SHARED-FACTOR DISTRIBUTION SHIFT IN MULTI-TASK LEARNING: IDENTIFICATION AND ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-task models serve tasks whose labels arrive at different speeds. Clicks arrive within minutes and purchases only after weeks. After a shift, the slow tasks must be predicted from the fast tasks' early labels. Fine-tuning on those labels improves every task that can be checked, and worsens a slow one. Which part of a slow task's shift do the early labels determine, and what should be done about the rest? We formalize the shift as window-specific gains on a few factors shared by all tasks, and derive what the early labels determine. A slow task's shift is identified if and only if every factor it reads is read by some fast task, and two past windows identify the factors themselves. Because the part that no fast task reads is unknown, an update that reaches into it can lose by any amount; the best update that avoids it is the refit with its unread part removed. FLAG serves this projected update on a frozen multi-task model. Replayed on four datasets and 47 weeks of a question-answering site, it has the lowest slow-task error of eight baselines, and is worse than not adapting in none of 89 served windows.

1 INTRODUCTION

Multi-task models are often deployed where the data keep shifting. We divide time into *windows*, such as weeks or seasons. Each task must stay accurate in every window, yet labels arrive at different speeds. Clicks arrive within minutes and purchases after days; an answer comes within hours, its acceptance and upvotes over weeks; plant height is measured during the season and yield only at harvest. A task is *fast* if its labels arrive early enough to adapt the model within the same window, and *slow* otherwise. After a shift, the slow tasks must be predicted from the fast tasks' *early labels* before their own arrive.

The obvious remedy is to fine-tune the model on the early labels. On the question-answering site CrossValidated, replayed quarter by quarter, the early labels record whether a question is answered within an hour and within a day. Fine-tuning on them cut the log loss of these two tasks by 3.65% and 4.59%, and that of answer acceptance, a slow task, by 0.79%. Yet it raised the log loss of upvotes, the other slow task, by 1.53% (Figure 1). All metrics available at the time improved. The harm became visible only weeks later, with the upvotes. Week by week, it worsened upvotes in 30 of 47 weeks, and in 23 of these it improved acceptance (Figure 3b). A fast-to-slow relation from the past fails too, as it changes between windows (Section 2).

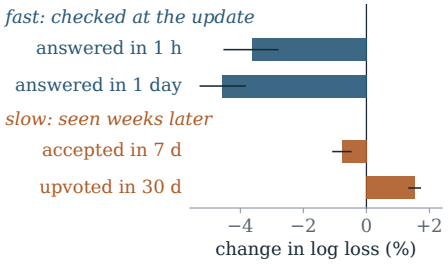


Figure 1: Fine-tuning on CrossValidated's early labels improves every task that can be checked at the time, and worsens upvotes.

The same update helped one slow task and hurt the other, and nothing in the early labels distinguished the two. We therefore ask which part of a slow task's shift the early labels determine, and what should be done about the rest. Existing methods address other questions. Delayed-feedback models (Chapelle, 2014; Yang et al., 2021) correct a task for its own labels still in transit, and sur-

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.

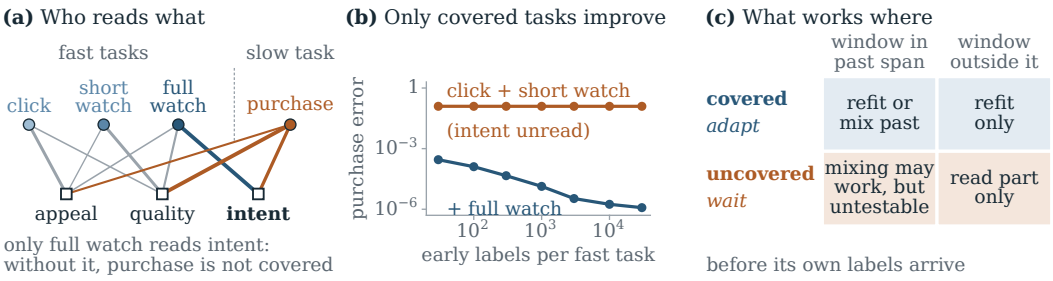


Figure 2: Running example. (a) Only full watch, a fast task, reads intent. (b) Purchase error after refitting the gains, simulated. (c) What recovers a slow task's shift before its labels arrive.

rogate indices (Athey et al., 2019) and synthetic controls (Abadie et al., 2010) reuse a relation or a mix fitted in the past. Robust training and test-time adaptation use no labels from the new window (Section 6). None of them tells which slow tasks the early labels can update.

We answer both questions in a model of a few factors shared by all tasks, such as an item's appeal, its quality and a user's purchase intent. Each task puts fixed weights on the factors and *reads* those with nonzero weight. A window rescales each factor by a number, its *gain* in that window. In our running example, the fast tasks click and short watch read only appeal and quality, so their early labels say nothing about how a campaign rescaled intent. Full watch, a third fast task, also reads intent, and so does purchase, the slow task (Figure 2a). By our main result, a slow task's shift can be recovered from the early labels if and only if every factor it reads is read by some fast task. We call this condition *coverage*. For a covered task, refitting the gains recovers the shift, even when no mix of past windows matches it. For an uncovered task, no number of early labels suffices (Figure 2b). The factors need no annotation, since two past windows identify them when no two factors change by the same ratio between the windows. Coverage can thus be checked in advance.

