools three query folds; $n_{\min}$ is the smallest class count and Δ is trained minus predicted score.

No.	Dataset	Label task	$n_{\min}$	Predicted	Trained	Δ
1	Waterbirds	bird_category	2,663	27.374	24.901	-2.473
2	CelebA	5_o_Clock_Shadow	22,515	9.405	9.333	-0.072
3	CelebA	Arched_Eyebrows	54,089	14.238	12.551	-1.686
4	CelebA	Attractive	98,760	10.919	8.086	-2.833
5	CelebA	Bags_Under_Eyes	41,442	5.473	4.272	-1.201
6	CelebA	Bald	4,546	16.345	16.363	+0.018
7	CelebA	Bangs	30,708	30.268	29.144	-1.125
8	CelebA	Big_Lips	48,783	8.636	7.943	-0.692
9	CelebA	Big_Nose	47,510	7.109	5.248	-1.862
10	CelebA	Black_Hair	48,467	19.955	18.307	-1.648
11	CelebA	Blond_Hair	29,981	33.972	33.705	-0.267
12	CelebA	Blurry	10,311	7.556	7.402	-0.154
13	CelebA	Brown_Hair	41,569	16.943	16.264	-0.679
14	CelebA	Bushy_Eyebrows	28,800	14.972	12.746	-2.227
15	CelebA	Chubby	11,660	11.383	10.804	-0.579
16	CelebA	Double_Chin	9,455	10.816	10.275	-0.541
17	CelebA	Eyeglasses	13,193	28.373	27.587	-0.786
18	CelebA	Goatee	12,715	14.494	14.819	+0.325
19	CelebA	Gray_Hair	8,497	21.672	20.821	-0.851
20	CelebA	Heavy_Makeup	78,386	8.407	7.783	-0.623
21	CelebA	High_Cheekbones	92,181	20.713	17.313	-3.400
22	CelebA	Male	84,424	52.614	52.861	+0.248
23	CelebA	Mouth_Slightly_Open	97,932	23.438	22.286	-1.152
24	CelebA	Mustache	8,416	9.798	9.936	+0.138
25	CelebA	Narrow_Eyes	23,327	5.937	5.893	-0.044
26	CelebA	No_Beard	33,438	18.920	18.804	-0.117
27	CelebA	Oval_Face	57,563	7.030	6.349	-0.680
28	CelebA	Pale_Skin	8,701	8.931	8.810	-0.120
29	CelebA	Pointy_Nose	56,207	8.675	7.640	-1.036
30	CelebA	Receding_Hairline	16,161	2.401	1.462	-0.940
31	CelebA	Rosy_Cheeks	13,315	18.158	17.236	-0.922
32	CelebA	Sideburns	11,447	13.229	13.539	+0.310
33	CelebA	Smiling	97,661	28.956	27.239	-1.717
34	CelebA	Straight_Hair	42,219	13.290	12.396	-0.894
35	CelebA	Wavy_Hair	64,739	16.480	13.786	-2.694
36	CelebA	Wearing_Earrings	38,273	16.555	15.240	-1.314
37	CelebA	Wearing_Hat	9,818	22.725	22.446	-0.279
38	CelebA	Wearing_Lipstick	95,711	4.886	4.080	-0.806
39	CelebA	Wearing_Necklace	24,911	11.652	11.088	-0.564
40	CelebA	Wearing_Necktie	14,729	11.470	11.271	-0.199
41	CelebA	Young	45,859	28.130	26.106	-2.024
42	MetaShift	cat_dog	1,542	39.860	36.309	-3.551
43	PACS	object_category	943	46.226	43.735	-2.491
44	PTB-XL	NDT	1,825	7.457	7.252	-0.205
45	PTB-XL	NST_	767	8.388	8.230	-0.158
46	PTB-XL	DIG	181	14.215	13.634	-0.581
47	PTB-XL	LNGQT	117	-6.004	-6.231	-0.227

Continued on next page

Table C.5 (continued)

