THE HIDING PROBLEM: WHEN UNREAD FEATURES DO AND DO NOT DRIVE HARMFUL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When a classifier adapts to new data without labels and its loss rises, the loss alone does not say why: the model may have acquired a harmful dependence on a component of its representation that it did not use before, or it may have become more confident in the mistakes it already made. We ask which of these happens when a later update reaches components that the current predictor does not read (the Hiding Problem). An exact binary construction (Theorem 1) shows that the first is possible: two predictors with identical calibrated predictions and zero population IRMv1 penalty take the same Euclidean entropy step, and the one whose unread coordinate is stretched climbs from 75% to 90% accuracy on the adaptation law and falls to 10% under a reversed nuisance association, while the other stays at 75% under both. Deliberate amplification produces the same behavior in synthetic and frozen-feature Camelyon17 representations. On six native ResNet-50 checkpoints for a naturally shifted hospital, three controls point to the second explanation instead: a label-free affine map of the initial score closely reproduces over 99% of the head-only loss increase, a selected unread direction keeps no mean advantage over random deletions once endpoint movement is matched, and the same parameters trained with 320 labels raise the ranking in every checkpoint, whereas label-free updates that reach the representation lower it in the three checkpoints that missed the most positives (a post hoc split). A preregistered replication on Waterbirds-95 with six fresh models reproduces the head-only account and the sign pattern of the representation arms; the widespread one-class collapse at a fixed learning rate did not replicate. Identical predictions do not imply identical responses to learning. An update can recruit an unused feature, increase confidence in existing mistakes, or appear less harmful simply because it changes the predictions less. Our results make these distinctions consequential: evaluating continued learning requires testing what an update changes and establishing which changes account for its harm; the checks here begin that test and lay the groundwork for safeguards that can be tested the same way.

1 INTRODUCTION

When a classifier adapts without labels and its loss rises, the loss does not say why. The model may have acquired a harmful dependence on a component of its representation that it did not use before, or it may have become more confident in the mistakes it already made, and a restriction that seems to protect it may only have made the update smaller. For a binary linear head the logit depends on one projection of the feature vector, so components orthogonal to the head do not affect it, and training need not remove them; the next label-free update can reach them. We call the retention of components the current predictor does not read the Hiding Problem. Accuracy, calibration or an invariance penalty measured on the current predictor describe that predictor, while the next update depends on components those measurements never saw (Theorem 1 makes this exact). Two questions follow: what can a later update make the predictor use, and what does that do to task risk? The two come apart: an update can avoid one-class collapse and still raise the loss, or reduce a nuisance response without adapting usefully. So we ask a narrower question: when we restrict an initially unread component and control how far the predictions move on a declared reference distribution, does the target risk change, and which diagnostics separate that effect from everything else the update changed?

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

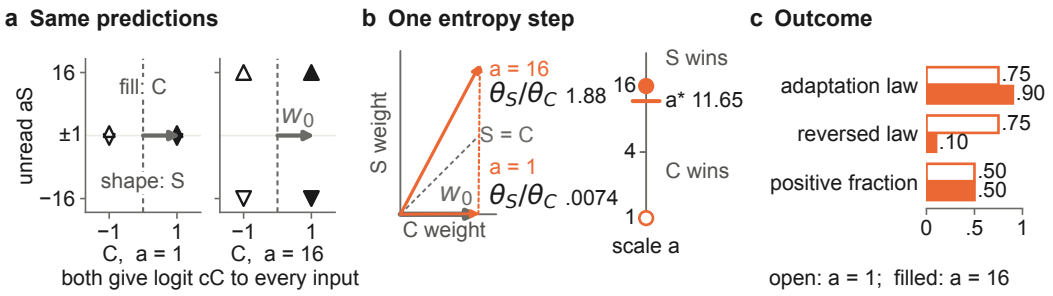


Figure 1: The construction of Section 3 ($p = .75$, adaptation agreement $r = .9$, reversed agreement .1, $\eta = .1$). (a) The four inputs $(C, S) \in \{\pm 1\}^2$ in feature coordinates $\phi_a = (C, aS)$ at scale $a = 1$ and $a = 16$ on one vertical scale, so the two markers at $a = 1$ overlap; fill marks C and shape marks S. Both models have the head $w_0 = (c, 0)$ (gray arrow) and the boundary $C = 0$ (dashed), so they give every input the same logit cC. (b) One population entropy step: the head before (gray) and after (orange) in raw coefficients $(\theta_C, \theta_S) = (c(1 + \eta h), \eta a^2 ch\rho)$, the update at $a = 16$ dotted; the ratio θ_S/θ_C is .0074 at $a = 1$ and 1.88 at $a = 16$ and passes 1 at $a^* = 11.65$ (logarithmic scale axis). (c) Accuracy after the step under the current adaptation law and the reversed law, and the positive fraction (Table 16).

Our empirical finding is negative. We went looking for unread-direction recruitment behind the harm in native Camelyon17 checkpoints and could not attribute the harm to it: an affine change of the score the head already had closely reproduced the head-only loss rise, and once the representation was allowed to adapt the checkpoints drifted toward the majority class. In the label-free updates we tested, recruitment of a selected unread direction showed up only where we built it in, in the exact construction and under deliberate amplification. Concretely, in the synthetic pilot and on six native checkpoints a label-free affine function of the initial score reproduces ordinary head-only entropy harm, whereas in the amplified settings the same update recruits the unread nuisance. Restricting a selected unread direction did not reduce the head-only harm, and the benefit it seemed to give natively disappeared once we matched movement (the representation arms received no such intervention). When normalization or all parameters of a native ResNet-50 adapted from pixels, checkpoints drifted toward their majority class, three of six at first-step-matched rates and five and six of six at a common rate, while the same parameters trained with 320 labels improved ranking in every checkpoint. These findings pin down what the measured interventions establish about component access, which common diagnostics fail to order risk here, and what is still unestablished about useful learning.

We study fixed-feature entropy adaptation because its update can be analyzed exactly: a change along a direction in the null space of the initial head preserves every initial logit and every statistic restricted to those logits, labels and environment assignments, yet changes the feature inner products a subsequent head update uses (Figure 1).

The mathematical ingredients have established explanations. Chen et al. prove that entropy minimization avoids a spurious feature that is independent of the label in the target domain, given a decently accurate source classifier (Chen et al., 2020); our construction keeps the nuisance associated with the label in the adaptation stream, where the same update recruits it. DeYO analyzes entropy updates using feature inner products and harmful logit changes (Lee et al., 2024), and optimizer behavior under reparameterization is a classical concern (Martens, 2020). Unlabeled adaptation beyond entropy includes test-time training, which updates a shared encoder through a self-supervised auxiliary loss (Sun et al., 2020), and continual test-time adaptation (Wang et al., 2022); fine-tuning can distort pretrained features, with conditional out-of-distribution tradeoffs (Kumar et al., 2022), and geometric and statistical skews can induce out-of-distribution failures of ERM (Nagarajan et al., 2021); the three-term split of Section 6 separates the head-side and representation-side parts of label-free updates.

Sections 4 and 5 test component access on retained synthetic and frozen-feature Camelyon17 (Koh et al., 2021) representations, including a pilot on four retained checkpoints that blocked a selected head-orthogonal nuisance-response direction at matched neutral-reference movement; Section 6

compares head-only and representation-reaching updates on two hospitals, and only its hospital-2 head study restricts an initially unread direction.

Relation to invariance and representation reuse. IRM encourages agreement through a particular prediction-level objective (Arjovsky et al., 2019). Its known limitations include failures to identify the desired invariant predictor (Rosenfeld et al., 2021; Kamath et al., 2021) and sensitivity to penalty formulation and conditioning (Khezeli et al., 2021). Information outside a current readout can also be useful, since last-layer retraining can improve robustness (Kirichenko et al., 2023); a task-relative nuisance is therefore not every unread component, and a practical restriction would need to preserve an independently demonstrated opportunity for beneficial adaptation.

Relation to concept erasure and test-time adaptation. Concept-erasure methods remove a concept's linear predictability from a representation so that downstream linear probes cannot recover it. INLP (Ravfogel et al., 2020), RLACE (Ravfogel et al., 2022), LEACE (Belrose et al., 2023) and JSE (Holstege et al., 2024) do so by iterative, adversarial, closed-form and subspace-specific projections. Our object is different: we restrict what a specified later update can access, not what a probe can predict, and in our tests such a restriction does not rescue adaptation. Tent adapts a trained network at test time by entropy minimization over normalization statistics and channel-wise affine parameters (Wang et al., 2021); our native study follows this recipe as a fixed development recipe, not as a reproduction of Tent's results. Entropy-based test-time adaptation is known to degrade under label shift, small or non-i.i.d. batches and long horizons (Niu et al., 2023; Zhao et al., 2023); we do not claim these failures are new, but decompose where the harm lives and test whether restricting unread directions changes it. DFR retrains only the last layer, with labels (Kirichenko et al., 2023); our head experiments update the same layer from unlabeled data by entropy descent.

2 Predictors, updates, and unread components

Let X be a raw input, $Y \in \{0,1\}$ a label, and e an environment identifier. A fixed representation $\phi : \mathcal{X} \to \mathbb{R}^d$ and head $w \in \mathbb{R}^d$ produce logit $f_w(x) = w^\top \phi(x)$. Binary logistic loss is $\ell(z,y) = \log(1+\exp z) - yz$. For a joint evaluation law Q, write $\mathcal{R}_Q(f) = \mathbb{E}_Q \ell(f(X), Y)$. Accuracy is reported separately, since neither an accuracy change nor a sensitivity change alone determines logistic-risk change.

The update is part of the system being studied. Unless stated otherwise, it is ordinary Euclidean descent of mean binary predictive entropy on a specified unlabeled stream, with a frozen representation and a freely updated head. With $p(z) = \sigma(z)$,

$$H(z) = -p(z)\log p(z) - (1-p(z))\log(1-p(z)), \qquad H'(z) = -z\sigma(z)(1-\sigma(z)). \tag{1}$$

Identical numerical update prescriptions in two representations need not produce identical function updates. We will name both the trainable parameter subset and the metric.

Definition 1 (Prediction equivalence and an initially unread component)**.** *Two representation–head pairs are initially prediction equivalent if* $w_0^\top \phi(x) = \widetilde{w}_0^\top \widetilde{\phi}(x)$ *for every* x*. A unit direction* u *is initially unread by* w_0 *when* $w_0^\top u = 0$*. Its later logit contribution is*

$$f_t^u(x) = (w_t^\top u)(u^\top \phi(x)). \tag{2}$$

Unreadness is generic, since every direction orthogonal to the head is unread, and it is a parameter-space property, not an information one: $u^\top \phi(x)$ can be correlated with, or equal to, the current logit. The question is which unread directions a given update recruits. An unread component is not an entire semantic variable: a nuisance can affect several representation directions, including directions the initial head already uses. The synthetic experiments select the component of a controlled input-perturbation response that is orthogonal to w_0; that component is exactly unread, while the entire nuisance's initial influence is generally nonzero. In Camelyon17 the selected direction is associated with class-conditional hospital differences, not a causally identified hospital effect. Appendix M states what a statistic restricted to initial logits, labels and environment assignments can observe, and Appendix K gives a fixed-readout family that changes response magnitude without changing current predictive use.

3 AN EXACT BALANCED FAILURE UNDER A SPECIFIED UPDATE

3.1 GEOMETRIC HIDING: AN EXACT BALANCED FAILURE

Let C be uniform on $\{-1, 1\}$. Independently draw signs B_c, B_s with $\mathbb{P}(B_c = 1) = p \in (1/2, 1)$ and $\mathbb{P}(B_s = 1) = r \in (0, 1)$. Define $T = CB_c$, $Y = (T + 1)/2$, and $S = TB_s$. Varying r changes the nuisance association while preserving the joint law of (C, Y). For a feature scale $a > 0$, set

$$\phi_a = (C, aS), \qquad w_0 = (c, 0), \qquad c = \log \frac{p}{1-p}. \tag{3}$$

All scales have initial logit cC, accuracy p, and the same logistic risk. The representation retains S, but the current head does not use it.

Write $h = p(1-p)$ and $\rho = (2p-1)(2r-1) = \mathbb{E}_r[CS]$. One population entropy step uses features but no labels:

$$w_1 = w_0 - \eta \mathbb{E}_r[H'(w_0^\top \phi_a)\phi_a], \qquad \eta > 0.$$

Let θ_C, θ_S denote its effective coefficients in the raw coordinates (C, S).

Theorem 1 (Geometric hiding). *For the law and features of (3), every scale $a > 0$ satisfies the following. (i) Calibration: for every environment parameter r the population scalar-dummy gradient $G_r = \partial_v \mathbb{E}_r \ell(vcC, Y)|_{v=1} = \mathbb{E}_r[(\sigma(cC) - Y)cC]$ is zero, so any finite sum of population IRMv1 squared dummy gradients is zero. (ii) Hidden energy: $\mathbb{E}\|\phi_a\|^2 = 1 + a^2$, while the initial logits cC, the labels and the environment assignments do not depend on a. (iii) Recruitment: the updated coefficients are*

$$\theta_C = c(1 + \eta h), \qquad \theta_S = \eta a^2 ch\rho, \tag{4}$$

and if $\rho > 0$ and $a^2 > (1 + \eta h)/(\eta h \rho)$ the classifier follows S on every input, its accuracy in an evaluation environment with nuisance agreement r_e is r_e, and, away from zero-logit ties, its positive prediction probability is exactly $1/2$.

Theorem 1 is elementary and claims no new kernel identity or nonidentifiability result; its value is an exact, balanced counterexample: (i) and (ii) leave the unread scale free, and (iii) makes the outcome of one identical update depend on it. It is a calibrated population statement, not an exactly zero finite-sample estimator or a causal-erasure or future-risk certificate; Appendix L gives the general finite calibration root and Appendix J the proof.

With $p = .75$, $r = .9$ and $\eta = .1$ the threshold is $a > 11.6548$: scale 1 retains the stable decision and scale 16 follows S (Table 16 in Appendix J: reversed risks .564840 and 1.627561 against .562335 initially), harmful recruitment without one-class collapse (Appendix J records the example's input support).

The scale-16 step improves accuracy, and in this particular step logistic risk, on the current adaptation law (its coefficients approach that law's Bayes logit $cC + \log \frac{r}{1-r} S$) and hurts the later reversal: a declared current/future tradeoff, not uniform degradation under every use, and no guarantee about arbitrary future label associations. Under either evaluation law the two systems share the same optimal action, so the construction demonstrates different risks under a specified update rather than an identifiability obstruction to choosing an action; Appendix M states what an auditor restricted to initial outputs can and cannot decide about the pair.

The classical update-kernel explanation. Under $\phi' = A\phi$, $w' = A^{-\top} w$, predictions initially agree but Euclidean head descent induces the original-coordinate update preconditioner $A^\top A$, the inverse of the parameter-space metric and the inner product of the induced feature kernel $\phi(x)^\top A^\top A \phi(x')$ (Appendices M and N). The square in (4) is the corresponding one-direction effect: one scale enters the gradient and another enters the resulting logit; in general, rescaling any unread coordinate by a while preserving predictions multiplies its one-step logit recruitment by a^2.

3.2 CENTERING AND ACCESS RESTRICTIONS

Mean centering uses $\phi_c = \phi - \mu$ with the compensating initial bias $b_0 = w_0^\top \mu$, preserving $w_0^\top \phi_c + b_0 = f_0$ while deliberately changing subsequent dynamics (Appendix O). For a unit $u \perp w_0$, projecting each head gradient as $g \leftarrow (I - uu^\top)g$ keeps $w_t^\top u = 0$ and blocks the contribution

(2); it is a valid privileged intervention when the head moves, whereas before a *fixed* head the same projection is an exact no-op (Appendix Q). Full second-moment preconditioning, a standard coordinate-invariance control, is treated in Appendix N.

4 RETAINED LEARNED-CHECKPOINT STUDIES

We retain four learned-checkpoint studies, Studies A–D (Appendix A). Study A is the synthetic scale study and Study B the frozen-feature Camelyon17 study. Study C runs long natural-scale streams with the collapse removed and also varies learning rate and batch size; Study D is a synthetic follow-up with separate direction-estimation samples and mean/intercept controls. In the central comparisons, original-scale updates moved toward predicting one class while the selected nuisance sensitivity barely changed. The prediction-preserving rescaling to scale 16 raised that sensitivity and lowered reversed accuracy (Table 1) and worst-hospital accuracy (median of the per-checkpoint minimum .8475 to .5569; Appendix A, Table 2). Our strongest present conclusion is that exact initial equivalence can coexist with distinct update outcomes in learned representations; at least two patterns we measured require different controls.

5 MATCHED COMPONENT-ACCESS DIAGNOSTICS

We tested whether restricting access to an independently selected, initially unread nuisance-response component reduces terminal harm more than a same-rank random restriction when reference prediction movement is matched. This diagnostic comparison of update directions uses known synthetic nuisance structure and is not a deployment verifier.

Fixed primary comparison. The primary setting used original-scale, uncentered synthetic features, step size .01 and 30 batches of 512 examples (ten each from the source-like, neutral and reversed nuisance laws of the retained synthetic generator, $\gamma = 1.8$, 0 and -1.5; Appendix A). The endpoint was reversed-law logistic risk at step 30. We estimated a rank-one direction from an independent sample of paired one-training-standard-deviation S interventions and projected it off w_0. We chose four random restrictions, orthogonal to both the initial head and the selected direction, independently of outcomes. The recorded contrast was $\Delta = R^{\text{targeted}}_{\text{target}} - \frac{1}{4}\sum_{j=1}^{4} R^{\text{random},j}_{\text{target}}$ (negative favors targeting). In the four checkpoints, the selected direction captured 16.6–27.6% of the head-orthogonal finite-difference response energy.