The model also explains how fine-tuning can hurt a task it never sees. Suppose we fine-tune the last shared layer on the fast tasks, with the heads fixed, by SGD or by gradient surgery such as PCGrad (Yu et al., 2020). Each slow task's prediction then changes by a fixed combination of the fast tasks' changes, with coefficients set by the heads at pretraining. This combination is a surrogate index (Athey et al., 2019) never validated on current slow labels. As each slow task has its own surrogate, one update can help one slow task and hurt another, as with acceptance and upvotes. A surrogate is exact only if the slow task's weights are a combination of the fast tasks' weights. Refitting the gains needs only coverage, a weaker condition, so fine-tuning can miss even a covered task. The harm does not come from conflicting gradients, since one setting of the gains suits every task at once.

We compare every update of a slow task with not adapting, which serves the model trained on past windows. Relative to it, an update changes the task's expected squared error by the update's squared size minus twice its overlap with the true change. The *unread part* of that change, the part no fast task reads, enters only through this overlap. It is unknown and can have either sign, so an update overlapping with it can lose by any amount. Our principle is to use only *safe* updates, whose improvement over not adapting does not depend on the unread part. The best safe update, the *projected update*, is the refit with its unread component removed, and its expected improvement is known in advance.

FLAG (Fast-Label Adaptive Gains) applies this principle to a frozen multi-task model. It learns the factors from past windows, refits the gains to the early labels, and serves each slow task its projected update. For a covered task, the projected update is the full refit, so the task adapts at once. Any other task is corrected where the fast tasks read and waits for its own labels on the rest (Figure 2c).

Our contributions are:

- **The problem.** We ask which part of a slow task's shift the early labels determine, and formalize the shift as per-window gains on a few shared factors.
- **The theory.** It identifies what the early labels determine, shows that fine-tuning the last shared layer is a surrogate index fixed at pretraining, and derives the safe updates, best among them the projected refit.

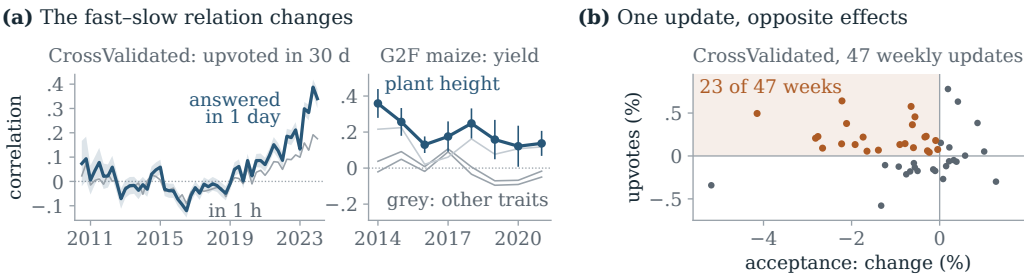


Figure 3: (a) Correlation of a slow task with each fast task, per quarter or year (± 1.96 s.e.). (b) One dot per week: change of acceptance's and upvotes' loss under fine-tuning, relative to not adapting; shaded, acceptance better and upvotes worse.

- **FLAG.** We propose FLAG, which serves the projected update on a frozen multi-task model. On four datasets and a 47-week replay it has the lowest slow-task error of eight baselines, and is worse than not adapting in none of the 89 served windows.

2 SLOW TASKS AFTER A SHIFT

The slow tasks are the ones a deployment cares about, and they are the ones whose labels arrive last. A fast task is usually a cheap proxy: a click stands in for a purchase, a quick answer for a useful one, plant height for yield, and the slow tasks are what the proxies stand for. Their labels arrive late because the outcomes themselves take time. In the e-commerce data of RetailRocket a purchase is counted over 14 days, so its label is final two weeks after the visit, twice the length of a window; on CrossValidated an upvote is counted over 30 days; and in the maize trials of Genomes to Fields (G2F) yield is known only at harvest. A model must therefore be updated long before a window's slow labels are complete.

Both cheap remedies fail, and during the window the failure is invisible. The first is to **fine-tune on the early labels**. It leaves the fast tasks improving, so every metric available at the time improves, while the slow tasks can move either way: on CrossValidated it improves acceptance yet worsens upvotes, in the quarterly replay and in 23 of 47 weekly updates, and on G2F it lowers the squared error for yield by 7% but raises it for grain moisture by 22%, both 95% intervals excluding zero. A slow task must therefore be judged on its own, and harm to it shows only when its labels arrive; averages over tasks, the usual summary in multi-task benchmarks, can hide the harm even then, since the slow tasks move in opposite directions. The second is to **carry the past relation over**: learn from past windows how the slow labels follow the fast ones. That relation changes from window to window (Figure 3a): in G2F the correlation of yield with plant height falls from .26 in 2015 to .13 in 2016, and on CrossValidated the correlation between being answered within a day and being upvoted within 30 days stays below .09 through 2020 and then climbs to .39 by 2023, higher than in any earlier quarter. In the replays of Section 5, a regression of the slow tasks on the fast ones, fitted on past windows, is worse than not adapting on all four datasets, by 0.5 to 7%, and pooling past windows does not help either, since a model trained on all of them learns their average relation rather than the current one (Appendix A).

What a method must do follows from the two failures. It must use the early labels, since past relations go stale; judge each slow task on its own and in each window, since averages can hide harm; and **never do worse than not adapting**, which is always available. For each slow task it must therefore find the part of the shift the early labels reveal, adapt that part, and leave the rest alone. Such a split is possible because shifts have few directions: in G2F three of the seven possible directions carry over 90% of the year-to-year change in how the traits covary (Appendix H.6), and Section 3.1 models them as shared factors and states the problem formally.