No.	Dataset	Label task	$n_{\min}$	Predicted	Trained	Δ
48	PTB-XL	NORM	9,514	9.726	7.893	-1.834
49	PTB-XL	IMI	2,676	5.608	5.371	-0.237
50	PTB-XL	ASMI	2,357	13.548	12.952	-0.596
51	PTB-XL	LVH	2,132	12.909	12.395	-0.514
52	PTB-XL	LAFB	1,623	15.070	14.813	-0.257
53	PTB-XL	ISC_	1,272	15.623	14.955	-0.668
54	PTB-XL	IRBBB	1,118	6.951	6.688	-0.262
55	PTB-XL	1AVB	793	10.533	10.341	-0.193
56	PTB-XL	IVCD	787	7.267	7.132	-0.134
57	PTB-XL	ISCAL	659	13.194	12.591	-0.603
58	PTB-XL	CRBBB	541	17.766	17.197	-0.569
59	PTB-XL	CLBBB	536	20.220	19.758	-0.463
60	PTB-XL	ILMI	478	10.955	10.604	-0.351
61	PTB-XL	LAO/LAE	426	8.803	8.585	-0.217
62	PTB-XL	AMI	353	-1.009	-1.018	-0.008
63	PTB-XL	ALMI	288	10.689	10.265	-0.424
64	PTB-XL	ISCIN	218	2.231	2.219	-0.012
65	PTB-XL	INJAS	214	5.360	5.274	-0.086
66	PTB-XL	LMI	201	1.309	1.293	-0.016
67	PTB-XL	ISCIL	179	6.883	6.742	-0.142
68	PTB-XL	LPFB	177	-5.784	-5.886	-0.102
69	PTB-XL	ISCAS	169	10.232	9.884	-0.349
70	PTB-XL	INJAL	145	-4.465	-4.595	-0.129
71	PTB-XL	ISCLA	140	7.913	7.744	-0.169
72	PTB-XL	RVH	126	5.233	5.188	-0.046
73	PTB-XL	ANEUR	104	7.823	7.589	-0.234
74	PTB-XL	RAO/RAE	99	2.217	2.139	-0.079
75	PTB-XL	EL	96	11.834	11.398	-0.436
76	PTB-XL	WPW	79	1.345	1.329	-0.017
77	PTB-XL	ILBBB	77	3.519	3.373	-0.146
78	PTB-XL	IPLMI	51	4.905	4.830	-0.075
79	PTB-XL	ISCAN	44	-2.467	-2.511	-0.044
80	PTB-XL	IPMI	33	-4.071	-4.130	-0.059
81	PTB-XL	SEHYP	29	-18.538	-19.317	-0.779
82	PTB-XL	INJIN	18	-19.793	-21.051	-1.258
83	PTB-XL	INJLA	17	-29.593	-31.533	-1.940
84	PTB-XL	PMI	17	-21.497	-22.966	-1.469
85	PTB-XL	3AVB	16	-24.942	-26.497	-1.555
86	PTB-XL	INJIL	15	-29.872	-32.224	-2.352
87	PTB-XL	2AVB	14	-31.928	-34.462	-2.534
88	PTB-XL	ABQRS	3,327	7.249	7.129	-0.121
89	PTB-XL	PVC	1,143	15.556	14.923	-0.633
90	PTB-XL	STD_	1,009	9.666	9.487	-0.180
91	PTB-XL	VCLVH	875	14.367	13.666	-0.701
92	PTB-XL	QWAVE	548	10.918	10.693	-0.225
93	PTB-XL	LOWT	438	12.424	12.086	-0.338
94	PTB-XL	NT_	423	11.625	11.311	-0.314
95	PTB-XL	PAC	398	11.914	11.424	-0.490
96	PTB-XL	LPR	340	13.018	12.586	-0.433
97	PTB-XL	INVT	294	12.518	12.114	-0.404
98	PTB-XL	LVOLT	182	14.663	13.954	-0.708
99	PTB-XL	HVOLT	62	11.062	10.664	-0.398
100	PTB-XL	TAB_	35	4.843	4.768	-0.075

Continued on next page

Table C.5 (continued)