Matching movement. Ordinary original-scale descent defined an anchored neutral-reference RMS schedule B_t. After each candidate restricted update we rescaled the entire displacement from w_0 to match B_t on that reference; this matches a scalar functional budget and leaves the update direction free. The frozen comparator stayed frozen. We computed budgets separately for raw and centered representations.

The pilot used the two retained IRM settings and initialization seeds 42 and 43, sharing the original source-training data. Appendix C gives the sample sizes, execution checks and stress controls.

The primary restriction did not provide consistent rescue. Table 4 (Appendix C) reports every checkpoint. Three contrasts were positive and one negative (equal-weight mean $+0.000622740$ nats), so the targeted restriction showed no consistent advantage over the declared random controls. In all four checkpoints, ordinary and targeted primary paths both produced one-class target predictions. Ordinary mean target risk rose from .798979 frozen to 1.745193. Accuracy increased from 45.03% to 50.90%, but under the prespecified constant-baseline comparison this was not useful adaptation.

The dominant change was nearly constant across inputs. On ordinary natural-scale paths, absolute mean reference-logit change divided by RMS change was .99969–.99975, and the affine fits had slopes 1.011–1.017 with large signed intercept shifts. Blocking the selected component succeeded numerically, yet this offset drift and collapse still occurred. The negative result concerns this selected rank-one intervention and these checkpoints, not every retained nuisance component.

Centering reduced movement. Centering with a compensating fixed bias prevented collapse in the tested natural-scale paths. Mean reference movement fell from 3.178554 to .030199, so the apparent protection came with much less learning (Table 3 in Appendix C). None of the 192 adaptive terminal reversed-target paths or 32 frozen bookkeeping paths met the prespecified useful-accuracy criterion, and no adaptive path met the joint criterion. These are four reused checkpoints, not independent replications.

Common budget, secondary restriction and affine references. A follow-up crossed geometry with a common movement budget (Appendix D). At the large raw budget, centering removed collapse but increased target risk in every checkpoint pair, by $+.144993$ nats on average (a claim about this diagnostic update, not native centering). Within these centered paths the selected restriction lowered risk relative to random restrictions by .143092 nats, yet its mean risk of 1.747783 stayed far above the frozen .798979. Saved affine references of the original score reproduce a large loss increase but not these contrasts ($+.031675$ of the extra centering loss between affine references, $+.113317$ between signed remainders).

6 Native-image adaptation

The head experiments do not establish that an unread-component restriction preserves beneficial native adaptation. We therefore ran a separately specified feasibility study on image inputs with existing ResNet-50 BatchNorm (BN) affine parameters and asked whether entropy learning improves hospital-1 query loss beyond both saved-statistics and current-batch-statistics frozen controls. We used Tent (Wang et al., 2021) as a fixed development recipe, not as a reproduction of its reported results or a semantic recruitment test.

Fixed recipe and panel. Two retained source checkpoints (seeds 42 and 43) used native 96×96 RGB images with ImageNet normalization and no augmentation. A deterministic, label-blind split of eight hospital-1 development patients gave four adaptation patients (80 patches each) and four query patients (128 each). The protocol fixed ten entropy-only Adam updates at rate 10^{-3} on the existing BN scale/offset tensors only (Appendix E).

Current-batch statistics: separate switch and learning effects. The source-frozen control kept checkpoint running statistics. The statistics-only control and the affine learner used each current batch's statistics during adaptation and query scoring without accumulating them, so their step-zero query outputs equaled each other but not the source-frozen predictor's. The primary feasibility convention required lower step-10 loss than both controls and accuracy no more than one percentage point below either, separately for each checkpoint.

The statistics switch increased mean loss from .232994 to 5.163395 before learning. The affine updates then added 1.402433 and .994757 nats and reduced accuracy relative to both controls (Table 8; accuracies in Table 6), so neither checkpoint satisfied the declared convention. Final positive-prediction fractions were .560547 and .556641, so this failure did not require one-class collapse. Query entropy fell in both seeds while proper loss and accuracy worsened. The hospital-1 panels differ sharply in tumor fraction (179/320 for adaptation, 29/512 for query), so query batch statistics differ from the class-balanced source-training rows. Current-batch outputs also depend on batch companions. At fixed weights, replacing an anchor's 31 companion images changed current-batch-statistics logits by an RMS of about 9 and flipped 2 to 3 of 32 signs; saved-statistics outputs were exactly invariant (Appendix E).

Retaining saved statistics does not supply a consistent remedy. A separately frozen diagnostic with the same recipe and saved BN statistics fixed was adverse for both checkpoints, $+.060795$ and $+33.723901$ nats (Table 8 in Appendix E). Seed 42 still improved accuracy from 92.9688% to 94.5313% while its logit RMS rose from 12.15 to 42.48; seed 43's accuracy fell to 11.1328%. Harm in this recipe therefore does not require first switching to current-batch statistics. The two conventions interact with opposite signs in the two seeds (Appendix E).

6.1 NATIVE HEAD AND PARAMETER-SUBSET ADAPTATION ON HOSPITAL 2

A second native study used six ERM checkpoints (seeds 42 to 47) that share ImageNet initialization and all source-training data and differ only in the seed (projection and classifier initialization, data order, per-epoch subsampling and augmentation draws): they replicate training noise and, with the patient rotation below, the choice of patients, not the source data or hospital. The panel has 384 reference patches (six source-hospital patients), 320 adaptation patches (four hospital-2 patients) and 640 query patches (five other hospital-2 patients; 33.3 percent tumor). We logged hospital-1 and hospital-2 accuracies every epoch. The sweep's gate step printed hospital-2 accuracy for manual review, while its written pass criteria concern representation norms and training accuracy. Every checkpoint is the final epoch and we use all six seeds the sweep produced, so the study is retrospective but not selected on hospital 2 within these runs (Appendix E); within-checkpoint contrasts do not depend on it.

Head-only entropy updates sharpen the score. With saved BN statistics, head-only entropy SGD at rate .01 for 30 steps started from frozen query NLL .81 to 2.66 and AUROC .79 to .93. The native update raised NLL by .10 to .25 nats in every seed (mean $+.155$) and in all 30 seed-by-patient cells (minimum $+.008$). Accuracy was unchanged, and AUROC changed by less than 10^{-4} in every seed. The harm grew with frozen NLL across the six seeds (Spearman .89). A label-free affine map of the initial score reproduces the native NLL to within 6×10^{-4} nats in every seed, over 99 percent of the harm. A one-seed ViT checkpoint shows the same affine reproduction; its AUROC change of $+.0017$ is just above the 10^{-3} unchanged-ranking bound (Appendix F). The update decomposes exactly (the head has no bias; a Euclidean projection of the head weight, distinct from the output-space affine reference above) into a scale factor a_T from 1.06 to 1.14 on the initial score plus a head-null component. Its query RMS is 10 to 13 percent of the scale change, and its ranking nearly matches the initial score's. Writing $w_T = a_T w_0 + n_T$ with $n_T \perp w_0$, every query logit change splits exactly as $\Delta z(x) = (a_T - 1) z_0(x) + n_T^\top \phi_0(x) + w_T^\top [\phi_T(x) - \phi_0(x)]$, a rescaled initial score, a head-null term and a representation-change term; head-only arms have no third term, BatchNorm-affine arms keep the head fixed and have only the third, and in the all-parameter arm this ordering assigns the cross term to the third. Deleting the head-null adaptation-minus-reference mean direction changed NLL by $-.0075$ relative to random deletions (declared criterion $-.01$). The advantage disappeared at matched endpoint movement, where the deletion was worse than random deletions in five of six seeds (mean $+.0007$). Scaling that direction's coordinate by four increased the calibration harm and left the ranking nearly unchanged. A hash-fixed rotation of the nine hospital-2 patients into two further partitions, on the same exports and recipe, kept the harm positive in all 18 seed-by-partition cells (minimum $+.048$), the six-seed mean deletion contrast above $-.01$ on both partitions ($-.006$ and $-.003$), and the ranking unchanged (largest AUROC change 7×10^{-5}). Centering with a compensating bias reduced the harm below .01 nats on the original panel in five of six seeds (magnitudes at most .003; seed 42: $+.113$). Under patient rotation it did not: centered minus frozen NLL exceeded $+.01$ in three of six seeds on each partition (seed 42 on one: $+.445$, above that cell's native harm of $+.169$). The transported full-moment preconditioned update moved the adaptation logits by under two percent of the native movement and was nearly harmless. The sharpening scales with the rate and appears under pseudo-label, information-maximization and diversity objectives, always with the ranking unchanged (post hoc analyses and per-seed values: Appendix F).

An invariance-penalized stratum shrinks its read scale. For a logit sg with read scale s, the IRMv1 scalar-dummy gradient of environment e is $s\, R_e'(s)$, where R_e is that environment's logistic risk of sg, so the penalty is $O(s^2)$ near $s = 0$ while the risk changes linearly; unless every environment shares a nonzero calibration root, as in Theorem 1(i), a large fixed penalty favors a small read scale. The six IRM checkpoints (penalty strength 1000) took that route: logit standard deviation about 10^{-3}, outputs uninformative in probability, and a frozen ranking that survives in the read direction (AUROC .84 to .94 in five seeds). Entropy at rate .01 was inert. At a stress rate matched to the ERM twin's first-step movement, every seed collapsed to one predicted class. Where the adaptation and query features lie on one line, a head update can only keep or reverse the query ranking, which it reversed in two seeds (AUROC .944 to .056 and .936 to .064). A two-parameter labeled recalibration of the standardized frozen logit reached NLL .33 to .49 with the frozen ranking (details: Appendix F).

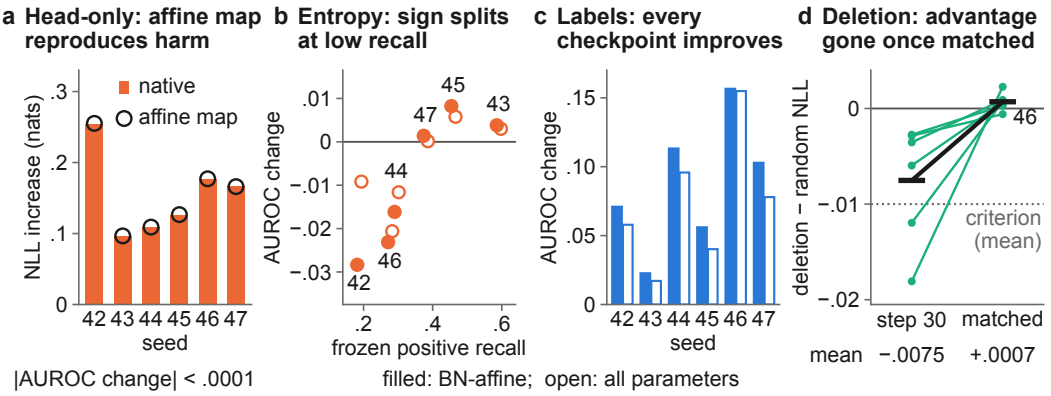


Figure 2: Native adaptation on the hospital-2 query panel (640 patches; six ERM checkpoints, seeds 42 to 47). (a) Head-only entropy, 30 SGD steps at rate .01: query NLL increase of the native update (bars) and of the label-free affine map of the initial score (rings), within 6×10^{-4} nats of each other in every seed; $|\Delta\text{AUROC}| \leq 4.4 \times 10^{-5}$. (b) Ten entropy steps at first-step-matched rates: AUROC change against the frozen model's positive recall for BN-affine (filled) and all-parameter (open) updates, labeled by seed and offset by $\pm.006$ in recall for legibility; the split at recall is post hoc. (c) The same subsets and rates trained with the 320 adaptation labels: AUROC change per seed, on its own scale. (d) Query NLL of the head-null deletion minus the mean of four random deletions, per seed (lines) and averaged over seeds (black), at step 30 and after every endpoint is rescaled to the group's smallest adaptation-panel logit movement, which is the deletion's own in every seed (matched; positive in five of six seeds, seed 46 at $-.0006$); dotted, the declared criterion $-.01$ for the step-30 mean.

Parameter subsets. The subset comparison uses plain SGD for ten steps, with rates matched label-free so that each subset's first step moves the reference logits by the same RMS as the head-only step at .01 (verified within 5 percent; 11 percent for seed 45's all-parameter subset). Labeled arms share the subset rates and are privileged comparators, not ceilings; the prior-shift arm is a transductive EM estimate (Table 10). Table 10 in Appendix F reports six-seed means and Figure 2 every seed. Head-only entropy raised NLL in every seed and left AUROC within 10^{-4} of frozen, consistent with the exact split, whose head-null term is small and ranks like the initial score. BatchNorm-affine and all-parameter entropy changed the ranking. BN-affine AUROC fell in seeds 42, 44 and 46 (by .028, .016 and .023; in seeds 42 and 44 in all five query patients). In 43, 45 and 47 it rose (by .004, .008 and .001, the last at the 10^{-3} bound used for an unchanged ranking; NLL also fell in 43 and 45, by .020 and .044). Post hoc, the sign splits at the frozen predictor's positive recall on the query panel (.19, .30 and .28 in the falling seeds against .59, .46 and .38), while the size of the change is not monotone in it. In the falling seeds the representation-change term $w_T^\top(\phi_T - \phi_0)$ alone ranked the query with AUROC .31, .06 and .08. The unlabeled update drifted the query toward the model's majority (negative) class, most strongly on the true positives (within the frozen model's predicted negatives it ranks positives below negatives, AUROC .08 to .29, against .89, .90 and .73 in the three rising seeds), flipping 23 of seed 42's 40 true-positive calls. The initial unread coordinates partly predict this drift. A label-free ridge on the head-null part of ϕ_0, fitted on three query patients and evaluated on the other two, explains the representation-change term beyond an isotonic function of the initial score in five of six seeds (held-out R^2 up to .75; random features never do). Which seed fails depends on the ridge strength, and in the falling seeds that predictable part ranks against the label. In every seed, the BN-affine and all-parameter subsets trained with 320 adaptation labels raised AUROC (BN affine: $+.02$ to $+.16$) and lowered NLL. The ten-step labeled head lowered NLL but left the ranking unchanged, while 200 labeled head steps raised AUROC by up to .05. So labels can recruit head-null information that entropy does not. A ridge-damped labeled recalibration of the frozen score, a post hoc comparator adopted after the prespecified temperature arm diverged in every seed, reaches the lowest mean NLL of any arm with the frozen ranking. The uncalibrated EM prior overshot toward the negative class and worsened NLL in every seed. At the common numeric rate .01 for 30 steps instead of first-step-matched rates (a separate run on A100 GPUs; Table 11 in Appendix F), the head arm reproduced the sharpening above, and score-only scalar and affine

controls moved little (mean NLL $+.04$). The BN-affine and all-parameter arms over-moved. BN-affine entropy lowered query AUROC in five seeds (to .26 to .80 from .79 to .93; below .5 in three) with nearly one-class prediction while the source-reference ranking survived (AUROC at least .968); all-parameter entropy predicted one class in every seed. Only seed 45 improved its ranking under BN-affine entropy in both conventions (AUROC .860 to .947 and NLL .865 to .616 at the common rate). Its frozen positive recall is the second highest (.46; seed 43, at .59, falls to AUROC .80 at this rate), so frozen recall does not single it out. At 30 matched steps the BN-affine split persisted and widened, and score-only entropy controls reproduced the head-only harm to three decimals (Appendix F, Table 13). The sign statements are exact for these six shared-source checkpoints and patients; any inference beyond them is descriptive. A controlled two-color MNIST family of thirty freshly trained pixel models (Appendix G) shows the same boundary with a known cue. The models retain color, and paired-color heads largely do not read it. Labeled all-parameter training recruited it ($+.106$ aligned, $-.106$ reversed, ten of ten seeds); entropy did not.

7 Scope, Limitations, and Implications

We tested a selected unread-direction restriction twice, in the original-scale synthetic pilot and in the native hospital-2 deletion, and neither shows a rescue. The pilot's targeted-minus-random contrast was positive and practically identical to random (largest absolute difference .0028 nats on risks near 1.75). The native deletion's advantage disappeared at matched endpoint movement. Neither result shows that retained nuisance response is irrelevant. Under the forced centered learner the favorable secondary restriction lowers risk without approaching frozen risk, and random restrictions match rank and movement, not every form of information removed (Appendix I). Three checks changed conclusions here, and we recommend them. Before attributing harm to features, score the affine reference of the initial score; it reproduces ordinary head harm to 3×10^{-5} nats in the synthetic pilot and 6×10^{-4} nats natively. Compare restrictions at matched movement, which turns the native $-.0075$ into $+.0007$. A labeled comparator on the same parameters and rates showed that the hospital-2 panel contains adaptable signal the unlabeled objective misses, so include one.

These studies do not establish useful label-free native adaptation. With 320 labels, the same BN-affine parameters raised AUROC in every hospital-2 checkpoint, so the failure is not for lack of adaptable signal. A successful feasibility test would establish room for benefit under its learner and task, but benefit alone could arise from recalibrating an existing score. A causal recruitment claim would additionally require a nontrivial intervention at a valid location, verified initial equivalence, controls for other information removed, and a task on which useful learning survives the restriction.