3 FROM THE EARLY LABELS TO THE PROJECTED UPDATE

3.1 MODEL AND PROBLEM

Setting. Tasks $t = 1, \dots, T$ are served over windows, and $\mu_{t,e}(x)$ is the mean of task t's label in window e given the input x. Past windows $e = 1, \dots, m$ have all their labels. In the current window $\star$, the model is adapted once the fast tasks $d \in D$ have early labels on the *arrived inputs*, independent given the inputs. A slow task $t^\circ \notin D$ has no labels yet, and its mean must be predicted on the *served inputs*. These can differ from the arrived inputs, as early users are not typical. For a binary task μ is the logit, and the results hold for a linearized working response (Appendix E.1).

Model. In the running example, click and short watch depend on an item's appeal and quality. Full watch and purchase also depend on a user's purchase intent. All tasks share k such factors:

$$\mu_{t,e}(x) = f_t^0(x) + w_t^\top \operatorname{diag}(s_e)\, \phi(x). \tag{1}$$

The *base* f^0 is a fixed multi-task model, such as an MMoE (Ma et al., 2018a). Task t weights the factors $\phi(x) \in \mathbb{R}^k$ by its *readout* w_t, and *reads* factor j if $w_{tj} \neq 0$. Window e rescales the factors by its gains $s_e \in \mathbb{R}^k$. With the task's features $z_t = w_t \odot \phi$, the model reads $\mu_{t,e} = f_t^0 + z_t^\top s_e$. Every task is thus a regression on its own features, with the window's gains as coefficients shared by all tasks. A slow task has no labels, so its gains must come from other tasks' labels.

The assumption. **(A0)** *The base, the factors and the readouts carry over across windows; only the gains move.* How the tasks co-move can still change, since the gains change (Figure 3a). (A0) rules out a change to one task's own base or readout, such as a new checkout step that alters purchases alone. No early label could reveal such a change (Appendix A.5).

Cost. For now the base, factors and readouts are known, and the only unknown is the current gain vector $s^\star$. Not adapting serves the mean past gain $\bar{s}$ and leaves the change $\theta = s^\star - \bar{s}$. Gains off by Δ cost task t the squared error $\|\Delta\|_{G_t}^2 = \Delta^\top G_t \Delta$, with $G_t = \mathbb{E}[z_t z_t^\top]$ over the served inputs. In the running example, purchase is the slow task. Table 7 in Appendix G.1 gives the readouts and gains, with a zero base and $\mathbb{E}[\phi\phi^\top] = I$. A new campaign moves the gains a distance r off the plane of two past weeks' gains, so no mix of the weeks reproduces it. Purchase has $G_t = \operatorname{diag}(.16, 1, .64)$, and not adapting to the campaign costs it $.350\, r^2$.

Problem. We judge an update δ of the gains, served as $\bar{s} + \delta$, by its *improvement* over not adapting,

$$\mathcal{I}_t(\delta; \theta) = \|\theta\|_{G_t}^2 - \mathbb{E}\|\delta - \theta\|_{G_t}^2, \tag{2}$$

with the expectation over the label noise. The early labels leave part of θ open, its unread part (Section 3.2). For each slow task we seek the update whose expected improvement is largest in the worst case over the unread part, with the read part drawn at the past's typical size v. This is a Γ-minimax criterion (Berger, 1985), made precise in Section 3.4.

3.2 WHAT THE LABELS IDENTIFY

The early labels. The early labels depend on the gains only through the fast tasks' features. Let $G_D = \sum_{d \in D} \mathbb{E}[z_d z_d^\top]$ over the arrived inputs and $U = \ker G_D$: moving the gains along U changes no fast task's mean on the arrived inputs. A change's component in U is its *unread part*, and the rest is its *read part*.

Lemma 1 (What the early labels identify)**.** *Assume that equation 1 and (A0) hold with known base, factors and readouts, and that each label's law given its input is determined by its mean, as for Gaussian noise or binary labels. Two current gain vectors give the early labels the same distribution if and only if their difference lies in U. So the current mean of t° on the served inputs is identified if and only if $\ker G_D \subseteq \ker G_{t^\circ}$; when $\mathbb{E}[\phi\phi^\top] \succ 0$ on both sets of inputs, this holds if and only if every factor t° reads is read by some fast task.*

We call t° *covered* when $\ker G_D \subseteq \ker G_{t^\circ}$. A change along U then leaves its predictions unchanged, and nothing it depends on is hidden. For an uncovered task, the change along U moves its

predictions, and no number of early labels reveals it. With click and short watch as the fast tasks, U is the intent axis. A least-squares refit of the gains to the early labels then leaves purchase's error at $.124\,r^2$, the cost of the campaign's change in intent, from 30 to 30,000 labels per task (Figure 2b). U depends only on the readouts and the inputs' second moments, so coverage can be checked before any slow label arrives (proof in Appendix A.1, via linear estimability, Scheffé, 1959).

The past windows. The factors are not observed, but the past windows supply them. By the uniqueness of the canonical polyadic decomposition (Harshman, 1970; Kruskal, 1977), as in identification through heteroskedasticity (Rigobon, 2003), two past windows identify the readouts and factors up to order and scale. This needs a known base, readouts of full column rank, a shared input law, and no two factors changing by the same ratio between the windows (Lemma 3 in Appendix D).

3.3 Why fine-tuning can hurt a slow task

The obvious alternative to refitting the gains is to fine-tune the network on the early labels.