No.	Dataset	Label task	$n_{\min}$	Predicted	Trained	Δ
101	PTB-XL	STE_	28	2.744	2.710	-0.033
102	PTB-XL	PRC(S)	10	-7.024	-7.338	-0.314
103	PTB-XL	SR	5,051	10.253	9.565	-0.688
104	PTB-XL	AFIB	1,514	23.287	22.575	-0.712
105	PTB-XL	STACH	826	14.292	13.873	-0.419
106	PTB-XL	SARRH	772	3.765	3.725	-0.040
107	PTB-XL	SBRAD	637	2.203	1.985	-0.217
108	PTB-XL	PACE	294	16.024	15.805	-0.218
109	PTB-XL	SVARR	157	12.583	12.008	-0.575
110	PTB-XL	BIGU	82	12.106	11.819	-0.287
111	PTB-XL	AFLT	73	-4.039	-4.194	-0.155
112	PTB-XL	SVTAC	27	1.814	1.737	-0.077
113	PTB-XL	PSVT	24	-19.846	-21.025	-1.179
114	PTB-XL	TRIGU	20	-6.953	-7.508	-0.555
115	NIH CXR14	Atelectasis	11,535	9.378	9.267	-0.111
116	NIH CXR14	Cardiomegaly	2,772	8.966	9.006	+0.041
117	NIH CXR14	Consolidation	4,667	4.002	3.903	-0.099
118	NIH CXR14	Edema	2,303	10.401	10.049	-0.352
119	NIH CXR14	Effusion	13,307	16.444	16.447	+0.003
120	NIH CXR14	Emphysema	2,516	7.651	7.691	+0.040
121	NIH CXR14	Fibrosis	1,686	5.090	5.039	-0.051
122	NIH CXR14	Hernia	227	4.961	4.910	-0.051
123	NIH CXR14	Infiltration	19,870	4.637	4.357	-0.280
124	NIH CXR14	Mass	5,746	6.883	6.909	+0.026
125	NIH CXR14	Nodule	6,323	3.393	3.381	-0.013
126	NIH CXR14	Pleural_Thickening	3,385	3.979	3.967	-0.012
127	NIH CXR14	Pneumonia	1,353	0.853	0.849	-0.003
128	NIH CXR14	Pneumothorax	5,298	11.239	11.303	+0.063
129	RSNA	malignancy	492	0.488	0.485	-0.003
130	ViNDr	released_abnormality	494	0.652	0.428	-0.224
131	UCI HAR	original_multiclass_task	1,406	21.311	21.518	+0.207
132	HHAR	original_multiclass_task	15,441	7.537	6.932	-0.604
133	WISDM	original_multiclass_task	6,413	1.878	1.782	-0.096
134	PAMAP2	original_multiclass_task	182	6.815	6.748	-0.067
135	GOOD-CMNIST	original_multiclass_task	2,645	-0.617	-0.826	-0.209
136	GOOD-Motif	original_multiclass_task	4,099	-0.233	-0.299	-0.066
137	GOOD-HIV	original_multiclass_task	486	5.584	5.300	-0.285
138	GOOD-SST2	original_multiclass_task	12,075	3.524	1.871	-1.652
139	SNLI	original_multiclass_task	185,999	3.484	2.495	-0.989
140	MultiNLI	original_multiclass_task	134,023	0.956	0.700	-0.256
141	QQP	original_multiclass_task	149,263	4.199	3.289	-0.909
142	CivilComments	toxicity	35,377	3.806	3.622	-0.184

C.3 Budget Dependence and Numerical Checks

Figure C.3 and Table C.6 compare the same assessment at three budgets. Dataset rankings agree exactly at $B = 0.5$ and $B = 2$, and remain closely aligned at $B = 8$. The score discrepancy increases as the trained path moves farther from the initial quadratic model.

For a complementary classification readout, a query is counted correct only if its true-class logit is strictly larger than every competing class under every candidate context. Endpoint all-context accuracy applies the same class/task balancing as the score. Figure C.1(b) shows why both quantities are useful: GOOD-HIV has a positive trained score of 5.300 while its endpoint all-context accuracy remains zero. Loss can improve before the classification boundary is crossed.