The evidence remains descriptive. The head-pilot checkpoints and the retained synthetic seeds each share one source-data draw. The frozen-feature Camelyon17 studies use constructed streams without patient-independent evaluation. The native-image studies separate adaptation and query patients but reuse development patients and shared-source checkpoints, with a role-composition mismatch on hospital 1 and retrospective monitoring of hospital 2, so they are not fresh-subject confirmation. We replaced no panel after observing failure. A generalization claim would need fresh data, a separately frozen analysis and the relevant independent sampling units (Appendix I). A same-day preregistered replication on Waterbirds-95 (Appendix H; six fresh ResNet-50 models, frozen NLL .49 to .92) supplies the first two: it reproduces the head-only harm with its affine reproduction in every seed, the sign pattern of the BN-affine arm within the 10^{-3} ranking bound and the labeled gains, but not the widespread one-class collapse at a fixed rate. The constructions and experiments cover binary logistic classification only, and we did not study shifts in the causal mechanism itself.

Prediction-based evidence applies to the predictor and evaluation law it measured; extending it to a later learner requires the trainable subset, the update geometry and the other state that affects the update. These studies establish neither a prospective monitor nor a practical defense. What they give a reader is the three comparisons above, to make before attributing harm, or protection, to a representation component. The next question is whether an intervention can prevent harmful recruitment while preserving useful learning beyond recalibration; this work says what such a test must distinguish and does not answer it. The hospital-2 checkpoints are already overconfident under shift (frozen NLL .81 to 2.66), and the affine reference closely reproduced the loss rise from label-free sharpening. The non-standard head, the missing filtering or prior-corrected test-time baselines and the unprobed IRM directions limit transfer to standard settings.

AI USE STATEMENT

In this work, we used generative AI tools to help organize and draft the manuscript; to review and challenge mathematical arguments and to derive and check equations; to propose experimental controls; to inspect research sources and reported prior work; and to assist with code design, implementation review, running and auditing analyses, and descriptive result analysis. We have not used generative AI tools for any other task with required disclosure. We have reviewed all AI-assisted work. Some mathematical and numerical checks were written with AI assistance, and AI assistants ran some later native analyses; the two hospital-1 BatchNorm studies were executed locally by the authors and audited from retained artifacts. This assistance is not independent replication or a guarantee of correctness, and the authors reviewed the sources, proofs and implementation and relied on independent recomputations of every reported number from the retained artifacts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The experiments concern model behavior under controlled shifts. Camelyon17 features and images are used as a distribution-shift research setting, not for clinical decisions or a clinical validation study. Hospital association is not treated as a causal clinical attribute. Privileged labels and group information used to construct the frozen-feature intervention are identified separately from the unlabeled learner's inputs; the native image panels were selected without labels. The paper makes no deployment-safety or medical-efficacy claim. Source-data access conditions, licenses and release permissions must be respected in any eventual artifact release; no checkpoints are assumed to be publicly distributable.

REPRODUCIBILITY STATEMENT

Our records separate claims, source provenance, retained-result summaries and run records. Exact constructions require only the displayed probability laws and update equations (Sections 2 and 3); Appendices J–Q give the proofs and supporting derivations. Tables in Appendix A can be traced to saved per-checkpoint rows and the aggregation rule stated there; Appendix B gives the reporting units. The head pilot has a separate protocol, retained per-checkpoint outputs and source/configuration hashes; its primary contrast was recomputed from the underlying risk records (Section 5 and Appendix C). The common-budget crossing reuses original diagonal cells and records crossed cells separately. The affine-reference scoring has its own array-integrity and numerical checks. The two hospital-1 native studies retain code/protocol and data hashes, ordered image identities, per-step logits, state audits, environment versions, implementation-check logs and execution records (Appendix E); the hospital-2 head study, the IRM stratum, the parameter-subset study and the natural-rate run retain sealed predictions and features hashed before labels were opened, per-arm training traces and state audits, and independent recomputations of every reported metric from those artifacts; the patient rotation, the exact head split, the AUROC statistics, the ridge recalibration and the penalty-strength sweep are post hoc analyses on the same sealed exports and retained checkpoints. The saved-statistics follow-up pins the original study's sources and artifacts rather than regenerating its controls. The Waterbirds-95 replication (Appendix H) retains its preregistration with five dated corrections in a hash chain bound before any query label was opened, the source-training and adaptation receipts, a global seal of 429 files, the sealed predictions, the score run and an independent result audit that recomputed every registered quantity. The supplementary material contains anonymized snapshots of the scoring and analysis code, the protocol files with machine paths redacted, the sealed result tables of five studies (the native head-only, parameter-subset, natural-rate, controlled-MNIST and Waterbirds studies) and the figure scripts with their plotted-value tables. Re-running the scoring or regenerating the figures requires inputs that are not included: prediction arrays, labels, checkpoints and feature exports.

REFERENCES

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

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063. Curran Associates, Inc., 2023. doi: 10.52202/075280-2884. URL https://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21061–21071, 2020. URL https://proceedings.neurips.cc/paper/2020/file/f1298750ed09618717f9c10ea8d1d3b0-Paper.pdf.

Kyle E. Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled Cayley transform. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1969–1978. PMLR, 2018. URL https://proceedings.mlr.press/v80/helfrich18a.html.

Floris Holstege, Bram Wouters, Noud van Giersbergen, and Cees Diks. Removing spurious concepts from neural network representations via joint subspace estimation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 18568–18610. PMLR, 2024. URL https://proceedings.mlr.press/v235/holstege24a.html.

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

Kia Khezeli, Arno Blaas, Frank Soboczenski, Nicholas Chia, and John Kalantari. On invariance penalties for risk minimization, 2021. URL https://arxiv.org/abs/2106.09777.

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://openreview.net/forum?id=Zb6c8A-Fghk.

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

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

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/cd0665986c1e9c15f6569ac944bcf88a-Paper-Conference.pdf.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://jmlr.org/papers/volume21/17-678/17-678.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://openreview.net/forum?id=fSTD6NFIW_b.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.12400.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BbNIbVPJ-42.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7201–7211, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

A Retrospective learned-checkpoint experiments

The results in this section come from retained studies A–D and a descriptive reanalysis of their saved rows. They are distinct from the exact population construction and from the new study in Section 5. No retained source-training run is described as newly executed. We report checkpoint-level summaries, without converting adaptation streams into independent source-model replications. Their strongest comparisons hold the learned checkpoint, raw stream and initial predictions fixed.

The synthetic setting supplies a controlled nuisance intervention. The Camelyon17 setting uses constructed hospital–label associations and a privileged hospital-associated direction; it does not identify a causal clinical nuisance. These studies also separate increased response from near-constant prediction drift.

Configuration labels pair a retained encoder with its source-training penalty. Cayley denotes retained encoders built with the Cayley parameterization (Helfrich et al., 2018), whose affine-free LayerNorm is discussed in Appendix N; GL denotes the unconstrained general-linear encoder family, whose square weight matrices range over $\mathrm{GL}(d)$ rather than the orthogonal group $\mathrm{O}(d)$ of the Cayley family. IRM denotes the IRM penalty (Arjovsky et al., 2019), reported with its penalty coefficient, and V-REx denotes the risk-extrapolation penalty of Krueger et al. (2021). Thus Cayley+IRM, GL+IRM and Cayley+V-REx name retained configurations.

A.1 Protocol and measurement

Synthetic setting. The retained study contains nine configurations with 25 final source-training checkpoints each. Within a configuration, training seeds change initialization and minibatch order on one fixed synthetic source-data draw. Each checkpoint receives five independent adaptation streams, each comprising 30 batches of 512 unlabeled examples: ten each from source, neutral, and reversed settings with generator parameters $\gamma = 1.8, 0, -1.5$. Fixed evaluation sets contain 5,000 examples per setting. The main results use learning rate .01 and step 30; retained scans also use .1 and earlier evaluation times.

Let $d_S(x) = [\phi(x^{S+\delta}) - \phi(x^{S-\delta})]/(2\delta)$. Study A uses $\delta = 1$ raw input unit and selects the normalized mean response projected off the initial head. Study D uses separate direction-estimation samples and perturbations of one training standard deviation. These units must not be interchanged. The measured logit sensitivity is $\mathbb{E}[(w^\top d_S(X))^2]$. It is a finite-difference response diagnostic, not mutual information or complete nuisance recovery.

For the selected unit direction $u \perp w_0$, the transformation

$$A_s = I + (s-1)uu^\top, \qquad \phi_s = A_s\phi \tag{5}$$

preserves every initial logit. The entropy scan uses $s \in \{1, 2, 4, 6, 8, 12, 16, 24\}$; method comparisons use $s = 1, 16$. This is an intentional intervention on learned representation directions, not evidence that spontaneous scale changes of this size occur in deployment. The initial entire predictor still has nonzero synthetic nuisance sensitivity.

Frozen-feature Camelyon17 setting. The clinical feature study uses eight configurations and 25 source-training checkpoints per configuration, with three streams per checkpoint. The data loader uses pre-extracted features, a fixed 2048-to-256 orthogonal projection, and standardization from training hospitals. Learned encoders are trained on hospitals $0, 3, 4$; hospitals $1, 2$ supply the constructed adaptation pools. Each hospital evaluation has 3,000 examples per class, disjoint from those adaptation pools at patch level. Patient- or slide-independent separation is not established by these records. The study is related to the Camelyon17 distribution-shift setting in WILDS (Koh et al., 2021), but its balanced hospital evaluations and worst-hospital endpoint are not the official WILDS evaluation protocol.

The selected direction averages, across classes, the hospital-2 minus hospital-1 representation means and projects this vector off w_0. Both labels and hospital identities are used to construct the direction and stream. The learner receives features only. Each 512-example batch is label balanced, with $\mathbb{P}(h = 2 \mid Y = 1) = r$ and $\mathbb{P}(h = 2 \mid Y = 0) = 1 - r$; r takes values $.9, .5, .1$ for ten batches each. Scales are $1, 4, 16$. Hospital-associated response is not a causally isolated nuisance, and balanced per-hospital risks do not reconstruct the class-conditional mixture risk of each stream phase.

Controls and aggregation. Retained controls include frozen behavior, head-norm reset, diagonal second-moment scaling, full second-moment solves, a reference-RMS guard, targeted gradient projection, diversity regularization, and recentering with bias compensation. Study D adds free-bias, centered/free-bias and bias-only conditions, as well as principal and random *amplification* directions. Those random directions are not random-deletion controls. Full moments are computed from the current unlabeled batch, with $\lambda = 10^{-8}\,\mathrm{tr}(G)/d$ in $(G + \lambda I)^{-1}g$.

Table 1: Retained synthetic Cayley+IRM, penalty coefficient 10^3. Accuracy and binary logistic risk on declared evaluation laws. Same checkpoints, initial predictions, raw streams, learning rate .01, and 30 updates; feature scales differ. Medians of 25 checkpoint means over five streams.

	Source		Neutral		Reversed	
Condition	Acc.	Risk	Acc.	Risk	Acc.	Risk
Initial, either scale	.6788	.5930	.5552	.7299	.4462	.8528
Scale 1, step 30	.5026	1.4245	.5044	1.5664	.4894	1.7249
Scale 16, step 30	.7115	.9417	.5064	1.9604	.3323	3.1139

Table 2: Retained frozen-feature Camelyon17 Cayley+IRM. Per-hospital accuracy/risk, medians of 25 checkpoint means over three streams, rate .01. Hospital 0 is a source hospital; 1 and 2 are held-out pools. The intervention and stream construction are privileged stress tests.

	Hospital 0		Hospital 1		Hospital 2	
Condition	Acc.	Risk	Acc.	Risk	Acc.	Risk
Initial, either scale	.9292	.1688	.8492	.3550	.8688	.3136
Scale 1, step 30	.9304	.1682	.8445	.3830	.8713	.3197
Scale 16, step 30	.8070	.9267	.5569	3.1208	.8049	.6615

For Tables 1–2, we average streams within each checkpoint and then take the median over 25 checkpoint means. These are descriptive endpoint summaries, not confidence intervals. Code inspection and row recomputation do not independently reproduce checkpoint generation. Architecture labels identify retained configurations; differences between them cannot isolate architectural causes because source-training clipping and other optimization choices were not fully matched.

A.2 Recruitment and collapse are different outcomes

At the original scale, synthetic Cayley+IRM often moves toward predicting one class. In the central coefficient-10^3 comparison, the positive prediction fraction reaches approximately .975, while selected raw-unit nuisance sensitivity changes only from about .03034 to .03233. Under scale 16, sensitivity reaches approximately 1.84305 and reversed accuracy falls further (Table 1). The latter improvement in source accuracy is accompanied by worse source logistic risk, illustrating the need to name the metric.

The original-scale increase in reversed accuracy from below chance toward chance is not evidence of useful nuisance removal. Across the two synthetic IRM arms and two learning rates, a later descriptive check finds 474 of 500 checkpoint–stream conditions improve reversed accuracy by more than one point, but none beats both its frozen accuracy and the best empirical constant by more than one point. Logistic risk rises in all 500. These are correlated descriptive conditions, not 500 independent training-data replications. The best empirical constant is an evaluation diagnostic, not a label-free operational method.

Mean/intercept controls explain part of this behavior. At rate .01 and 30 steps, Study D reports collapse in 22/25 and 25/25 checkpoint summaries for the two synthetic IRM arms under ordinary and uncentered free-bias updates, with zero under the centered fixed-bias and centered free-bias controls at that horizon. Bias-only updates also show zero under that criterion. The changes align with the source representation mean. Near zero logits, entropy descent approximately amplifies the eigenmodes of the uncentered second moment $\mathrm{Cov}(\phi) + \mu\mu^\top$ (Appendix O); this provides an explanation, not a theorem that every entropy collapse is caused by the mean. Centering and diversity can suppress the measured collapse without consistently preventing recruitment after amplification.

A.3 Hospital-associated degradation without global class collapse

Worst-hospital accuracy in the central comparison changes from an initial median .8475 to .8445 at scale 1 and .5569 at scale 16. The median of the per-checkpoint minimum across five hospitals need not equal the minimum of the displayed marginal medians. This is controlled susceptibility to a

prediction-preserving intervention, not a prevalence estimate for naturally occurring hospital shifts. A post hoc check averaging positive prediction fractions over all five balanced hospital evaluations finds no strict global one-class collapse in the 75 Cayley+IRM checkpoint–stream runs at either scale and rate .01 (minority threshold .05). Other configurations do collapse. Thus hospital-associated harm is not automatically the synthetic mean-collapse mechanism.

Natural-scale gains and harms also coexist in some retained clinical arms. In GL+IRM at rate .01 on hospital 2, eight of 25 checkpoint means improve accuracy by more than one point and six lose more than five points. None of the corresponding improving checkpoint–stream episodes lowers logistic risk. Such gains remain real if zero-one accuracy is the declared utility, but do not establish general beneficial adaptation or a clinical safety improvement. Sparse evaluations also cannot reconstruct cumulative deployed mistakes.

A.4 What the intervention controls establish

Full second-moment preconditioning nearly removes dependence on the tested scales. In the central synthetic and clinical comparisons, scale-1 and scale-16 checkpoint-mean accuracies agree, with residual response discrepancies at small numerical/ridge scale. This is consistent with the known congruence calculation and supports an approximate invariance control. It does not show that preconditioning eliminates all nuisance dependence, improves risk, or remains invariant under arbitrary singular/damped transformations.

The targeted rank-one projected-gradient condition does not generally rescue original-scale clinical performance. At rate .01 and step 30, median worst-hospital accuracies for ordinary versus projected updates are approximately $.845/.845$ for Cayley+IRM, $.812/.811$ for Cayley+V-REx, and $.837/.820$ for GL+IRM. Reducing a selected hospital logit gap therefore does not imply improved risk. These negative controls are retained rather than replacing the target direction after observing failures. Section 4 states the conclusion drawn from these retained studies. The main rescaling results do not match reference movement, and the older head/encoder comparison does not supply paired equal-budget evidence. Section 5 tests a stronger original-scale component restriction at matched neutral-reference movement and reports its negative primary result. Its subsequent forced-budget comparisons distinguish component access from collapse suppression and useful adaptation.

B Retained-result provenance and reporting units

The exact population table is evaluated from the eight states of (C, B_c, B_s). It is separate from the earlier finite-batch binary and digit experiments and from the learned-checkpoint studies. The digit study used one trained source classifier with an appended nuisance coordinate; it is not counted as an additional independently trained representation campaign here.

The synthetic and clinical tables summarize source-checkpoint means over their adaptation streams, followed by medians over checkpoints. Different column medians can be attained by different checkpoints. Differences between displayed medians are descriptive and are not paired-effect confidence intervals. Clinical worst-hospital rows contain no risk value; per-hospital risks are reported directly rather than inventing a worst-hospital risk from those rows.

Post hoc collapse and benefit analyses are labeled as such. The counts of 125 synthetic or 75 clinical checkpoint–stream conditions are not independent data-draw counts. The executed pilot uses existing checkpoints with fresh role-specific samples and does not repair that source-data limitation. A complete new replication requires fresh source data, retained failures, a frozen analysis and uncertainty clustered at the appropriate source-data unit.

C Head-diagnostic sampling, execution checks and stress controls

This appendix collects protocol constants, execution checks and the scale-16 stress comparison for Section 5.