Proposition 1 (Fine-tuning is a surrogate index fixed at pretraining)**.** *Let a multi-task network predict $f_t(x) = b_t + h_t^\top M\psi(x)$ with a frozen representation ψ, a last shared layer M and fixed heads h_t, and let H_D have the fast tasks' heads as columns. If M is trained on the early labels by steps that linearly combine per-example gradients of the fast tasks' losses, as in SGD, PCGrad (Yu et al., 2020) or MGDA (Sener & Koltun, 2018), then at every step, with Δf_D the fast tasks' changes, for all t and x*

$$\Delta f_t(x) = c_t^\top \Delta f_D(x), \qquad c_t = (H_D^\top H_D)^+ H_D^\top h_t. \tag{3}$$

Fine-tuning thus predicts each slow task's change from the fast tasks' changes with fixed weights c_t. These weights form a surrogate index (Athey et al., 2019), set by the heads at pretraining and fitted to no current slow label. Each slow task has its own c_t, so one update can improve every fast task, help one slow task and hurt another, the surrogate paradox (Chen et al., 2007; VanderWeele, 2013). Adam, weight decay and trained heads break the exact form; training all shared layers keeps it to first order (Appendix B).

Support, not span. In equation 1, the heads play the role of the readouts. A fixed combination of the fast tasks' changes is exact only if the slow task's readout is a combination of theirs. Otherwise no number of labels removes its error (Proposition 4). The refit needs only coverage. Short and full watch cover purchase, but purchase's readout is not a combination of theirs. With them as the fast tasks, the refit is exact. Yet for an appeal-only campaign, fine-tuning errs by $.490\,r^2$, three times the error of not adapting. Fine-tuning is thus neither exact for a covered task nor, in general, safe for an uncovered one. None of this is a conflict between tasks: under (A0) the current gains minimize every task's loss at once. FLAG therefore moves slow tasks only through their fixed readouts.

3.4 Which update is safe

Expanding equation 2 gives $\mathcal{I}_t(\delta;\theta) = 2\langle \mathbb{E}\delta, \theta\rangle_{G_t} - \mathbb{E}\|\delta\|^2_{G_t}$: an update gains through its overlap with the true change and pays for its own size. A shift of θ along U leaves the law of the early labels unchanged (Lemma 1), and with it the law of any update computed from them, so for every $u \in U$

$$\mathcal{I}_t(\delta;\theta+u) = \mathcal{I}_t(\delta;\theta) + 2\langle \mathbb{E}\delta, u\rangle_{G_t}. \tag{4}$$

The unread part enters linearly. Let B have orthonormal columns spanning U. Unless the overlap $B^\top G_t \mathbb{E}\delta$ is zero, the improvement can be made as negative as desired. A zero overlap makes the improvement the same for every $\theta + U$ (Lemma 2), and we call such an update *safe*.

The natural candidate refits what the early labels read. Let Z be the arrived design, with row $z_{d_i}(x_i)^\top$ for label i, and $F = Z^\top Z/\sigma^2$, with σ^2 the noise variance, the labels' information about the gains. Shrinking the least-squares estimate $\hat\theta$ toward zero gives $\theta_W = (v^{-1}I + F)^{-1}F\hat\theta$, the posterior mean when the read part of θ has variance v in every direction. θ_W leaves the unread directions at their past value. Yet it is not safe if the task's features couple read and unread directions (Proposition 7). Let $P_t = B(B^\top G_t B)^+ B^\top G_t$ project onto U in the task's geometry. $P_t\theta$ is then the unread part of the change for task t, and the *projected update* removes it from the refit:

$$\delta_t^* = (I - P_t)\,\theta_W. \tag{5}$$

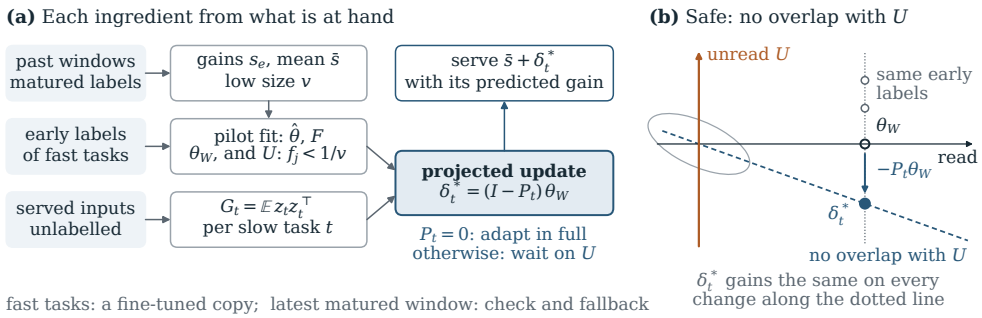


Figure 4: FLAG in one window. (a) Where each ingredient comes from (every constant: Algorithm 1, Appendix F). (b) Task t's geometry, with a level set of $\|\cdot\|_{G_t}$; every change on the dotted line gives the same early labels.

Theorem 1 (The projected update is the best safe update)**.** *For Gaussian noise, let Γ_v hold every law of θ whose read part is $\mathcal{N}(0, vI)$, with any law for the unread part. (i) Over Γ_v, δ_t^* has the largest worst-case expected improvement of all updates, uniquely up to changes that task t does not see. (ii) Its improvement is the same for all $\theta + U$, and its average under every law in Γ_v is $\mathbb{E}\|\delta_t^*\|_{G_t}^2 \geq 0$, set by F, G_t and v.*

A covered task has $P_t = 0$, so $\delta_t^* = \theta_W$: it adapts in full. An uncovered task is corrected on the read directions and waits for its own labels on the rest. The past supplies only the size v, and a larger true size keeps the guarantee (Proposition 10). In the example with click and short watch, take two campaigns with the same change in appeal and quality. In one, intent follows the two weeks' co-movement ($+.33\,r$), and in the other it falls ($-.44\,r$). Filling intent from the past co-movement improves on not adapting by $.294\,r^2$ in the first campaign and loses $.028\,r^2$ in the second. The projected update improves by $.226\,r^2$ in both, a number known before any purchase label.