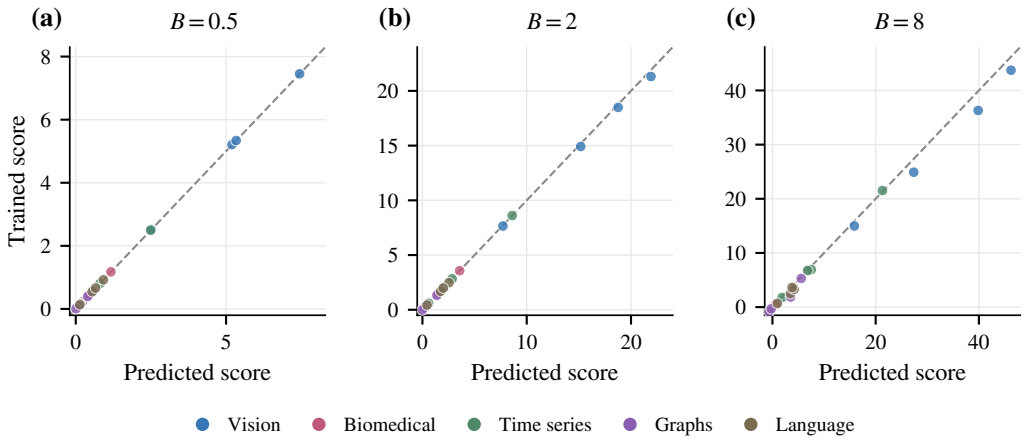


Figure C.3: **Validation across training budgets.** Each panel compares predicted and trained scores for all 20 datasets at the indicated budget. Dashed lines mark equality. The panels use different axis ranges; budgets share the same training trajectories.

Table C.6: Paired agreement across budgets. Each row weights its stated units equally. Error is trained minus predicted score.

B	Equal-weight unit	Spearman	MAE	Max. $\lvert\Delta\rvert$	Mean Δ
0.5	20 datasets	1.0000	0.0027	0.0137	+0.0027
0.5	142 tasks	1.0000	0.0055	0.0312	+0.0035
2	20 datasets	1.0000	0.0745	0.6041	-0.0737
2	142 tasks	0.9998	0.0426	0.6041	-0.0379
8	20 datasets	0.9910	0.7841	3.5508	-0.7634
8	142 tasks	0.9977	0.6070	3.5508	-0.5870

Figure C.4 follows each dataset across the three validated budgets, complementing the aggregate agreement statistics.