Table 3: Ordinary-update controls from the same executed pilot. Entries average the same four checkpoints. Collapse counts use target predicted minority fraction below .05. Matched budgets are shared across scales and restrictions *within* each centering condition; raw and centered budgets differ substantially. Centering therefore does not supply an across-centering equal-movement comparison in this pilot.

Geometry	Update/scale	Target risk	Accuracy (%)	Collapse	Ref. RMS
Frozen	Initial	.798979	45.03	0/4	0
Raw	Matched/1	1.745193	50.90	4/4	3.178554
Centered	Matched/1	.804933	45.09	0/4	.030199
Raw	Native/16	2.730912	32.44	0/4	3.745437
Raw	Matched/16	2.374649	32.60	0/4	3.178554
Centered	Native/16	2.833006	31.74	0/4	3.767140
Centered	Matched/16	.807473	44.29	0/4	.030199

Sampling. Direction, source-mean, movement-reference and affine-fitting samples each contained 2,048 examples with separate role seeds. Labeled source, neutral and reversed evaluations each contained 4,096 examples; their intervention-response samples contained 2,048. Initial source-training standardization was retained.

Execution and checks. All 224 condition paths completed, yielding 4,032 evaluation records. The job completed with PyTorch CPU and eight threads in 33.92 seconds; this is measured wall time for the pilot, not new source training. Original metrics passed on all twelve checkpoint/evaluation-law pairs. Initial-equivalence discrepancies were at most 3.11×10^{-15}, projected-component leakage at most 1.46×10^{-16}, and the recomputed matched-budget discrepancy at most 8.89×10^{-16}. These checks support implementation consistency; they do not validate the scientific hypothesis.

Stress effects and centering have different boundaries. At scale 16, ordinary matched-budget target risk was 2.374649, with 32.60% accuracy and no global target collapse. Removing access to the deliberately amplified component reduced risk to 1.745825. The restricted scale-1 and scale-16 risks agreed within 4.45×10^{-16}, as the projection algebra predicts. The restricted models still collapsed and remained substantially worse in logistic risk than their frozen versions. This establishes control of the constructed amplification path, not a useful general safeguard. Table 3 lists the ordinary-update controls, including centering. At centered scale 16, a large native-rate risk effect became small under the centered movement budget (Table 3).

Common-budget follow-up execution. All features, streams, direction estimates, checkpoints and endpoints were reused; the original own-budget cells were not rerun. All 56 added paths completed: 48 adaptive paths and eight frozen bookkeeping paths, with 1,008 evaluation records and 1,736 trace records. The follow-up took 3.54 seconds on the same CPU environment after 18 numerical and pipeline checks passed. Actual reference radii matched the correct saved donor schedules within 8.89×10^{-16}; the restriction bases and cached features remained unchanged. The original primary mean $+.000622740$ is retained and is not replaced by a favorable follow-up outcome.

Affine-reference scoring check. Across 336 terminal comparisons, initial risks recomputed from the numeric feature arrays agreed with saved risks within 3.11×10^{-15}. No adaptation was rerun.

Residual-response decomposition. For the secondary comparison in Section 5, if $D_t = \alpha_t D_0 + e_t$ denotes the paired intervention response, then

$$\mathbb{E}[D_t^2] = \alpha_t^2 \mathbb{E}[D_0^2] + \mathbb{E}[e_t^2] + 2\alpha_t \mathbb{E}[D_0 e_t].$$

The cross term is signed. After targeting, the normalized L^2 inner product between D_0 and e_t was approximately $-.922, -.901, -.872, -.927$: a larger residual could cancel more of the initial response.

Signed affine-reference accounting. For the affine-reference comparisons in Section 5, the identity

$$R_t - R_0 = (R_{\text{aff}} - R_0) + (R_t - R_{\text{aff}})$$

Table 4: Executed primary pilot: original feature scale, uncentered representation, matched neutral-reference movement, step 30. Each random comparator averages four prespecified rank-one restrictions. All values are reversed-law logistic risk in nats. The four checkpoints comprise two initialization seeds and two IRM penalties on shared training data; the mean is descriptive, without an independent-data or significance claim. Negative Δ favors the targeted restriction.

IRM penalty	Seed	Targeted risk	Mean random risk	Δ
1,000	42	2.016637	2.013810	+0.002827
1,000	43	1.776970	1.776849	+0.000120
5,000	42	1.758859	1.759359	-0.000499
5,000	43	1.430834	1.430791	+0.000043
Descriptive mean		1.745825	1.745202	+0.000623

Table 5: Executed budget crossing and affine-reference check. All entries are equally weighted means over the same four retained checkpoints at step 30 and feature scale 1. B_{ctr} and B_{raw} are the checkpoint-specific saved centered and raw schedules. Actual and affine-reference risks use the reversed target law. Minority is the smaller predicted class fraction; balanced class frequencies do not imply low risk. The affine reference uses saved unlabeled-fit coefficients and is not a separate budget-matched learner.

Learner	Budget	Actual risk	Affine risk	Minority	Ref. RMS
Frozen	None	.798979	.798979	.402283	0
Raw	B_{ctr}	.800756	.800745	.385498	.030199
Centered	B_{ctr}	.804933	.805112	.406738	.030199
Raw	B_{raw}	1.745193	1.745163	0	3.178554
Centered	B_{raw}	1.890186	1.776839	.488159	3.178554

is signed accounting for one declared reference, not unique causal mediation. The fitted reference need not itself obey the learner's movement budget.

Scope of the selected component. The selected latent component is associated with S but is not identified as a pure semantic channel; same-rank random restrictions are not matched in every form of removed information.

D BUDGET CROSSING, SECONDARY RESTRICTION AND AFFINE ACCOUNTING

This appendix gives the head-pilot comparisons summarized in Section 5.

Affine reference and residual response. We fit $f_t \approx \alpha_t f_0 + \beta_t$ on the separate unlabeled calibration sample and reported the residual response $\mathcal{D}_S f_t - \widehat{\alpha}_t \mathcal{D}_S f_0$ as well as absolute response. This measures departure from the fitted global affine response. A nonzero residual can also arise from a nonlinear function of the old score; it does not establish new information use or nuisance erasure.

Crossing geometry with a common movement budget. The follow-up added the two missing scale-1 cells: raw geometry at the saved centered budget and centered geometry at the saved raw budget. The new primary was ordinary centered-minus-raw target risk at the common raw schedule. Across the four checkpoints, the differences were $+.076750$, $+.188589$, $+.096936$ and $+.217695$ nats, averaging $+.144993$. Centering removed one-class collapse but increased target logistic risk in every pair. At the reciprocal small budget, raw adaptation also avoided collapse and had lower risk in all four pairs, with mean difference $+.004177$ against centering (Table 5).

Thus the earlier protection did not persist at equal reference movement. This does not invalidate reducing movement as a practical effect of centering. It identifies a limit of the stronger explanation that centering protects risk independently of movement. The forced centered trajectories required displacement multipliers as large as 38.83. They are diagnostic updates, not native gradient descent or a faithful published test-time adaptation method.

Table 6: Native BN feasibility, specified before execution. Loss is binary logistic risk in nats; accuracy is a percentage. Both seeds use one fixed query panel with 29 positive and 483 negative patches from four patients. The equal-seed mean is descriptive, not an independent-patient estimate. Switching statistics already changes the predictor. Neither seed meets the declared loss-and-accuracy convention.

	Saved-stat frozen		Batch statistics only		BN affine, step 10	
Seed	Loss	Acc.	Loss	Acc.	Loss	Acc.
42	.228001	92.97	4.556555	57.62	5.958988	49.61
43	.237986	94.34	5.770235	54.69	6.764992	49.61
Mean	.232994	93.65	5.163395	56.15	6.361990	49.61

Table 7: Post-pilot saved-statistics diagnostic on the unchanged panel. Loss is query logistic risk in nats and accuracy is a percentage. The declared primary Δ_{saved} is learned-minus-frozen loss; positive is adverse. I is the retained batch-statistics learning effect minus Δ_{saved}. Its opposite signs preclude a consistent descriptive ranking across these two checkpoints.

	Saved-stat frozen		Saved-stat learned			
Seed	Loss	Acc.	Loss	Acc.	Δ_{saved}	I
42	.228001	92.97	.288796	94.53	+.060795	+1.341638
43	.237986	94.34	33.961888	11.13	+33.723901	-32.729144

Component access matters in one secondary comparison. Within the centered large-budget paths, the targeted-minus-mean-random target-risk differences were $-.224302$, $-.155559$, $-.098841$ and $-.093667$ nats. Their mean was $-.143092$. Targeting also reduced total interventional S sensitivity in all four cases, with neither the ordinary nor targeted condition collapsing. This demonstrates an effect of restricting the selected component under this specified forced-budget learner.

It does not demonstrate a useful adaptation safeguard. Mean targeted risk was 1.747783, still far above frozen risk .798979, and targeted accuracy remained below the empirical best constant. Moreover, the squared affine-residual S response increased in three of four cases. A larger residual could cancel more of the initial response (Appendix C); its norm is therefore not a monotone measure of harmful recruitment.

Separating affine score change from additional response. We independently scored the saved reference $g_t(x) = \widehat{\alpha}_t f_0(x) + \widehat{\beta}_t$ on the cached held-out evaluations. The coefficients had been fitted on an unlabeled reference; evaluation labels were used only for scoring. For the raw large-budget learner, the affine reference had mean target risk 1.745163, close to the actual 1.745193. For the centered learner at the same budget, affine risk was 1.776839 and actual risk 1.890186. The additional centered-minus-raw loss $+.144993$ decomposes into $+.031675$ between affine references and $+.113317$ between their signed remainders. A large total loss increase is already present in an affine function of the original prediction, while that reference does not reproduce the extra centering loss. The targeted centered learner had affine risk 1.789350 and actual risk 1.747783. Relative to random restrictions, its affine risk increased by $.012375$ while its actual risk fell by $.143092$. This further rules out interpreting the residual as an additive nonnegative share of harm; Appendix C states the signed accounting identity behind these comparisons.

E Native study protocol and verification details

Tables 6 and 7 give the full native results with accuracies. The descriptive interaction in Section 6 is

$$I = (R_{\text{batch,learned}} - R_{\text{batch,frozen}}) - (R_{\text{saved,learned}} - R_{\text{saved,frozen}}).$$

The descriptive interaction I is the batch-statistics learning effect minus the saved-statistics learning effect, each relative to its own frozen control. Its signs disagree: $+1.341638$ and -32.729144

nats. The mean saved-statistics loss increase, 16.892348, and mean interaction, -15.693753, are dominated by seed 43 and do not justify a general ranking of normalization conventions.

The native-image architecture is a ResNet-50 feature extractor, adaptive average pooling, a 2048–256–128 projection with an intermediate ReLU, and the retained unconstrained scalar classifier. Tensor-only full checkpoints were loaded strictly; no pretrained weights were downloaded. Native images were decoded at 96×96, converted to RGB and normalized with the fixed ImageNet channel means and standard deviations. No resize or augmentation was introduced. The ten entropy-only Adam updates used rate 10^{-3}, $\beta = (.9, .999)$, $\epsilon = 10^{-8}$ and zero weight decay; the 106 existing BN scale/offset tensors (53,120 elements) were the only trainable parameters.

Source training. The six checkpoints (seeds 42 to 47; the hospital-1 studies use seeds 42 and 43) were fine-tuned end to end by ERM for 20 epochs from an ImageNet-pretrained torchvision ResNet-50 (IMAGENET1K_V1), with the 2048–256–128 projection and a bias-free linear classifier initialized by the seed, on the metadata split-0 rows of hospitals 0, 3 and 4. Each step concatenated 32 patches from each hospital, each epoch subsampled every hospital to the size of the smallest, native 96×96 patches received random horizontal and vertical flips and ImageNet normalization, and AdamW used learning rate 10^{-4} and weight decay $.01$ without a learning-rate schedule; the saved checkpoint is the final epoch, not a best-validation or best-test epoch. These are the settings of the training script as committed on March 13, 2026, which the sweep invoked with only the configuration name and the seed; the runs themselves record their configuration name, seed, 20 epochs and per-epoch hospital-1 and hospital-2 accuracies, not the optimizer, device, precision or library versions, and the seed-42 run predates that commit. Whether any other seed was trained and discarded is not recorded.

Selection used only hospital-1 rows with the recorded split value zero. Eight numeric patient identifiers were sorted and shuffled with seed 62000100; four patients were assigned to each role. Adaptation sampling selected 80 patches per patient with seed 62000200 and shuffled the result with seed 62000201. Query sampling selected 128 per patient with seed 62000300 and shuffled with seed 62000301. Patient/node/coordinate order was fixed before sampling. These are development masks and roles, not a claim of official WILDS test-split equivalence. Adaptation and query patients, including their slides, were disjoint.

The original protocol and its affine-reference amendment were fixed before the current-batch study's outcomes. Both source checkpoints were retained regardless of result. Loss and accuracy were recorded at steps 0, 1, 5 and 10, with step 10 fixed as the endpoint. For the secondary score reference, initial and final adaptation logits were replayed at their respective fixed parameter states using exactly the same batch companions. Ordinary least squares with an intercept fitted final on initial logits in float64. Neither query logits nor any labels entered fitting. Initial-score variance below 10^{-12} would make the fit unavailable; neither checkpoint encountered that case. Query labels were used only after predictions and reference maps were fixed.

The execution environment was Python 3.12.13, PyTorch 2.13.0, torchvision 0.28.0, NumPy 2.4.4 and Pillow 12.2.0, on CPU with eight intra-op threads and one inter-op thread. There was no GPU or mixed precision. The current-batch study and the follow-up completed in approximately 203 and 139 seconds; their recorded peak process memory was 1.75 and 1.69 billion bytes, respectively. The implementation suites passed 20 and nine named checks before the corresponding studies. These are implementation checks, not scientific replications or statistical sample sizes.

The artifact audit of the current-batch study independently reproduced the identity-only selection and compared all 832 selected image-file hashes, both checkpoint hashes, metadata and approved source hashes. It recomputed retained predictions' losses and other metrics without executing another neural forward pass. Recorded hash maps permitted only the specified BN mode/buffer configuration and affine-parameter changes. The two nonlearning controls remained immutable and the two batch-statistics branches had identical initial predictions. In the follow-up, panel identities, labels and normalized-batch fingerprints matched the current-batch study; initial score replay was exact. All saved BN buffers and all non-BN parameter hashes remained unchanged, and every query evaluation preserved its preceding state.

The follow-up audit recomputed all 40 condition-step score records of both studies and checked 853 declared file-hash comparisons. Its maximum discrepancy under the recorded score convention

Table 8: Native BN query loss in nats under both normalization conventions, on one fixed query panel with 29 positive and 483 negative patches from four patients. Δ_{saved} is saved-statistics learned-minus-frozen loss; positive is adverse. I is the batch-statistics learning effect minus Δ_{saved}; its opposite signs preclude a consistent descriptive ranking across these two checkpoints. Equal-seed means are descriptive, not independent-patient estimates. Tables 6 and 7 in Appendix E add accuracies.

	Saved statistics		Batch statistics			
Seed	Frozen	Learned	Stats only	Learned	Δ_{saved}	I
42	.228001	.288796	4.556555	5.958988	+.060795	+1.341638
43	.237986	33.961888	5.770235	6.764992	+33.723901	−32.729144
Mean	.232994	–	5.163395	6.361990	+16.892348	−15.693753

was 7.82×10^{-14}. These are audit assertions on existing artifacts. They do not recreate unlogged tensor states, rerun gradients or add independent patients. The protocol's remaining limits—old development subjects, only two checkpoints, role-composition mismatch, fixed query batches and no semantic intervention—remain despite those implementation checks.

Additional current-batch results. Retrospective confusion counts show 29/29 and 28/29 positives detected, but 258/483 and 257/483 negatives misclassified. Accuracy near 50% therefore does not mean random guessing or absence of discrimination on this panel.

A prespecified secondary reference fitted one affine map from initial to final logits on the same fixed unlabeled adaptation batches, then applied it to the initial batch-statistics query logits. Its query losses were 5.307853 and 6.499503 nats; actual learner-minus-reference differences were $+.651135$ and $+.265489$. These gaps reject reproduction by this fitted reference only. They do not rule out another old-score map, identify internal unread-feature recruitment, or measure a uniquely mediated share of harm.

Execution. The statistics-only control and the batch-statistics affine learner depend on batch companions. The initial and frozen scores of the saved-statistics diagnostic replayed the current-batch study's saved-statistics scores exactly. The current-batch study passed its 20 implementation checks and completed in 202.984 seconds; the saved-statistics diagnostic passed nine checks and completed in 138.993 seconds. Both used CPU/eight threads, float32 model updates and float64 scoring. Subsequent artifact audits checked saved scores, source/data identities and recorded state invariants; these were not second neural-model executions. In the saved-statistics run, all 159 BN buffers and 58 non-BN parameter tensors retained their recorded hashes, and query evaluations preserved state.

Additional panel and saved-statistics details. Patient and slide roles were disjoint. Both checkpoints and every condition used the same ordered panels and batches of 32. In the saved-statistics diagnostic, seed 43 predicted positive on 484/512 images, including 455 false positives, and accuracy fell to 11.1328%. This is severe positive skew; both classes are still predicted. The convention changes both adaptation gradients and query scoring; this comparison cannot uniquely locate the cause in either phase.