4 FLAG: THE PROJECTED UPDATE ON A FROZEN MODEL

Section 3 took the factors as known and the tasks as linear. A deployed model names no factors, and most of its tasks are binary. FLAG keeps the trained model frozen and estimates each ingredient of Theorem 1 from what is at hand (Figure 4a). The past windows give the features z_t and the size v, the early labels give the information F and the unread directions U, and the served inputs give each G_t. Where the guarantee rests on an uncertain estimate, the estimate errs toward not adapting.

Window gains on a frozen model. FLAG attaches q window gains $s \in \mathbb{R}^q$ to the frozen model in one of three ways. (i) *Expert gains* rescale the experts of an MMoE (Ma et al., 2018a). (ii) A *residual branch* adds $w_t^\top \operatorname{diag}(s)\,\phi(x)$, with ϕ a learned map of a frozen hidden layer. Trained on the past windows with one gain vector per window, it learns readouts and factors, which Section 3.2 shows to be identified. (iii) A *window-embedding gate* scales a frozen hidden layer by a function of s, as in FiLM and PEPNet (Perez et al., 2018; Chang et al., 2023). Each past window e has its own gains s_e, scaled by their root-mean-square. Their mean $\bar{s}$ reproduces the frozen model, which is what not adapting serves. The features z_t are the gradient of task t's prediction, a logit if binary, in s. Equation 1 holds exactly for the residual branch and to first order for the others (Appendix E.2).

Size and information. A larger true size keeps the guarantee of Theorem 1 (Proposition 10), and even a third of v leaves a positive expected improvement (Appendix C.1). FLAG therefore takes v low: the lower 10% quantile of the past changes' mean variance over bootstrap resamples of the past windows. A pilot fit of the gains to the early labels, penalized toward $\bar{s}$, fixes the point where binary tasks are linearized. The gradients of the early labels' predictions at this point form the rows of Z, and one step of iteratively reweighted least squares gives $\hat{\theta}$ and F. Labels of one user or question are correlated, so F is cluster-robust (Liang & Zeger, 1986) and capped at the model-based information.

Weakly read directions. Learned readouts are rarely exactly zero, so the early labels read some directions only weakly. FLAG counts a direction as unread when the labels read it less precisely than

a typical change. U is thus spanned by the eigenvectors of F with eigenvalues $f_j < 1/v$. Along them, θ_W errs more from shrinking toward the past than from label noise (Appendix C.3). As U is spanned by eigenvectors of F, δ_t^* uses only the information off U. Theorem 1 (ii) thus holds for this U, and the weak information on U could improve the worst case only slightly (Propositions 8 and 9).

Serving and checks. G_t needs no labels, so FLAG estimates it on the served inputs. Each slow task is served $\bar{s} + \delta_t^*$, with the predicted improvement $\mathbb{E}\|\delta_t^*\|_{G_t}^2$ in closed form (Appendix C.1). A waiting task collects its own labels, and these read U in proportion to the task's weight there (Appendix F.2). The fast tasks are served from a copy fine-tuned on their own labels, the slow tasks never (Proposition 1). Theorem 1 takes the fitted readouts as given. Their error enters, like a violation of (A0), as model error that no early label reveals (Appendix C.1). The latest window with matured slow labels therefore checks each slow task. A task falls back to the frozen model if FLAG is worse there by more than one standard error. The same window chooses the construction, q and a multiplier of v in $\{\frac{1}{4}, \frac{1}{2}, 1\}$.

5 EXPERIMENTS

The replays test three predictions of the principle: FLAG is never worse than not adapting, its gain comes from the projection, and its waiting follows coverage.

5.1 SETUP

Data and windows. On RetailRocket (Zykov, 2017) (weekly windows), views and add-to-carts are fast and purchases within 14 days slow. On CrossValidated (Stack Exchange, Inc., 2024) (quarterly), answers within an hour and a day are fast, and acceptance and upvotes slow. On G2F maize (Lima et al., 2023) (yearly), flowering dates and heights are fast, and yield, moisture and test weight slow. On KuaiRand-Pure (Gao et al., 2022) (two-day windows), click and long view are fast, and five rare engagements, held back until each window ends, are slow (Appendix H.3). The last twelve windows of each dataset are served in turn (G2F: its last six years), each after retraining on all earlier windows' matured labels. A window's early part supplies the early labels, and its later part scores the slow tasks. With the weekly replay of Section 5.3, 89 windows are served (Appendix H).

Baselines and metric. Eight baselines cover three responses to a shift (Table 1). *No adaptation* serves the model trained on the past windows, by standard training (not adapting) or by GroupDRO (Sagawa et al., 2020). *Adaptation* to the early labels fine-tunes a shared-bottom network or an MMoE (Ma et al., 2018a), fits a window embedding after PEPNet (Chang et al., 2023), or fits a full adapter on a shared layer. *Reuse* of the past combines past windows' predictions or regresses the slow tasks on the fast ones, a surrogate index (Athey et al., 2019). All methods are tuned as in DomainBed (Gulrajani & Lopez-Paz, 2021), with 20 configurations chosen on the latest matured window. We report normalized entropy (He et al., 2014), the log loss divided by that of the training base rate. The base rate scores 1; G2F uses squared error over the past variance. It is averaged over slow tasks, then over windows or taken in the worst window. A method is *worse than not adapting* in a window if its slow-task mean is higher by more than their summed standard errors over three seeds.