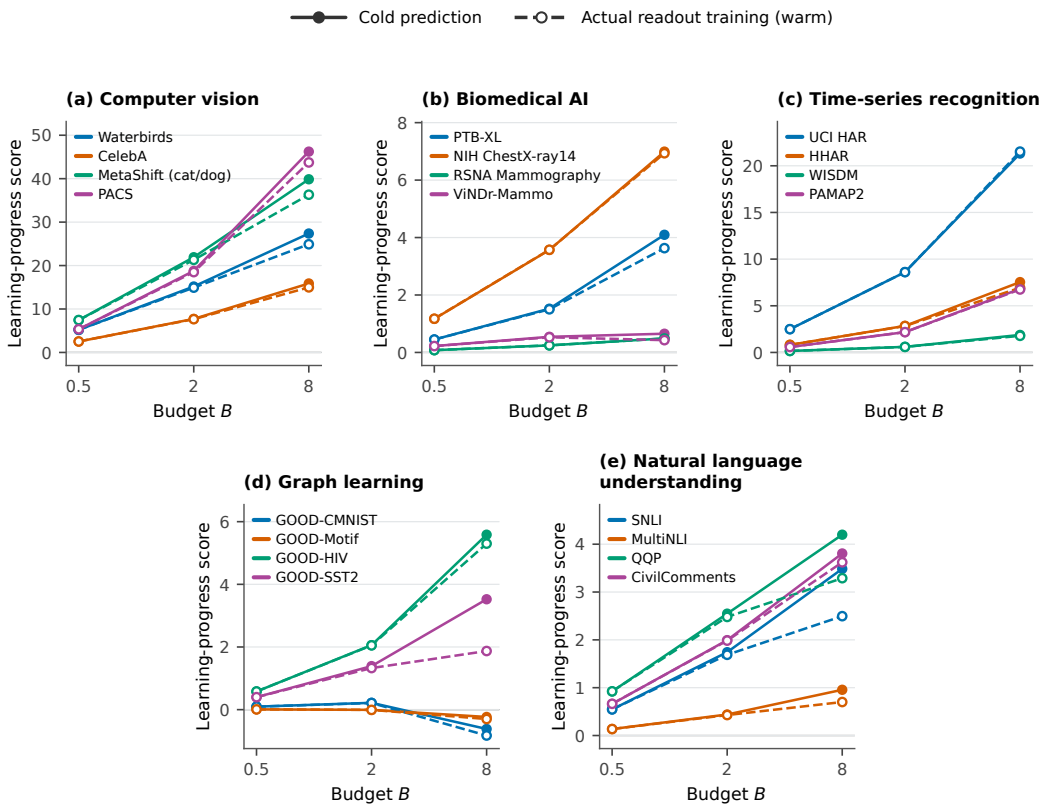


Figure C.4: **Dataset scores across budgets.** Predicted and trained scores for each dataset at $B = 0.5, 2, 8$, grouped by domain. The shorter budgets are prefixes of the same trained trajectory; connecting lines join the reported budgets only.

Numerical sensitivity. The numerical checks cover nine selected tasks and all their query folds. Selection includes the largest discrepancies, near-zero and negative scores, and the high-score end. The checks halve the Euler step to 0.0125, compare with DOP853 using relative tolerance 10^{-10} and absolute tolerance 10^{-12}, and double the number of scoring-grid intervals. Table C.7 reports the task-level results at $B = 8$.

Across the checked tasks and budgets, the largest score changes are 0.039160 points under step halving and 0.078370 points against the adaptive solution. Grid refinement changes trained scores by at most 0.001189 points and predicted scores by at most 0.001324 points. At $B = 8$, the largest checked Euler–adaptive difference is 0.064578 points. These are measured sensitivities of the stated checks; the primary paired scores remain unchanged.

The fixed task logits have zero source mean within each context. At context-only initialization, the softmax probabilities and context logits are constant within each such group. Consequently the aggregate initial cross-Hessian between the task and context coefficients is zero, up to numerical precision. Separate sample-group contributions can still be nonzero and cancel in their sum. The SK validation tests this specified readout prediction. The evolving majority-feedback mechanism in Sections 3–4 is established by the preceding analysis and channel interventions.

Code and data assets. The machine-readable result tables retain all 142 task scores and 426 task–fold rows, including per-class scores and counts for exact fold pooling. The scope and task rosters specify the record units, label sets, representations, and contexts. The protocol lock records

Table C.7: Numerical checks at $B = 8$, in score points. Half step and adaptive subtract the primary Euler score; grid refinement subtracts the adaptive score on the original grid.

Dataset	Task	Primary	Half step	Adaptive	Grid
MetaShift	cat_dog	36.3093	-0.02571	-0.05142	+0.00098
CelebA	High_Cheekbones	17.3130	-0.01031	-0.02064	+0.00055
CelebA	Attractive	8.0859	-0.00583	-0.01166	+0.00036
GOOD-Motif	original_multiclass_task	-0.2993	+0.00013	+0.00026	+0.00001
RSNA	malignancy	0.4852	-0.00043	-0.00086	+0.00001
GOOD-CMNIST	original_multiclass_task	-0.8259	+0.00021	+0.00041	+0.00008
PTB-XL	2AVB	-34.4624	+0.03232	+0.06458	-0.00090
PTB-XL	INJIL	-32.2241	+0.02980	+0.05953	-0.00082
PACS	object_category	43.7348	-0.02325	-0.04650	+0.00119

solver settings and hashes of the original predicted scores. The code files specify source construction, trajectory prediction, actual-gradient validation, aggregation, and numerical checks.

Replaying an individual training trajectory requires its fixed source and query logits, context prototypes, labels, and initial coefficients. Reconstructing these logits additionally requires the pinned feature assets and record/fold assignments. The result tables support exact aggregation checks; trajectory replay uses the separate fixed-logit cache package.