Composition of the hospital-1 panels. The label-blind selection yielded 179/320 adaptation positives and 29/512 query positives (post-run audit), so the empirical majority-class query accuracy was 94.3359%; neither quantity entered selection or entropy updates. The statistics-only query predictor normalizes each query batch by that batch's own statistics; since the query panel is 94.3% negative, its batch statistics differ from the source-training rows that are 50.06 percent tumor (302,436 rows from hospitals 0, 3 and 4). Batch-composition sensitivity of current-batch normalization is therefore the untested ordinary explanation for the initial switch loss; the adaptation-panel mismatch does not bear on it. A later check confirms the dependence on batch context, though not its role in the switch loss: at fixed weights, replacing the 31 companion images of an anchor changed current-batch-statistics logits by an RMS of about 9 and flipped 2 to 3 of 32 signs, while saved-statistics outputs were exactly invariant. A prespecified affine score reference also fails to reproduce the learned losses (Appendix E).

Saved-statistics diagnostic, full statement. After the current-batch study, a separately frozen diagnostic completed the missing saved-statistics learning cell. It reused the exact checkpoints, images, patients, ordering, batches, optimizer and ten-step endpoint. All modules remained in evaluation mode; saved BN means, variances and counters remained fixed while the same affine subset learned. This was an exploratory follow-up on the same development subjects, not independent confirmation or a search over rates, panels or checkpoints.

The declared primary, learned-minus-frozen query loss, was adverse for both checkpoints: $+.060795$ and $+33.723901$ nats (Table 8). Seed 42 nevertheless improved accuracy from 92.9688% to 94.5313%, with unchanged 20/29 positive recall and eight fewer false positives. Its logit RMS rose from 12.15 to 42.48; better classification accuracy coexisted with worse proper loss. Seed 43's accuracy fell to 11.1328% (Appendix E).

The descriptive interaction between the two conventions has opposite signs in the two seeds and does not justify a general ranking of normalization conventions (Appendix E).

Scope. Appendix E gives execution, protocol and audit details. The failed feasibility question remains the result under both conventions. Harm in this recipe does not require first switching to current-batch statistics. No native semantic component was blocked, no hospital association was causally manipulated, and no fresh subjects or source training were introduced. These outcomes limit the attempted bridge to useful native learning; they do not invalidate all entropy-adaptation methods or replace the head-only primary.

F HOSPITAL-2 PER-SEED RESULTS

AUROC statistics, the exact head split, the patient rotation and the ridge recalibration were computed after the sealed outcomes were scored; the parameter-subset ten-step arms of seed 42 were first scored in a smoke run before the six-seed run. Table 9 lists the hospital-2 head results by seed for the original panel, the two rotated partitions and the IRM stratum. Table 12 lists the single-seed penalty-strength observation cited in Section 6.1.

A one-seed ViT check. Three ViT checkpoints (one seed each, ten epochs; a 768-dimensional class-token representation with a linear head whose bias was kept frozen during adaptation) were exported on the same 1,344 images; the input pipeline was recovered from the training wrapper and confirmed by exact reproduction of the recorded hospital-1 validation count (26,549 of 31,430; a zero-padding control gives 26,151). On the ERM checkpoint, head-only entropy SGD at .01 for 30 steps raises query NLL from 1.717 to 1.865 in all five patients, the affine reference of the initial score reproduces the whole increase, the scale factor is 1.07 with a head-null share of 24 percent of the logit change, and AUROC changes by $+.0017$ (frozen .713), which misses the preregistered 10^{-3} bound for an unchanged ranking. The IRM checkpoint has logit standard deviation .0011, NLL within .001 of $\log 2$ and AUROC .844, and the update is inert. A spectrally normalized IRM variant (logit scale .001) responds with a bias-like shift that predicts one class. Eighty-five percent of the ResNet head harm falls in the two tumor-rich query patients.

Additional head-only analyses on the same exports. With the same 128-dimensional features, head and adaptation batches, 30 entropy SGD steps at rates .001, .003, .01, .03 and .1 raise query NLL by mean $+.017$, $+.050$, $+.155$, $+.397$ and $+.958$ nats while the largest absolute AUROC change at any rate and seed is 1.3×10^{-4} (12 reordered pairs of the 213×427; seed 47 at rate .1); pseudo-label self-training (hard labels from the current sign), information maximization and diversity-regularized entropy at rate .01 raise NLL by mean $+.078$, $+.135$ and $+.148$ with the ranking again unchanged (largest absolute AUROC change 1.4×10^{-4}, under information maximization). Paired bootstrap intervals over query rows, stratified by patient (2,000 resamples; conditional on the five query patients), exclude zero for the native head harm in every seed (seed 42: $+.22$ to $+.29$; seed 43: $+.07$ to $+.12$). With the first 10, 30, 100 and 320 adaptation labels in the frozen panel order, the ridge recalibration changes mean NLL by $-.58$, -1.12, -1.13 and -1.07 and a ten-step labeled head (ten steps regardless of the label count) by $-.61$, $-.83$, $-.78$ and $-.47$, the ridge with the frozen ranking and the labeled head with AUROC changes of at most $+.011$ (seed 47 with 30 labels; below 10^{-3} only at 320 labels); those first thirty labels lower NLL in every seed, whereas a

random thirty-label draw does so in about three of four draws, and ten labels fail in one seed. These analyses were run after the sealed outcomes were scored.

Invariance-penalized stratum: details. The six IRM checkpoints' outputs are uninformative in probability (NLL within .01 of $\log 2$); their hospital-2 features span one to four directions, and their frozen ranking survives in the read direction (AUROC .84 to .94 in five seeds, better than the ERM twins in four of those five and worse in seed 43). At the stress rate every seed collapses to one predicted class (query NLL 4 to 16); in the three seeds whose adaptation and query features lie on one line (centered rank one; query rows on it to relative residual below 2×10^{-3}), a head update can only keep or reverse the query ranking up to that residual, and it reversed it in two (AUROC .944 to .056 and .936 to .064); in the two seeds with two to four directions it lowers AUROC from .860 to .763 and from .835 to .757. The two-parameter labeled recalibration reaches NLL .33 to .49 with the frozen ranking in those five seeds. In one seed, raising the penalty strength shrinks the fraction of feature energy the head reads from .58 to 1.3×10^{-7} under a fixed-norm bottleneck while 26 feature directions remain and the read direction still ranks well (AUROC .876); a ten-epoch warm-up variant collapses the scale as well (AUROC .627), and without the bottleneck the same strength collapses the representation to one direction; whether those unread directions carry label information was not measured.

Table 9: Hospital-2 per-seed results. Original panel: head-only entropy SGD at rate .01 for 30 steps with saved BN statistics. H is native minus frozen query NLL; D is the deletion of the head-null adaptation-minus-reference mean direction minus the mean of four random deletions; a_T and the null-over-scale ratio come from the exact split of the final head into a rescaled initial score and a head-null part; the amplified arm scales the direction's feature coordinate by four. Rotated partitions reuse the same exports and head recipe with two further four-plus-five patient roles. IRM stratum: penalty strength 1000, eval mode; the feature rank counts centered directions above 10^{-8} of the top eigenvalue on the adaptation rows (the query rows of seeds 42 and 45 lie on the same line to relative residual 5×10^{-7}, those of seed 43 to 1.8×10^{-3}); "deg." marks seed 44, whose hospital-2 query features are a single repeated vector. Values are rounded to three decimals.

Seed	42	43	44	45	46	47
Original panel, ERM checkpoints						
Frozen NLL	2.663	.808	1.565	.865	1.638	2.026
Frozen AUROC	.792	.929	.824	.860	.797	.802
H: native minus frozen NLL	.255	.096	.109	.126	.177	.167
D: deletion minus random NLL	−.018	−.003	−.006	−.004	−.003	−.012
Scale factor a_T	1.087	1.114	1.064	1.144	1.100	1.077
Null over scale, query RMS	.120	.114	.130	.133	.116	.098
Amplified minus native NLL	.255	.039	.084	.050	.042	.166
Centered minus frozen NLL	.113	.000	−.003	.001	−.002	−.003
Preconditioned minus frozen NLL	.003	.002	.002	.005	.002	.002
Rotated partitions, ERM checkpoints						
H, partition 1	.169	.075	.082	.139	.131	.120
H, partition 2	.106	.048	.055	.056	.068	.067
D, partition 1	−.012	−.003	−.004	−.004	−.002	−.010
D, partition 2	−.007	−.002	−.003	−.001	−.001	−.005
Centered minus frozen, partition 1	.445	−.016	.018	−.006	.055	−.097
Centered minus frozen, partition 2	.090	.014	.002	.020	−.001	−.039
IRM stratum						
Adaptation feature rank (centered)	1	1	deg.	1	4	2
Frozen AUROC	.944	.884	.498	.936	.860	.835
Matched-rate AUROC	.056	.884	.502	.064	.763	.757
Recalibrated NLL	.332	.444	.644	.332	.429	.493

Per-seed facts behind the drift, predictability and centering statements. At the first-step-matched rates, ten BN-affine entropy steps change the query positive-prediction fraction from .066, .202, .100, .167, .097 and .128 (seeds 42 to 47) to .028, .203, .087, .175, .091 and .128: it falls in seeds 42, 44 and 46, rises slightly in 43 and 45 and is unchanged in 47 (the all-parameter subset

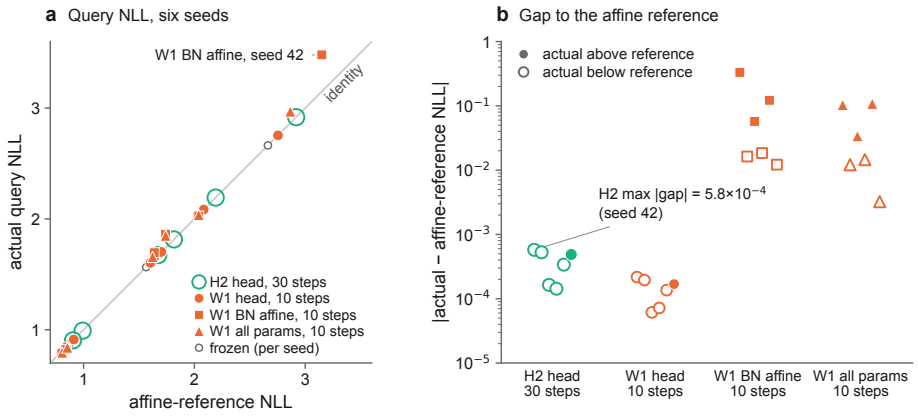


Figure 3: Affine reproduction of head-only harm. (a) Actual query NLL against the NLL of the label-free affine reference of the initial score, per seed, for the hospital-2 head arms (30 steps at .01) and the ten-step matched arms; the identity line is drawn. (b) Absolute gap between actual and reference NLL on a log axis: the head arms lie below 10^{-3} nats (largest 5.8×10^{-4}, seed 42); the BN-affine and all-parameter entropy arms, which change the representation, lie between 3×10^{-3} and .3 nats.

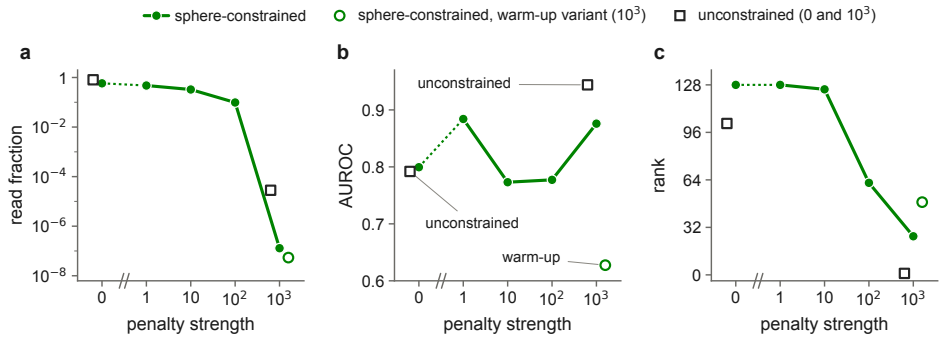


Figure 4: Penalty-strength sweep, seed 42, hospital-2 roles (Table 12): (a) the fraction of adaptation feature energy the head reads, (b) query AUROC and (c) the number of feature directions above 10^{-8} of the top eigenvalue, against IRMv1 penalty strength for the sphere-constrained family, its ten-epoch warm-up variant and the two unconstrained models. One seed; descriptive.

behaves alike). At the common rate .01 the BN-affine fraction is at most .052 in five seeds and .227 in seed 45, and the all-parameter fraction is zero in every seed. Label-free BN-affine entropy improved two checkpoints at matched rates (seed 43: NLL $-.020$, AUROC $+.004$; seed 45: NLL $-.044$, AUROC $+.008$) and seed 45 at the common rate (NLL .865 to .616, AUROC .860 to .947); we report these as instances, not as useful adaptation. Within the frozen model's predicted negatives, the representation-change term ranks positives above negatives with AUROC .889, .900 and .731 in the rising seeds 43, 45 and 47 and .290, .076 and .078 in the falling seeds. In the predictability analysis, the isotonic baseline alone has negative held-out R^2 in four seeds (42, 44, 46 and 47), so "beyond an isotonic function" partly measures a weak baseline; the ridge on the head-null coordinates reaches held-out R^2 of .18, .26, .75, .69 and .40 in seeds 42 to 46 and $-.45$ in seed 47, so the five-of-six count also holds against a zero baseline. For centering, the sealed matched center pair rescales the native endpoint's logit displacement to the centered arm's smaller adaptation radius; the centered arm's query NLL is below that rescaled native arm's in every seed (by .018 to .158), so on the original panel the centered arm's advantage is not only its smaller movement, although under patient rotation it is not stable (Table 9).

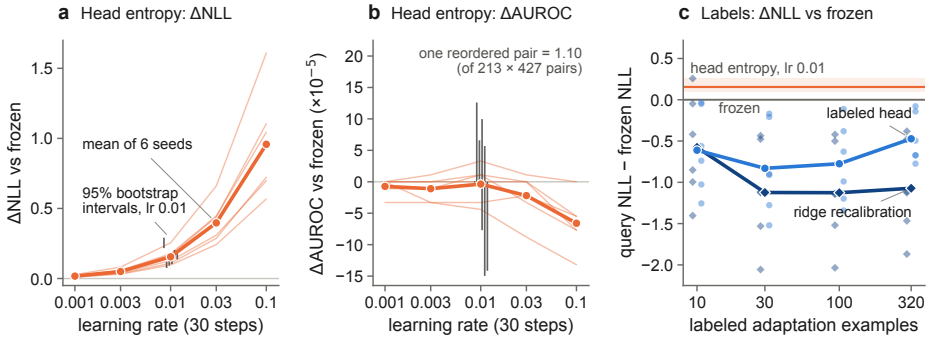


Figure 5: Head-only analyses on the sealed hospital-2 exports (six ERM seeds). (a) Query NLL change of 30 head-entropy steps against the learning rate, per seed and mean, with the 95 percent paired bootstrap interval at rate .01; (b) the corresponding AUROC change in units of 10^{-5}, at most 12 reordered pairs of the 213×427; (c) query NLL minus frozen NLL for the ridge recalibration and a ten-step labeled head fitted on the first n adaptation labels in frozen order, against n, with the frozen level and the head-entropy harm at rate .01 as references. Post hoc analyses; the labeled series use adaptation labels only.

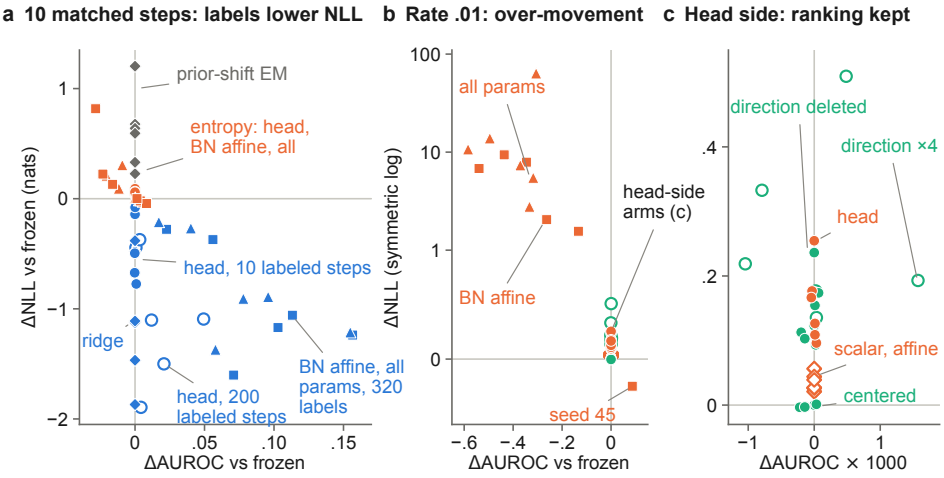


Figure 6: Query loss change against ranking change on the hospital-2 panel, every seed of every arm (six ERM checkpoints). Shape marks what is trained (circle, head; square, BN affine; triangle, all parameters; diamond, a score-only map) and colour the family (orange, label-free entropy; blue, labeled; gray, the transductive prior-shift EM control; aqua, hospital-2 head variants). (a) Ten first-step-matched steps: every labeled arm lowers NLL in every seed, and the BN-affine, all-parameter and 200-step head labeled arms also raise AUROC in every seed. (b) Thirty steps at rate .01 (symmetric-log NLL axis, linear within ± 1): BN-affine entropy moves to the upper left in five seeds and all-parameter entropy in six; seed 45's BN-affine point improves on both axes. (c) The head-side arms of (b) with AUROC changes multiplied by 1000: every arm leaves the ranking within 2×10^{-3}; the head arm coincides with the hospital-2 native head arm, which is therefore not drawn separately; the centered arm stays within .004 of frozen NLL except in seed 42 ($+.113$).