5.2 RESULTS

FLAG is never worse than not adapting. The replays repeat the problem of Section 2. Fine-tuning worsens the slow tasks on CrossValidated and G2F. Every baseline that adapts or reuses is worse than not adapting in 6 to 38 of the 89 windows (Table 1, last column). FLAG is worse in none. It has the lowest slow-task error on all five replays, 0.6 to 4.8% below the best baseline and 0.6 to 6.9% below it in the worst window, with 95% bootstrap intervals excluding zero. The fallback to the frozen model fired once, in G2F's first served year; without it, FLAG is still never worse.

The gain comes from safe updates. Refitting the gains without the projection, θ_W, gives half to three fifths of the mean gain over the best baseline on every replay. Yet it is worse than not

Table 1: Slow-task normalized entropy (G2F: relative squared error; lower is better), mean over served windows and worst window, and the number of the 89 windows in which a method is worse than not adapting. Best baseline in italics.

	KuaiRand-Pure		RetailRocket		CV, quarterly		G2F maize		CV, weekly		worse
method	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst	count
not adapting	1.024	1.199	*.829*	.962	.966	.993	*.879*	*1.603*	.971	1.004	—
GroupDRO	1.047	1.254	.832	*.955*	.974	1.006	.901	1.668	.973	1.004	19
shared-bottom, fine-tuned	1.022	*1.196*	.830	.964	.971	1.007	.941	1.834	.968	1.005	7
MMoE, fine-tuned	*1.015*	1.212	*.829*	.968	.974	1.038	.938	1.674	.969	*1.001*	*6*
window-embedding gate	1.134	1.351	.830	.966	*.930*	*.979*	.968	1.867	.972	1.004	14
pooled full adapter	1.020	1.208	.831	.968	.976	1.014	1.088	1.962	*.967*	1.004	9
past-window combination	1.130	1.343	.831	.961	1.224	1.935	.978	1.687	.976	1.081	29
fast-to-slow regression	1.096	1.312	.842	.965	.971	1.015	.941	1.719	.975	1.027	38
FLAG	**.978**	**1.147**	**.801**	**.911**	**.885**	**.911**	**.842**	**1.510**	**.961**	**.995**	**0**
vs. best, %	−3.6	−4.1	−3.4	−4.6	−4.8	−6.9	−4.2	−5.8	−0.6	−0.6	
ablation: θ_W, no projection	.995	1.177	.814	.939	.905	.952	.858	1.568	.964	1.005	4
ablation: always fill U	1.007	1.188	.822	.945	.917	.962	.870	1.622	.965	1.010	15

adapting in 4 windows. The projection gives the rest of the gain and most of the worst-window gain: without it, the worst window is .010 to .058 higher. Filling U from the past co-movement in every window, an unsafe update, makes FLAG worse than not adapting in 15 windows. Where the truth is known, the pattern is the same. We simulate 2,000 windows with random tasks and correlated factors (Appendix G.3). The projected update is worse than not adapting in 0.9% of the slow-task windows, all within label noise, against 4% for θ_W and 25% for fine-tuning.

The harm sits in the unread part. Once the slow labels mature, a method's gap to the projected update splits in two. One part is its error on the directions the early labels read. The other is its fill of the unread part (Appendix H.4). Over the 114 slow-task windows of the four short replays, the fill makes up 43% of fine-tuning's gap and 91% of reuse's. What model equation 1 does not represent is under a tenth. (A0) also holds closely on matured windows. Between consecutive windows, refitting only the gains removes 81 to 93% of the slow tasks' squared logit error (Appendix H.4).

Coverage decides who waits. A slow task's *coverage share* $1 - \operatorname{tr}(P_t G_t)/\operatorname{tr} G_t$ is the part of its weight on the gains that the early labels read. It is .31 to .58 for purchases, upvotes and KuaiRand's engagements. These tasks largely wait, and the projection leaves them under 45% of θ_W's change. It is .87 to .96 for acceptance and G2F's traits, which adapt and keep over 90%. Predicted and realized improvements have Spearman correlation .62. FLAG lost to not adapting in 11 of the 114 slow-task windows, each time by less than label noise. FLAG adds 0.4 to 2.1 s per window on a CPU.

5.3 CASE STUDY: A TRIAGE QUEUE FOR NEW QUESTIONS

On CrossValidated, a new question is worth flagging if it will lack an accepted answer after a week. We replay this weekly from April 2023 to February 2024 (47 weeks, 8,704 scored questions). Answers in each week's first two days are the early labels, and each day the 20% most at-risk new questions enter the queue. All methods share one base network with a weekly embedding (Appendix H.5).

Table 2: Triage queue: entropy, unresolved questions caught per 100 queued, and task-weeks (of 94) worse than not adapting. Italics: best baseline. Last two rows: with slow outcomes seen early.

method	slow	accept.	upv.	caught	worse
not adapting	.971	.952	*.990*	91.9	—
fine-tuned	*.968*	*.945*	.991	91.8	15
window embedding	.972	.950	.995	91.9	24
level recalibration	.969	.946	.992	91.9	19
fast-to-slow regression	.975	.957	.993	90.8	3
FLAG	**.961**	**.940**	**.982**	**92.6**	**0**
+ priced fill	.958	.934	.982	93.0	0
delayed feedback	.962	.948	.976	92.0	8
FLAG	**.953**	**.935**	**.971**	**93.1**	**0**

Fine-tuning's harm is a surrogate's, not a conflict. Over the 47 weeks, the split of Section 1 recurs. Adapting the embedding lowers acceptance's error by .0024 [.0012, .0039] but raises that of upvotes by .0044 [.0024, .0065]. Every baseline that adapts to the fast labels worsens upvotes in 28 to 30 of the weeks, and in 23 of

them fine-tuning also improves acceptance. Proposition 1's first-order prediction explains 83% of the variance of fine-tuning's weekly change in upvotes. The two parts of its gap to the projected update rank its weekly harm with Spearman correlation $.64$, against $.07$ for the cosine between the fast and slow tasks' gradients.