Thirty matched steps and score-only controls. Repeating the subset comparison for 30 first-step-matched steps on A100 GPUs (Table 13), two score-only entropy controls that update only a scale, or a scale and an offset, on the frozen logit reproduce the head-only arm's NLL change to three decimals ($+.154$ and $+.155$ against $+.155$) with AUROC, positive fraction and reference retention unchanged. BatchNorm-affine entropy keeps the ten-step sign split and widens it: AUROC falls in seeds 42, 44 and 46 by .049, .043 and .042 (ten steps: .028, .016, .023) and rises in 43, 45 and 47 by

Table 10: Parameter-subset adaptation on the hospital-2 panel: six-seed means of query NLL and AUROC after ten plain SGD steps at label-free first-step-matched rates (Section 6.1). Frozen: NLL 1.594, AUROC .834. Labeled arms use the 320 adaptation labels at the same subset rates and are privileged comparators, not ceilings; the 200-step head arm is a labeled reference; the two-parameter ridge recalibration of the frozen score is a post hoc labeled comparator specified after the prespecified temperature arm diverged numerically in every seed; the prior-shift arm is transductive on query inputs with a declared source prior of .5.

	Entropy (label-free)		Labeled	
Trainable subset	NLL	AUROC	NLL	AUROC
Head	1.650	.834	1.122	.834
BN affine	1.779	.825	.642	.921
All parameters	1.691	.828	.784	.908
Frozen	1.594	.834		
Prior shift, EM (transductive)	2.206	.834		
Head, 200 labeled steps			.527	.849
Affine recalibration of the frozen score (ridge)			.522	.834

Table 11: Parameter subsets at the common numeric rate .01 for 30 plain SGD steps with saved BN statistics (one A100 job per seed): query NLL / AUROC, and for the two representation-changing arms the source-reference (R) AUROC. Scalar and affine denote the score-only entropy controls. The endpoint mixtures that shrink every arm to the smallest adaptation radius stay within .01 AUROC of frozen in every seed and are omitted. Seed 42's all-parameter arm reached an exactly zero float32 gradient after 26 steps.

Seed	42	43	44	45	46	47
Frozen	2.663/.792	.808/.929	1.565/.824	.865/.860	1.638/.797	2.026/.802
Scalar	2.719/.792	.829/.929	1.591/.824	.892/.860	1.682/.797	2.065/.802
Affine	2.720/.792	.829/.929	1.592/.824	.892/.860	1.683/.797	2.065/.802
Head	2.918/.792	.904/.929	1.674/.824	.992/.860	1.815/.797	2.193/.802
BN affine	10.56/.446	2.367/.796	3.615/.561	.616/.947	8.445/.258	11.41/.365
All params	66.32/.486	6.294/.611	4.346/.491	11.56/.276	8.994/.426	15.87/.306
BN affine, R AUROC	.993	.969	.998	1.000	.999	.994
All params, R AUROC	.351	.946	.989	.955	.997	.853

.009, .020 and .007; the step-10 values of this GPU run reproduce the CPU ten-step study to three decimals. The representation-change term ranks against the label in the falling seeds (AUROC .06 to .34) and with it in the rising seeds (.85 to .91). Reference retention is unchanged under every entropy subset at matched movement ($|\Delta\mathrm{AUROC}| < 10^{-3}$), so the reference damage at the natural rate is a magnitude effect. The labeled BN-affine and all-parameter arms raise AUROC in every seed at steps 1, 10 and 30 (30 steps: $+.033$ to $+.168$ and $+.028$ to $+.165$; at the protocol's ten steps the means are $+.087$ and $+.074$). Matching held within 5 percent except seed 45's all-parameter arm (11 percent), as in the ten-step study; two seeds ran on 80 GB A100 variants with bit-identical initial outputs.

G CONTROLLED PIXEL-LEVEL REPLICATION WITH A KNOWN CUE

Design and result. To test the retained-versus-read distinction where the cue is known, thirty fresh pixel models (a 1568-256-128-1 ReLU network, no normalization layers; ten initialization seeds each of ERM, IRMv1 with penalty weight 100 after a ten-epoch warm-up, and paired-color ERM that renders each training image in both colors and averages the two losses) were trained for 200 epochs on a two-color MNIST law: the target is $\mathbf{1}[\text{digit} \geq 5]$ with 25% label noise, and the color agrees with the noisy target with probability .9 and .8 in the two source environments. Adaptation used 2,048 unlabeled images shifted four pixels with color agreement .9; the query is the 10,000 test images, shifted, scored as an exact expectation under color agreements $p \in \{.9, .5, .1\}$ (the .1 law reverses the cue). Five arms started from each final source state with first-step movement matched on 1,024 clean reference images: head-only and all-parameter entropy descent, and the same two

Table 12: Penalty-strength sweep, seed 42, eval mode, hospital-2 roles (single seed, one architecture family; descriptive). The sphere-constrained family ("Sphere") fixes the feature norm at 11.31 and normalizes the head; the last two rows use the unconstrained general-linear family, and the warm-up row trains the sphere-constrained model for ten epochs before applying the penalty. Logit standard deviation is on the adaptation panel; read fraction is the fraction of adaptation feature energy along the head; directions count eigenvalues above 10^{-8} of the top one; AUROC and NLL are on the query panel.

Family	IRM strength	Logit s.d.	Read fraction	Directions	AUROC	NLL
Sphere	0	5.9	.581	128	.799	1.574
Sphere	1	5.9	.474	128	.884	.913
Sphere	10	4.0	.327	125	.773	1.285
Sphere	100	2.7	.098	62	.777	.413
Sphere	1000	.0010	1.3×10^{-7}	26	.876	.692
Sphere, warm-up	1000	.0021	5.4×10^{-8}	49	.627	.693
Unconstrained	0	13.2	.809	102	.792	2.663
Unconstrained	1000	.0007	.00003	1	.944	.693

Table 13: Thirty first-step-matched SGD steps on the hospital-2 panel (six ERM seeds; step 30 minus frozen; A100 run). Score-only controls update the frozen logit's scale (α) or scale and offset (α, b). Per-seed AUROC changes in seed order 42 to 47; R is the source-hospital reference.

Arm	ΔNLL	ΔAUROC	ΔAUROC per seed	Δpos. frac.	R ΔAUROC
Head entropy	+.155	.000	all .000	.000	.000
Scalar control (α)	+.154	.000	all .000	.000	.000
Affine control (α, b)	+.155	.000	all .000	.000	.000
BN-affine entropy	+.376	−.016	−.049 +.009 −.043 +.020 −.042 +.007	−.013	.000
All-parameter entropy	+.264	−.012	−.033 +.007 −.031 +.017 −.033 +.001	−.012	.000
BN-affine, 320 labels (30 steps)	−1.121	+.105	+.086 +.033 +.125 +.078 +.168 +.143	+.120	.000
All parameters, 320 labels (30 steps)	−1.021	+.094	+.073 +.028 +.113 +.065 +.165 +.122	+.103	.000
Head, 200 labeled steps	−1.067	+.015	+.004 +.001 +.049 +.003 +.012 +.021	+.010	.000
Ridge recalibration	−1.072	.000	all .000	+.066	
Prior-shift EM	+.612	.000	all .000	−.040	

subsets with the 2,048 noisy labels, for 30 SGD steps. Table 14 reports means over ten seeds; every sign statement holds in all ten seeds unless stated.

Figure 7 shows the retained-versus-read picture on 1,024 shifted probe rows: every model retains color in its features (RMS color difference relative to the feature RMS .71 for the paired models against 1.6 for ERM and 2.0 for IRMv1), while the paired heads' mean absolute cosine with the color direction is .18 against .79 for ERM and .87 for IRMv1. Paired-color models are invariant in accuracy across the three laws (.636, .633, .629 unshifted) although their logits still respond to color (paired-color logit contrast RMS 14 on the reference against 38 for ERM). Label-free entropy raises NLL in 47 of 49 eligible seed-arm cells at the reversed law (48 of 49 at the neutral law) and moves configuration-mean accuracy by under one point everywhere except IRMv1's all-parameter entropy ($+.025$ aligned, $-.040$ reversed); head-only entropy could not be rate-matched within the preregistered rate ceiling in 8 of 10 ERM and 3 of 10 paired models, whose saturated logits weaken its first-batch gradient three- to twentyfold. The labeled head-only arm lowers NLL at every law in all 30 cells, by .06 to 18.7 nats, including the reversed one, with accuracy unchanged: recalibration of an overconfident source model, not adaptation. The one recruitment is the all-parameter labeled arm on the paired models: accuracy $+.106$ under the aligned law and $-.106$ under the reversed law (ranges $+.083$ to $+.132$ and $-.080$ to $-.130$), with the shifted-query color contrast changing by an RMS of 24 to 38 in every seed and its magnitude rising in nine of ten (by -6 to $+22$, mean $+9$), whereas the head-only labeled arm on the same models moves accuracy by under 1.5 points. The cue is retained by the representation, unread by the head, unreachable by entropy or by 30 head-only labeled steps, and recruited once the representation trains on labels. IRMv1 in this recipe reduced color reliance without collapsing scale (unshifted reversed-law accuracy $.461$ against $.293$ for ERM and $.629$ for paired-color ERM), the non-collapsed regime of the penalty. Lowering the paired models' head scale by up to thirtyfold, so that head-only entropy moves the reference logits by 15 to

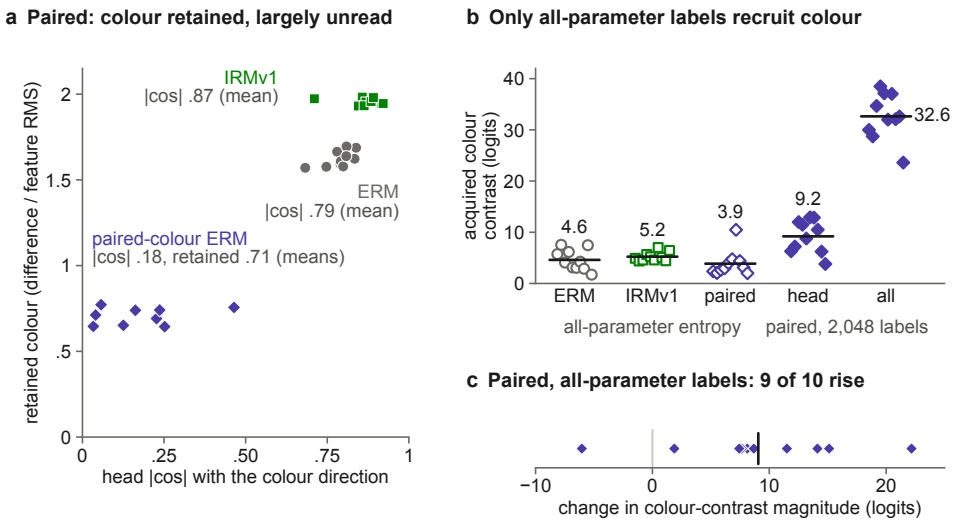


Figure 7: Retained versus read in the controlled-MNIST family (ten seeds per source configuration). (a) Frozen source models on the 1,024 shifted-query probe rows: the head's absolute cosine with the principal direction of the colour difference $d_i = \phi(x_i \text{ in colour } 1) - \phi(x_i \text{ in colour } 0)$, against the RMS colour difference relative to the feature RMS; gray circles, ERM; green squares, IRMv1; violet diamonds, paired-colour ERM. Every model retains colour in its features, and the paired heads largely do not read it (means: cosine .18 and ratio .71, against cosine .79 for ERM). (b) Acquired colour contrast on the shifted query after 30 steps (RMS over rows of the change in the colour-0 minus colour-1 logit), per seed, with the mean: all-parameter entropy in each configuration (open markers) and, on the paired models, head-only and all-parameter training with the 2,048 adaptation labels (filled). (c) Paired models, all-parameter labels: change in the colour-contrast magnitude on the shifted query per seed (bar: mean $+9.1$); it rises in nine of ten seeds.

33 percent instead of 2 to 5, does not let it recruit the retained cue (reversed-law accuracy within .6 points and the head's alignment with the color difference unchanged at every scale and at 300 steps), whereas the same scale reduction lets head-only entropy amplify the cue that ERM models already read ($+.009$ aligned and $-.012$ reversed at scale .3, ten of ten). A CPU rerun of all thirty cases with the same code and data reproduces every ten-seed statement above (paired all-parameter labeled arm $+.112$ and $-.107$, ten of ten seeds each) although individual seeds differ between devices by up to .12 in accuracy on single views, so only seed means and sign counts are reported. This is a toy cue intervention on one fixed data partition (ten seeds are initialization replications); it identifies a controlled mechanism and establishes neither real-world causal recovery nor a safeguard.

Table 14: Controlled-MNIST family: shifted-query accuracy under the aligned ($p = .9$) and reversed ($p = .1$) color laws, mean over ten seeds; Δ columns are step 30 minus frozen. Color logit contrast RMS on shifted query rows: frozen, and the RMS of its per-row change at step 30 (unsigned). Entropy rows are means over the rate-matched seeds only; head-only entropy rows are omitted because it could not be rate-matched in 8 of 10 ERM and 3 of 10 paired models (text).

Source	Arm	Acc. p=.9	Acc. p=.1	ΔAcc. .9	ΔAcc. .1	Contrast frozen	Acquired contrast (RMS)
ERM	frozen	.844	.185			161	
ERM	all-param. entropy			$+.006$	$-.008$		4.6
ERM	head, labeled			$+.004$	$-.006$		43.7
IRMv1	frozen	.834	.238			19	
IRMv1	all-param. entropy			$+.025$	$-.040$		5.2
Paired	frozen	.551	.544			47	
Paired	all-param. entropy			$+.001$	.000		3.9
Paired	head, labeled			$+.008$	$-.012$		9.2
Paired	all-param., labeled			$+.106$	$-.106$		32.6

H PREREGISTERED REPLICATION ON WATERBIRDS-95

This appendix reports a preregistered replication of the native hospital-2 studies of Section 6 on a second natural-image group-shift dataset, run on the day of submission from a frozen preregistration with six freshly trained source models and scored after a global seal. Every prediction was registered before any source model was trained; every reading of the registered text used by the scorer was fixed before the seal or before any query label was opened (five dated corrections in a hash chain; reproducibility statement); an independent auditor recomputed every registered quantity from the sealed arrays and the label file and confirmed every verdict.

Dataset, laws and roles. Waterbirds-95 (Sagawa et al., 2020): 11,788 landbird and waterbird photographs from CUB pasted on land or water backgrounds from Places. The positive class is waterbird, the minority. The release fixes the training law, where the background agrees with the class in 95.0 percent of images, and the balanced validation and test splits (50 percent); a reversed law was sampled from the test split with fixed group quotas that flip the training shares (25, 467, 141 and 7 rows for landbird on land, landbird on water, waterbird on land and waterbird on water). The roles mirror the hospital-2 design: reference rows R (384 training images), adaptation rows A (320 validation images), query rows Q (640 test images in five hash clusters of 128, standing in for the five query patients) and the reversed-law rows V (640 further test images). R, A and Q were fixed by SHA-256 rank before any label was read; V was then drawn from the remaining test rows by group quota and hash rank. No fit stage reads a label of R, Q or V; the labeled comparators read the 320 labels of A inside their own fit stage.

Source models and arms. Six ResNet-50 models (ImageNet initialization; the paper's 2048–256–128 projection and bias-free $128 \to 1$ head) were trained by AdamW at rate 10^{-4} with weight decay .01, batch 32, 20 epochs at a constant rate, horizontal flips, float32 on Apple Metal, seeds 1001 to 1006, final epoch kept, with no validation or query forward during training; every run reproduced the values the auditor had computed prospectively (permutation and flip hashes, initial-state digests, export input hashes). The frozen models have query AUROC .90 to .95, query NLL .49 to .92, positive fraction .225 to .334 and waterbird recall .71 to .82; none is at the registered .99 ceiling, so no seed is excluded. The arms are the paper's recipes (Appendix F for the Camelyon17 versions); the network arms run one seed at a time on the device of the export: the head-only study in float64 on CPU from the exported features, after a one-seed check that a cached-feature head arm reproduces the full forward pass on the device (float32 exactly; the float64 head arithmetic within 10^{-6} absolute plus 10^{-5} relative, largest gap 3.6×10^{-6}); the first-step-matched representation arms for 30 steps with snapshots at steps 1, 10 and 30, whose step-10 snapshots are the ten-step values (every verified first-step ratio lay within $1 \pm .05$: BN-affine 1.000 to 1.001, all parameters .992 to 1.035, so no matched-rate prediction is suppressed); and the common-rate arms (SGD at .01 for 30 steps). The registered A100 profile guard became a per-step resource stop on Metal; no stop fired.