FLAG in the queue. FLAG worsens upvotes in 9 weeks, none of them beyond label noise (Table 2). It lowers the slow tasks' entropy to $.961$, $.007$ below the best baseline, and catches 0.8 more unresolved questions per 100 queued. Matured weeks also reveal past unread changes, so replaying them can price a fill of U from the past co-movement (Appendix F). Priced on the 52 matured weeks of 2022, the fill is used for acceptance in 19 weeks and never for upvotes, whose relation to the fast labels drifted (Figure 3a). It lowers the entropy to $.958$ but is not safe in the sense of Section 3.4, so we report it apart.

When some slow outcomes arrive early. A delayed-feedback model (Chapelle, 2014) uses the 45% of acceptances and 77% of upvotes seen by the cutoff to reach $.962$, but is worse than not adapting in 8 of the 94 task-weeks (Appendix H.5). With these outcomes in Z as the slow tasks' own labels, FLAG reaches $.953$ and is worse in none.

6 RELATED WORK

Negative transfer. It is usually attributed to conflicting gradients, which optimizers reweight or project (Chen et al., 2018; Sener & Koltun, 2018; Yu et al., 2020; Navon et al., 2022). Proposition 1 gives a harm that no such optimizer removes. One setting of the gains suits every task, yet fine-tuning moves a slow task by a relation fixed at pretraining (cf. Kumar et al., 2022).

Late outcomes and new windows. Delayed-feedback models correct a task for its own labels in transit (Chapelle, 2014; Ktena et al., 2019; Yang et al., 2021), and FLAG takes such seen outcomes as labels (Section 5.3). Surrogate indices (Athey et al., 2019), conversion models (Ma et al., 2018b; Yang & Zhan, 2022) and synthetic controls with past windows as donors (Abadie et al., 2010; Agarwal et al., 2020) carry late outcomes from early ones through a relation fitted in the past. That relation drifts (Section 2), and a fixed combination needs span, not only coverage (Section 3.3). Robust objectives (Sagawa et al., 2020; Duchi & Namkoong, 2021; Arjovsky et al., 2019; Krueger et al., 2021) and test-time adaptation (Sun et al., 2020; Wang et al., 2021; 2022) use no labels from the new window. Feature-wise scaling (Perez et al., 2018; Liu et al., 2022; Lian et al., 2022) fits scales to a task's own labels, which a slow task lacks.

Factor models and safe decisions. Nowcasting reads latent factors from early-released series through fixed loadings (Giannone et al., 2008; Bańbura & Modugno, 2014). It fills even an unloaded factor from the past co-movement, and such a fill can lose by any amount (Section 3.4). Partial identification (Manski, 2003; Giacomini & Kitagawa, 2021), minimax regret against a status quo (Manski, 2004; Stoye, 2009) and safe policy improvement (Thomas et al., 2015; Laroche et al., 2019) ask a decision not to lose to a baseline where the data identify a target only in part, as early labels do for a slow task.

7 CONCLUSION

We studied how the early labels of fast tasks can update slow tasks after a shift. With shared factors and moving gains, the early labels recover a slow task's shift if and only if the fast tasks read every factor it reads. Two past windows identify the factors. Fine-tuning moves each slow task by a surrogate fixed at pretraining. It can thus hurt a slow task while every task checked at the time improves. The best safe update is the refit with its unread component removed, and its expected improvement is known in advance. FLAG serves it on a frozen model and, in our replays, was never worse than not adapting.

The guarantees assume that only the gains move and take the fitted readouts as given. A change in one task's readout is invisible to the early labels, and only the check on matured windows can catch it. Natural next steps are slowly drifting readouts and ranking requests for slow labels by the predicted improvement.

AI USE STATEMENT

Generative AI tools assisted with reading and organizing the related work, with locating and preparing the public datasets used in 5, and with polishing the prose and producing the figures. All modelling choices, theoretical results, experiments and interpretations are the authors' own, and we take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix F specify FLAG with every constant, guard and fallback (Algorithm 1); Appendix H lists data sources, windows, label timing, splits, tuning and baselines; Appendix G specifies the known-truth experiments; and Appendices A–E prove every stated result.

ETHICS STATEMENT

All four datasets are public releases. We use them to predict aggregate outcomes of recommendations, questions and field plots, and make no predictions about identifiable individuals.

REFERENCES

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi: 10.1198/jasa.2009.ap08746.

Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic interventions. *arXiv preprint arXiv:2006.07691*, 2020.

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

Susan Athey, Raj Chetty, Guido W. Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Working Paper 26463, National Bureau of Economic Research, 2019.

Marta Bańbura and Michele Modugno. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, 29(1):133–160, 2014. doi: 10.1002/jae.2306.

K. E. Basford, P. M. Kroonenberg, and I. H. DeLacy. Three-way methods for multiattribute genotype $\times$ environment data: an illustrated partial survey. *Field Crops Research*, 27(1-2):131–157, 1991. doi: 10.1016/0378-4290(91)90027-S.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010. doi: 10.1007/s10994-009-5152-4.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2 edition, 1985.