Outcomes against the registered predictions. Table 15 lists every registered prediction with its verdict under the readings fixed before scoring, and Figure 8 shows the anatomy of Figure 2 on this dataset. The account-level rules of the preregistration give: A1 (head-only entropy is sharpening of the current score) not refuted; A2 (deleting a selected unread direction does not reduce the harm beyond movement) not refuted; A3 (representation-reaching entropy moves the ranking with a sign tied to the frozen minority recall through majority drift, while labels on the same parameters raise it) not refuted; the over-movement reading of the common-rate study refuted. In words:

- *The head-only account replicates in full.* Head-only entropy raises query NLL in every seed (by .021 to .066, mean .043) with the ranking and accuracy unchanged ($|\Delta\mathrm{AUROC}| \leq 5.4 \times 10^{-5}$; accuracy within .002); the label-free affine map of the initial score reproduces the native NLL within 3×10^{-4} nats in every seed on Q and within 4.1×10^{-4} on the reversed-law rows (S1); the exact split gives a_T 1.04 to 1.08 and ρ .22 to .31 (P2); deleting the head-null mean direction changes NLL by $-.005$ relative to random deletions on average and, at matched endpoint movement, is worse than random in all six seeds (P3; the native result was five of six); the harm is monotone in the rate with the ranking unchanged at every rate (S5), survives both patient-style rotations in all 18 cells (S4), and the 30-step head arm of the matched study equals the head-only study within 4.3×10^{-8} nats (S8). Score-only scale and affine controls at the common rate raise NLL by less than the head

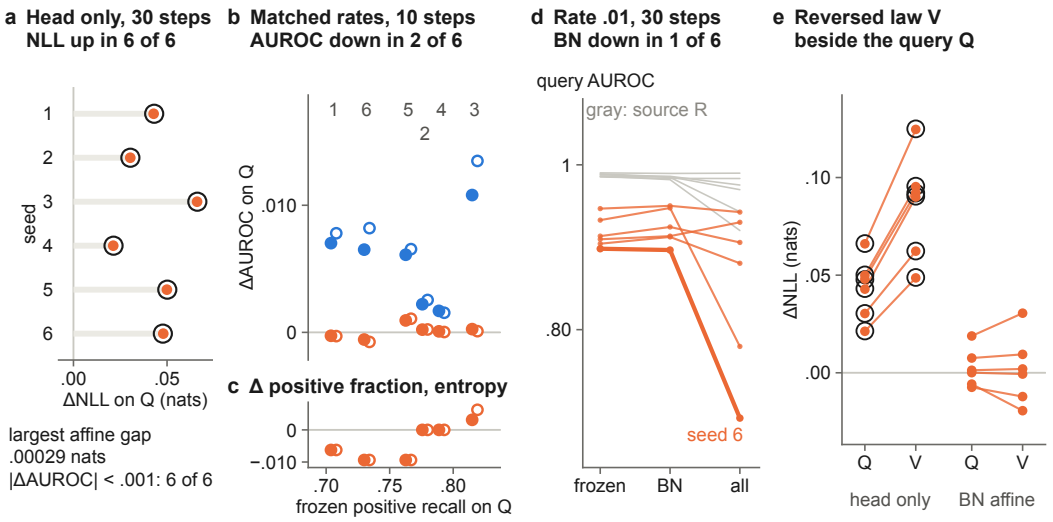
Table 15: Waterbirds-95 replication: every registered prediction (preregistration Section 5, readings fixed in Corrections 4 and 5) with its verdict from the sealed score run; six seeds, no exclusion (no seed at the AUROC ceiling, no degenerate frozen prediction, no missed first-step match). "k of n" counts seeds satisfying the criterion; verdicts follow the allowed-failures reading of count criteria.

ID	Criterion (short)	k of n	Verdict
P1a–d	head-only: NLL up, affine within 10^{-3}, ranking and accuracy unchanged	6 of 6 each	held
P2	exact split $a_T > 1$, $\rho < 1$	6 of 6	held
P3a, P3b	deletion's mean advantage above the $-.01$ criterion; at matched movement	mean $-.005$; 6 of 6 positive	held
P4a	$\lvert\Delta\mathrm{AUROC}_{\mathrm{BN}}\rvert \geq 10^{-3}$ at ten matched steps	0 of 6	not held
P4b, P4c	sign split at frozen recall; all-parameter sign agrees	$\lvert F\rvert = 2$, $\lvert U\rvert = 4$; 6 of 6	held
P5	majority drift in falling seeds (D11 below .5; above .5 in rising)	6 of 6	held
P6a, P6b	labeled BN-affine and all-parameter up; labeled head recalibrates	6 of 6	held
P6c	200 labeled head steps raise AUROC	2 of 6	not held
P7	prior-shift EM raises NLL	5 of 6	not held
P8	ridge available and lowest NLL of the ten-step arms	6 of 6; lowest	held
P9a, P9b	common rate: head sharpens; score-only controls below head	6 of 6; means	held
P9c	common rate: BN-affine AUROC down in ≥ 5, one-class in ≥ 5, R kept	1, 0, 6 of 6	not held
P9d	common rate: all-parameter one-class in every seed	1 of 6	not held
P10a	score-only controls equal head at 30 steps within 10^{-3}	NLL gap .0013, .0015	not held
P10b	BN-affine sign persists and widens at 30 steps	6 of 6	held
P10c	every entropy arm keeps the R ranking within 10^{-3}	5 of 6 (all-parameter)	not held
P10d	labeled arms raise AUROC at steps 1, 10 and 30	5 of 6	not held
P11a–d	head-null coordinates predict the representation term	6 of 6; falling 2 of 2	held
S1, S2	reversed law: affine and ranking hold; drift dominates background	6 of 6; 2 of 2	held
S4, S5, S8	rotations 18 cells; rate sweep; matched-versus-head identity	all	held

arm (means .014 against .043; P9b) and, at 30 matched steps, reproduce the head arm's mean NLL to .0015 and .0013 nats, just outside the registered 10^{-3} (P10a not held on that part; its ranking parts hold in every seed).

- *Representation arms replicate in sign, not in size.* At ten first-step-matched steps, BatchNorm-affine entropy leaves the query ranking within the 10^{-3} unchanged-ranking bound in every seed (P4a not held: -5.6×10^{-4} to $+9.4 \times 10^{-4}$) and all-parameter entropy in five (largest $+.0011$), below every Camelyon17 BN-affine change (.001 to .028 in magnitude). Within those small changes the registered structure holds: the two falling seeds (1001 and 1006) have the two lowest frozen waterbird recalls (.706 and .732 against .765 to .817; P4b), while frozen confidence and frozen NLL do not separate them; the all-parameter sign agrees in every seed (P4c); the representation-change term ranks positives below negatives within the frozen model's predicted negatives in the falling seeds (AUROC .38 and .47 against .63 to .78 in the rising seeds; P5); drift dominates the background contrast in both falling seeds and no seed shows background recruitment under the registered group contrast (S2); the sign persists and widens at 30 steps in every seed (P10b); and the initial head-null coordinates predict the representation-change term out of sample in all six seeds (incremental R^2 .04 to .26, total .26 to .66, Gaussian null negative in every seed; P11). Labels on the same parameters and rates raise AUROC by .002 to .013 and lower NLL in every seed at ten steps (P6a), and the labeled head lowers NLL with the ranking unchanged (P6b); but 200 labeled head steps raise AUROC in only two of six (P6c not held), and the labeled representation arms are not above zero at every one of steps 1, 10 and 30 in one seed (P10d not held by seed 1004 at step 1).

- *The common-rate collapse does not replicate.* At rate .01 for 30 steps, BatchNorm-affine entropy raises query AUROC in five of six seeds (by $-.002$ to $+.015$) with positive fractions .12 to .34, and all-parameter entropy predicts one class in one seed only (positive fractions 0, .002, .09, .14, .19 and .23), so P9c and P9d are not held and the over-movement reading is refuted on this dataset; the source-reference ranking survives in every seed (R AUROC at least .982). The Camelyon17 checkpoints that collapsed at this rate started from frozen NLL .81 to 2.66; these start from .49 to .92.
- *Recalibration.* The ridge recalibration is available in every seed and has the lowest six-seed mean query NLL of the ten-step arms (P8); prior-shift EM raises NLL in five of six seeds and lowers it by 4×10^{-5} in the sixth (P7 not held by that seed), shifting the prior up in four seeds and down in two.



orange: label-free entropy; blue: 320 labels; filled: BN affine; open: all parameters; ring: affine map; label k is seed 1000 + k

Figure 8: Waterbirds-95 replication of the native anatomy of Figure 2 (six fresh ERM ResNet-50 models, seeds 1001 to 1006, labeled by their last digit; query Q: 640 test images with balanced backgrounds; V: 640 test images with the background association reversed; the positive class is waterbird). (a) Head-only entropy, 30 SGD steps at rate .01: query NLL change per seed (dots) and that of the label-free affine map of the initial score (rings); the largest gap is .00029 nats and $|\Delta\text{AUROC}| < .001$ in 6 of 6 seeds. (b) Ten first-step-matched steps: AUROC change against the frozen model's positive recall on Q for BN-affine (filled) and all-parameter (open) updates trained by entropy (orange) or with the 320 adaptation labels (blue); entropy lowers AUROC in 2 of 6 seeds and both labeled subsets raise it in 6 of 6. (c) Change in the fraction of Q predicted positive under entropy. (d) Rate .01 for 30 steps: Q AUROC of the frozen model and of BN-affine and all-parameter entropy, one line per seed (bold: seed 1006, the one whose BN-affine AUROC falls), with the reference-role AUROC in gray; BN-affine entropy lowers it in 1 of 6 seeds and all-parameter entropy in 5 of 6. (e) NLL change on V beside Q for the head-only arm, with its affine rings, and for the ten-step BN-affine entropy arm, one line per seed.

What this replication is and is not. It is a same-day, preregistered, independently audited test of the paper's native account on a second dataset with fresh source models, at zero metered cost on a single Apple Metal device. It supports the head-only account without qualification and the majority-drift reading of representation updates only as a sign pattern among changes at or below the paper's own unchanged-ranking bound (BN-affine below it in every seed, all-parameter in five); it contradicts the generality of the common-rate collapse, consistent with the higher initial NLL of the Camelyon17 checkpoints without establishing it as the cause. It is not evidence of useful label-free adaptation or of a safeguard, and its seeds replicate training noise only (one architecture family, one recipe, one query panel, one day). Receipts, seals and the preregistration

are listed in the reproducibility statement (preregistration 0316e635..., launch A e99f5f1b... with run seal bbb2dc5a..., launch B 46ef4b36..., head-only study db1fac8c..., global seal 02298593..., score run wbS-20260926-r1 with predictions table e305c105..., result audit 1ece9675... with receipt 68045d60...).

I Additional discussion of response, risk and native outcomes

The budget crossing shows why collapse suppression is an incomplete explanation of protection. Centering at the native rate greatly reduced reference movement. Forcing centered updates to follow the raw learner's larger reference schedule removed collapse while increasing target risk in all four pairs. This intervention changes the learner: its displacement multipliers reached 38.83, and equal movement on the neutral reference need not equal movement on the reversed target. The experiment supports a claim about the specified diagnostic update, not an estimate of the typical effect of native centering.

Response and risk measurements also require separate interpretations. The total nuisance response can fall while the norm remaining after an affine fit rises, because the residual can cancel part of the original response. The saved affine-logit references reproduce a large loss increase over frozen predictions, while failing to reproduce the full centering and targeted-restriction contrasts. These are signed comparisons with one fitted reference family. They establish neither a nonnegative allocation of harm to mechanisms nor the impossibility of explaining an outcome with a different function of the old prediction.

No terminal reversed-target condition in the head pilot met the declared useful-accuracy or joint-gain criterion. The subsequent native feasibility study also failed under its fixed recipe. Changing BN statistics caused a large loss increase before learning; affine entropy updates added harm. Retaining saved statistics still increased proper loss in both checkpoints, with opposite interaction signs and an accuracy increase in one checkpoint. A normalization-state comparison is therefore distinct from an initially unread-component comparison, and improved accuracy alone is distinct from the declared loss endpoint.

The favorable secondary restriction is similarly bounded. It changes component access and lowers risk under the forced centered learner, but the targeted predictor remains substantially worse than frozen. Random restrictions match rank and reference movement; they do not match every form of feature energy, gradient drive, or useful information removed. The selected direction is associated with a controlled nuisance response but is not a pure semantic channel. The comparison therefore identifies the effect of this intervention on this learner without uniquely identifying nuisance-only mediation.

Supervised readouts on the hospital-1 checkpoints meet the same boundary: readouts fitted with 320 labels across four patient folds and scored on 512 query rows did not beat an affine recalibration of the frozen score in both seeds (seed 42 lost by .011 nats with identical confusion matrices; seed 43 gained .022).

Differences between retained architecture families include source-training and clipping choices, and orthogonal weight layers do not make the entire implemented encoder an isometry.

These follow-ups narrow the empirical claim to collapse suppression, component access and risk under controlled update geometry. They do not establish a prospective monitor or useful target adaptation. Fresh source-data replication beyond the same-day Waterbirds study of Appendix H remains outstanding. The separate native normalization-parameter studies in Section 6 are complete, but fail to demonstrate the intended useful-learning opportunity.

A final head-null projection applied after all adaptable operations is an exact no-op when the final head is fixed, so it cannot supply the causal recruitment test described in Section 7.

J Proofs for Section 3

Proof of Theorem 1. (i) The construction gives $\mathbb{E}[Y \mid C] = \sigma(cC)$, independently of r. Taking conditional expectation makes $G_r = 0$. (ii) $\mathbb{E}\,\|\phi_a\|^2 = \mathbb{E}C^2 + a^2\mathbb{E}S^2 = 1 + a^2$ because $C, S \in$

$\{-1, 1\}$; the initial logit cC, the label Y and the environment parameter r do not involve a. (iii) Initially $|f_0| = c$, so $\sigma'(f_0) = h$ and $H'(f_0) = -chC$. Therefore the gradient is $-ch(1, a\rho)$, giving the updated head $(c(1 + \eta h), \eta ach\rho)$; multiplication of its second coordinate by a gives (4). Since $C, S \in \{-1, 1\}$, the sign of $\theta_C C + \theta_S S$ equals the sign of S on every input exactly when $\theta_S > \theta_C$, which is the stated threshold; the prediction then equals $S = TB_s$, which is correct exactly when $B_s = 1$, an event of probability r_e; and $\mathbb{P}(S = 1) = 1/2$ because C is uniform on $\{-1, 1\}$ and independent of (B_c, B_s). □

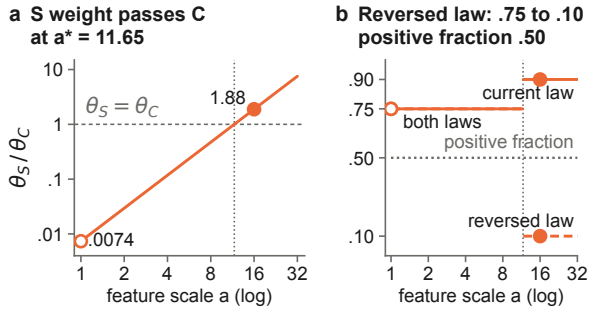


Figure 9: The threshold of Theorem 1 in the example of Section 3 ($p = .75, r = .9, r_e = .1, \eta = .1$). (a) The ratio $\theta_S/\theta_C = \eta a^2 h\rho/(1+\eta h)$ of the coefficients after one population entropy step, against the feature scale a (both axes logarithmic); it crosses 1 at $a^* = \sqrt{(1+\eta h)/(\eta h\rho)} = 11.65$. The open and filled markers are the scales 1 and 16 of Figure 1 (ratios .0074 and 1.88). (b) Accuracy of the updated classifier under the current and the reversed law, and its positive fraction: below a^* the sign follows C and the accuracy is $p = .75$ under both laws; above it the sign follows S, the accuracy is $r = .90$ under the current law and $r_e = .10$ under the reversed one, and the positive fraction stays .50. At a^* two inputs tie at logit zero; the accuracy curves omit that point, while the positive-fraction line passes through it at .50, its value on both sides (a strict positive rule at the tie itself would give .35 and .15).

Support of the numerical example. All four raw input states have positive probability. The raw second moments under $r = .9$ and $r_e = .1$ have correlations $+.4$ and $-.4$, with eigenvalues $.6, 1.4$; absent raw-input support is not needed for this example.

Table 16: Exact population quantities after one head update, with $p = .75$, $r = .9$, $r_e = .1$, and $\eta = .1$. Initial accuracy is .75 and initial logistic risk is .562335 on both laws. This is arithmetic for the construction, not a learned-checkpoint experiment.

Scale	Current accuracy	Reversed accuracy	Reversed risk	Positive fraction
1	.75	.75	.564840	.50
16	.90	.10	1.627561	.50

K Linear invertibility and unread energy

Proposition 2 (A fixed-readout family)**.** *Let $w_0 \neq 0$, $u \perp w_0$, and $s : \mathcal{X} \to \mathbb{R}$. Define $\phi_\alpha(x) = \phi_0(x) + \alpha u s(x)$. Then $w_0^\top \phi_\alpha(x) = w_0^\top \phi_0(x)$ for every x, α. If u is unit, $\mathbb{E}\,\|\phi_0(X)\|^2 < \infty$, and $0 < \mathbb{E}s(X)^2 < \infty$, then*

$$\mathbb{E}\,\|\phi_\alpha(X)\|^2 = \alpha^2 \mathbb{E}s(X)^2 + O(|\alpha|) + O(1).$$

Proof. The logit identity follows from $w_0^\top u = 0$. Expanding the squared norm gives the second claim; the cross term is finite by Cauchy–Schwarz. □

This elementary construction changes response magnitude without changing current predictive use. It does not establish increasing information content, optimizer reachability, or harm. Invertible rescaling preserves unrestricted recoverability, and the linear invertible version of the construction follows.

For input $x = (C, S, \ldots) \in \mathbb{R}^d$, let an invertible W_0 satisfy $w_0^\top W_0 = ce_C^\top$, with $w_0 \neq 0$ and $d \geq 2$. Choose a unit $u \in \ker(w_0^\top)$ and define $W_\alpha = W_0 + \alpha u e_S^\top$. The initial logit is cC for all α. Select the sign of u so that $e_S^\top W_0^{-1} u \geq 0$. The determinant lemma gives

$$\det W_\alpha = \det W_0 \left(1 + \alpha e_S^\top W_0^{-1} u\right) \neq 0 \qquad (\alpha \geq 0).$$

Under finite input second moments with $\mathbb{E}S^2 > 0$, expansion yields $\mathbb{E}\,\|W_\alpha X\|^2 = \alpha^2 \mathbb{E}S^2 + O(\alpha) + O(1)$. This is an expected second-moment statement. It is not pointwise growth for every input, increased mutual information, or a training-dynamics theorem. Every W_α remains invertible and therefore preserves the raw input's recoverability.

L General finite positive calibration

Suppose all environments share the law of (C, Y), with finite second moments, $\mathbb{E}[(Y - 1/2)C] > 0$, and positive weighted sign error

$$\mathbb{E}\left[|C|\,\mathbf{1}\{Y \neq \mathbf{1}[C > 0]\}\right] > 0.$$

Define $R(c) = \mathbb{E}\ell(cC, Y)$. Then

$$R'(c) = \mathbb{E}[(\sigma(cC) - Y)C], \qquad R''(c) = \mathbb{E}[\sigma'(cC)C^2] > 0.$$

Positive association gives $R'(0) < 0$. By dominated convergence,

$$\lim_{c \to \infty} R'(c) = \mathbb{E}[(\mathbf{1}[C > 0] - Y)C] = \mathbb{E}\left[|C|\,\mathbf{1}\{Y \neq \mathbf{1}[C > 0]\}\right] > 0.$$

Continuity and strict convexity imply a unique finite $c_* > 0$ with $R'(c_*) = 0$. For initial logit $c_* C$, every environment's scalar-dummy gradient is $c_* R'(c_*) = 0$.

Equal environment laws alone make their dummy gradients equal, not zero at an arbitrary chosen scale. Perfectly sign-separable labels can lack a finite positive calibration root. Empirical penalties remain identical across an exact prediction-equivalent family but need not equal their population value. These distinctions are necessary when moving from the constructed law to trained checkpoints.

M Prediction evidence and trajectory equivalence

What an initial prediction statistic observes. Consider a statistic whose representation access is restricted to initial logits, supplied labels, and fixed environment assignments. Exact prediction equivalence makes this statistic identical, even with arbitrary input queries. The restriction is about its inputs, not its scalar output: a scalar function of features or full head gradients can observe additional information. We use this boundary to interpret initial evidence; it does not exclude clone-and-update evaluation or establish that every sound audit must fail.

The update kernel. For fixed features and mean batch loss $n^{-1} \sum_i L_i(f_w(x_i))$, ordinary head descent has the exact finite-step identity

$$f_{t+1}(x) = f_t(x) - \frac{\eta_t}{n} \sum_i L_i'(f_t(x_i)) K(x, x_i), \qquad K(x, x') = \phi(x)^\top \phi(x'). \tag{6}$$

This standard identity, including its role in entropy adaptation, is explanatory background (Lee et al., 2024); we do not claim it as a new theorem. Equation (6) follows by multiplying the head update by $\phi(x)^\top$. For an objective coupling batch logits, its full logit-gradient entries replace the individual derivatives and then carry the batch averaging themselves, so the $1/n$ is not applied twice.

Initial predictions together with the relevant kernel determine the complete fixed-feature descent trajectory. A training Gram matrix alone determines training logits; target predictions also need target/training cross-kernels.

Equation (6) is exact because the head is linear in its parameters. If two systems have equal initial predictions and equal kernels on every point that can enter adaptation and evaluation, induction on updates gives equal subsequent predictions. The same statement covers deterministic objectives of the batch logits by substituting their full logit gradients $\partial L_{\text{batch}}/\partial z_i$ for $n^{-1}L_i'$. It requires identical batches, schedule and all other operations. A free bias can be represented by appending a constant feature; a normalized head, moving encoder or Adam state requires a different update analysis.

Equal training Gram matrices need not imply equal target/training cross-kernels. Conversely, for one particular trajectory, a kernel difference may be harmless because it annihilates all encountered logit-gradient vectors. Full kernel equality is a convenient sufficient condition, not a necessary condition for one realized run.

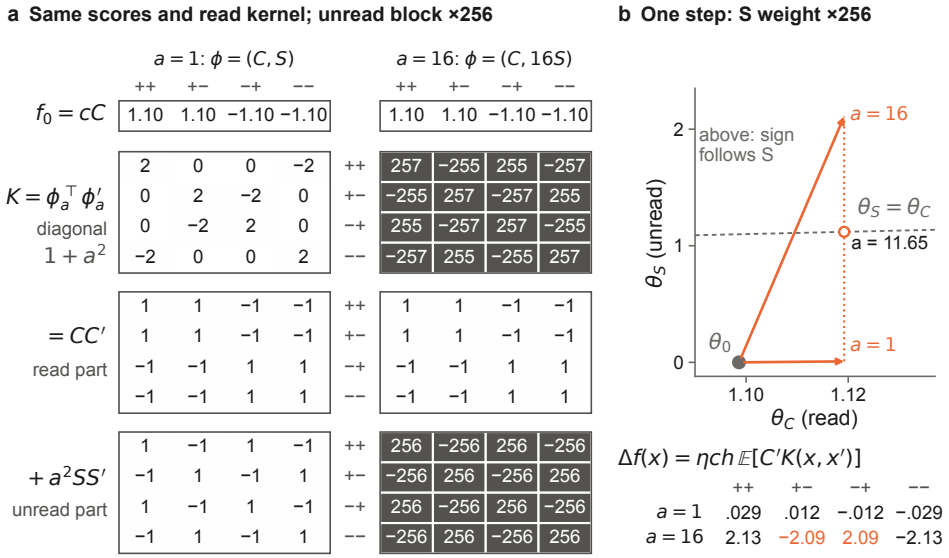


Figure 10: The update-kernel view of the construction in Section 3 ($p = .75$, $r = .9$, $\eta = .1$, $c = \log 3$, $h = .1875$, $\rho = .4$). (a) For the four inputs $(C, S) \in \{\pm 1\}^2$ (rows and columns) and $\phi_a = (C, aS)$ at $a = 1$ and $a = 16$: the initial scores $f_0 = cC$, the fixed-feature kernel $K(x, x') = \phi_a(x)^\top \phi_a(x')$ of Equation (6), and its exact split into the read part $(w_0^\top \phi)(w_0^\top \phi')/\|w_0\|^2 = CC'$ for $w_0 = (c, 0)$ and the unread part $a^2 SS'$; shade is proportional to the absolute entry on one scale. Scores and read parts are identical at both scales, the unread part differs by the factor $a^2 = 256$, and the diagonal is $1 + a^2$, the hidden energy of Theorem 1. (b) One population entropy step moves the head by $\Delta w = -\eta \mathbb{E}[H'(f_0)\phi_a] = \eta ch(1, a\rho)$; in raw coordinates the coefficients move from $\theta_0 = (c, 0)$ (gray) to $(\theta_C, \theta_S) = (1.119, .008)$ at $a = 1$ and $(1.119, 2.109)$ at $a = 16$ (orange; dotted, $1 \le a \le 16$). Above the dashed line $\theta_S = \theta_C$, reached at $a = 11.65$, the sign of the updated logit follows S on every input. The table gives the resulting logit change $\Delta f(x) = \eta ch\, \mathbb{E}[C'K(x, x')]$, the expectation over x' under the adaptation law; orange entries reverse the sign of f_0.

An elementary evidence limit follows. If two admitted systems have identical initial functions but risks $R_1 < R_2$ after a fixed procedure on the same stream and evaluation law, an initial-output-only auditor has the same transcript on both. At a threshold between R_1 and R_2, it cannot both accept the first and reject the second uniformly. A randomized auditor with common acceptance probability t has false acceptance t on the second and false rejection $1 - t$ on the first. An always-reject rule remains sound. These are standard indistinguishability consequences of the restricted observation model, not a new impossibility of all useful auditing. An auditor with access to representations, update kernels, or clone-and-update outcomes has additional information.

For the constructed pair, under either evaluation law the two systems share the same optimal action (keeping the initial head under the reversed law, taking the step under the current law). For any threshold between the two post-update risks, an auditor restricted to initial outputs cannot decide whether post-update risk exceeds it; and for the two-action choice between keeping the initial head and taking this step, with the evaluation law reversed with probability q and the adaptation law

otherwise, the two systems' expected-risk-optimal actions differ exactly for $q \in (.2113, .4907)$ (scale 1 prefers the step, scale 16 prefers to keep). A selector restricted to initial outputs must apply one action, possibly randomized, to both systems; over that interval its minimax regret is at most .0012 nats (.0014 if the action must be deterministic, attained by keeping the head), for example .0012 nats at $q = 1/4$, where the step changes expected risk by $-.0012$ (scale 1) and $+.0523$ (scale 16).

N FULL MOMENTS: INVERTIBILITY, RIDGE, AND QUERY SUPPORT

Let $M = \mathbb{E}[\phi\phi^\top]$ be an uncentered second moment and g a head gradient. If M is invertible, the update $M^{-1}g$ transforms covariantly under every invertible linear feature transformation: $M' = AMA^\top$, $g' = Ag$, and $(M')^{-1}g' = A^{-\top}M^{-1}g$. This is a standard coordinate-invariance control (Martens, 2020), not a general natural-gradient or safety claim. It is equally valid for a nonsingular empirical moment. Isotropic ridge, singular moments and off-span target queries require additional care, as follows.

In the constant-absolute-margin binary example, the entropy gradient is proportional to Mw, so exact full-moment preconditioning preserves the zero nuisance coefficient. This extra suppression property does not hold for general varying-margin data. Moreover, increasing the calibrated stable coefficient can worsen logistic risk while preserving accuracy. Coordinate invariance, nuisance suppression and useful adaptation are different properties.

For nonsingular M and invertible A, $(AMA^\top)^{-1}Ag = A^{-\top}M^{-1}g$. This establishes identical function updates when features, head and moment are transformed consistently. It is a finite empirical identity whenever the empirical moment is nonsingular; population estimation is not required for the algebra.

The implemented control instead uses a batch moment G and isotropic ridge λI, with $\lambda = 10^{-8}\,\mathrm{tr}(G)/d$. Under feature transformation, its effective original-coordinate direction is

$$A^\top(AGA^\top + \lambda_A I)^{-1}Ag = \left(G + \lambda_A A^{-1}A^{-\top}\right)^{-1} g.$$

It is generally different from $(G + \lambda I)^{-1}g$. A small ridge is not a uniform approximation near singularity. For positive definite G, a simple bound is

$$\left\|G^{-1} - (G + \lambda I)^{-1}\right\|_{\mathrm{op}} \le \frac{\lambda}{\lambda_{\min}(G)(\lambda_{\min}(G) + \lambda)}.$$

The retained Cayley encoder's affine-free LayerNorm also imposes a zero-sum feature constraint, making its ideal moment singular in the ambient coordinates. Observed near-invariance is therefore reported for the tested transformations, not inferred from a nonexistent full-rank ambient inverse.

The Moore–Penrose inverse does not restore arbitrary global congruence. For example, take

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad g = e_1, \quad A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Then $M^\dagger g = (1, 0)^\top$, whereas $A^\top(AMA^\top)^\dagger Ag = (1, 1/2)^\top$. Queries in $\mathrm{range}(M)$ agree; an e_2 query does not. More generally, when $g \in \mathrm{range}(M)$, the discrepancy lies in $\ker(M)$, so on-span predictions agree. Whether a target lies in that span is consequential. A globally fixed feature relation can make this condition benign; insufficient reference coverage need not.

For the binary construction with constant absolute margins, $g_H = -\sigma'(c)Mw$. Exact full preconditioning therefore changes w only radially. This special identity explains nuisance exclusion there, but does not extend to varying-margin examples, where the sample weights $\sigma'(f_i)$ differ. It also does not make increasing c beyond its calibrated optimum beneficial for logistic risk.

O MEAN AND INTERCEPT EFFECTS

The local entropy expansion is

$$H(z) = \log 2 - z^2/8 + z^4/64 + O(z^6), \qquad H'(z) = -z/4 + z^3/16 + O(z^5).$$

With uniformly small logits, or suitable moment and remainder bounds, a head step begins as

$$w_{t+1} \approx (I + \eta M/4)\, w_t, \qquad M = \mathrm{Cov}(\phi) + \mu\mu^\top.$$

A large mean can supply a dominant near-constant mode, but covariance directions can dominate instead. Initial alignment and later saturation also matter. The expansion does not justify a universal mean-collapse theorem.

Centering with $\widetilde{\phi} = \phi - \mu$ and fixed $b_0 = w_0^\top \mu$ preserves the initial function and fixes $\mathbb{E}_{\mathrm{ref}} f_t = b_0$ when the same population mean defines the reference. A trainable bias is different. For centered features near zero logits, its leading update is $b_{t+1} \approx (1+\eta/4)b_t$, so a bias instability is still possible. Finite-horizon suppression by a centered/free-bias control is evidence about the tested dynamics, not a guarantee at arbitrary horizons.

Freezing b_0 fixes the mean logit on the centering sample, or on a population whose feature mean is μ. For a different law Q, $\mathbb{E}_Q[f_t] = w_t^\top \big(\mathbb{E}_Q[\phi] - \mu\big) + b_0$, which can change with w_t; the pilot's source-mean sample and neutral movement reference are distinct. A free intercept restores a constant-response update channel whose finite-horizon behavior is an empirical question.

P Equal reference drift does not fix direction

Matching numerical learning rates does not match functional movement, and matching a scalar reference RMS change also does not match its direction relative to margins. This appendix gives a full-support, equal-reference-budget version of the binary example. It respects ordinary same-input risk bounds and serves as a control on interpretation, not a separate novelty claim.

In the binary construction, take a neutral reference with nuisance agreement $1/2$. Its raw second moment is identity. For a common RMS budget $b > 0$, set

$$\eta_a = \frac{b}{ch\sqrt{1 + a^4\rho^2}}, \qquad \delta\theta = \frac{b(1, a^2\rho)}{\sqrt{1 + a^4\rho^2}}.$$

Every scale then has exactly the same reference RMS displacement b. With $p = .75$, adaptation agreement .9, reversed agreement .1, and $b = 1.2$, the scale-1 effective coefficients are approximately $(2.212784, .445669)$, while scale 16 gives $(1.110330, 1.199943)$. The former follows C and the latter S. Their reversed accuracies are .75 and .10, respectively. Evaluation on the neutral reference itself gives accuracies .75 and .50 despite equal movement norms.

The example has full reference and target raw support. The effect is direction relative to margins, not necessarily a missing reference direction. It uses different numerical rates to implement the same budget-normalized direction rule and is not the equal-rate comparison in the main theorem. Its budget exceeds the initial stable margin $c = \log 3$; for this positive-coordinate update family $b \le c$ prevents nuisance dominance. It does not refute a small-budget margin guarantee.

Indeed, logistic loss is 1-Lipschitz in its logit. For any fixed joint law Q,

$$|\mathcal{R}_Q(f_t) - \mathcal{R}_Q(f_0)| \le \mathbb{E}_Q|f_t(X) - f_0(X)| \le \sqrt{\mathbb{E}_Q(f_t(X) - f_0(X))^2}.$$

This bounds change, not improvement or exact risk. A finite empirical reference need not control an arbitrary target marginal. The bound remains valid on its own stated law.

Q Access interventions and measurement checks

For unit $u \perp w_0$, projected head descent gives

$$u^\top w_{t+1} = u^\top w_t - \eta u^\top (I - uu^\top) g_t = u^\top w_t = 0.$$

Rescaling the displacement about w_0 also preserves this property. The intervention blocks exactly the selected readout component, not every response to the nuisance. If $d_S(x)$ has components outside u, their later logit contributions remain accessible.

In contrast, for a fixed final head and any moving encoder ϕ_θ, a final projection $P = I - uu^\top$ satisfies

$$w_0^\top P \phi_\theta(x) = w_0^\top \phi_\theta(x) \quad \text{for all } \theta, x.$$

Losses depending only on these logits and their encoder gradients are identical. This rules out that placement as a suppression test for fixed-head normalization adaptation. Projection before an adaptable downstream affine map can be nontrivial: initial effective-readout orthogonality can preserve f_0 while later coefficients change. Such an extension must verify the actual forward and update maps rather than inherit the head-only argument.

For response measurement, an affine recalibration $f_t = \alpha f_0 + \beta$ changes the nuisance contrast to $\alpha \mathcal{D}_S f_0$. The residual $\mathcal{D}_S f_t - \widehat{\alpha} \mathcal{D}_S f_0$ measures departure from the fitted global affine response; it does not exclude every nonlinear function of the old score. It does not make the fitted model causal, and a harmful global calibration change can have zero residual. Risk and response must remain separate endpoints.