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing (STOC)*, pp. 594–603. ACM, 2014. doi: 10.1145/2591796.2591881.

Jianxin Chang, Chenbin Zhang, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, and Kun Gai. PEPNet: Parameter and embedding personalized network for infusing with personalized prior information. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 3795–3804. ACM, 2023. doi: 10.1145/3580305.3599884.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1097–1105, 2014. doi: 10.1145/2623330.2623634.

Hua Chen, Zhi Geng, and Jinzhu Jia. Criteria for surrogate end points. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):919–932, 2007.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018.

Simon S. Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3), 2021. doi: 10.1214/20-AOS2004.

Zhili Feng, Shaobo Han, and Simon S. Du. Provable adaptation across multiway domains via representation learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 3953–3957, 2022. doi: 10.1145/3511808.3557624.

Raffaella Giacomini and Toru Kitagawa. Robust Bayesian inference for set-identified models. *Econometrica*, 89(4):1519–1556, 2021.

Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, 2008. doi: 10.1016/j.jmoneco.2008.05.010.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. doi: 10.2307/1912775.

Richard A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, December 1970.

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. Practical lessons from predicting clicks on ads at Facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pp. 1–9, 2014. doi: 10.1145/2648584.2648589.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991. ISBN 978-0-521-30587-7. doi: 10.1017/CBO9780511840371.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/d305281faf947ca7acade9ad5c8c818c-Abstract.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.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 (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021.

Joseph B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95–138, 1977. doi: 10.1016/0024-3795(77)90069-6.

Sofia Ira Ktena, Alykhan Tejani, Lucas Theis, Pranay Kumar Myana, Deepak Dilipkumar, Ferenc Huszár, Steven Yoo, and Wenzhe Shi. Addressing delayed feedback for continuous training with neural networks in CTR prediction. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 187–195, 2019. doi: 10.1145/3298689.3347002.

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.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 3652–3661, 2019.

S. E. Leurgans, R. T. Ross, and R. B. Abel. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993. doi: 10.1137/0614071.

Sai Li and Linjun Zhang. Multi-dimensional domain generalization with low-rank structures. *arXiv preprint arXiv:2309.09555*, 2023.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 109–123, 2022.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. doi: 10.1093/biomet/73.1.13.

Dayane Cristina Lima, Jacob D. Washburn, José Ignacio Varela, Qiuyue Chen, Joseph L. Gage, Maria Cinta Romay, James Holland, et al. Genomes to Fields 2022 maize genotype by environment prediction competition. *BMC Research Notes*, 16(1):148, 2023. doi: 10.1186/s13104-023-06421-z.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130, 2018.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1950–1965, 2022.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1930–1939, 2018a. doi: 10.1145/3219819.3220007.

Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1137–1140, 2018b. doi: 10.1145/3209978.3210104.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4): 1221–1246, 2004.

Ankur Moitra. *Algorithmic Aspects of Machine Learning*. Cambridge University Press, 2018.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446. PMLR, 2022.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E. Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7957–7968, 2019.

Roberto Rigobon. Identification through heteroskedasticity. *Review of Economics and Statistics*, 85 (4):777–792, 2003.

Daniel Runcie and Hao Cheng. Pitfalls and remedies for cross validation with multi-trait genomic prediction methods. *G3: Genes|Genomes|Genetics*, 9(11):3727–3741, 2019. doi: 10.1534/g3.119.400598.

Jessica Rutkoski, Jesse Poland, Suchismita Mondal, Enrique Autrique, Lorena González Pérez, José Crossa, Matthew Reynolds, and Ravi Singh. Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3: Genes, Genomes, Genetics*, 6(9):2799–2808, 2016. doi: 10.1534/g3.116.032888.

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 (ICLR)*, 2020.

J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958. doi: 10.2307/1907619.

Henry Scheffé. *The Analysis of Variance*. Wiley, 1959.

J. Schur. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 140:1–28, 1911. doi: 10.1515/crll.1911.140.1.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, and Xiaoqiang Zhu. One model to serve all: Star topology adaptive recommender for multi-domain CTR prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 4104–4113. ACM, 2021. doi: 10.1145/3459637.3481941.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Stack Exchange, Inc. Stack Exchange data dump, 2024-04-02. Internet Archive, `https://archive.org/details/stackexchange`, 2024.

Jörg Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1): 70–81, 2009.

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, 2020.

Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. In *Fourteenth ACM Conference on Recommender Systems*, pp. 269–278, 2020. doi: 10.1145/3383313.3412236.

Aleksey Tetenov. Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165, 2012.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2380–2388, 2015.

Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7852–7862. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/59587bffec1c7846f3e34230141556ae-Abstract.html.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794.

Tyler J. VanderWeele. Surrogate measures and consistent surrogates. *Biometrics*, 69(3):561–565, 2013.

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 (ICLR)*, 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*, pp. 7201–7211, 2022.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. doi: 10.2307/1912526.

Jia-Qi Yang and De-Chuan Zhan. Generalized delayed feedback model with post-click information in recommender systems. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26192–26203, 2022.

Jia-Qi Yang, Xiang Li, Shuguang Han, Tao Zhuang, De-Chuan Zhan, Xiaoyi Zeng, and Bin Tong. Capturing delayed feedback in conversion rate prediction via elapsed-time sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Yongxin Yang and Timothy M. Hospedales. A unified perspective on multi-domain and multi-task learning. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.7489.

Yongxin Yang and Timothy M. Hospedales. Unifying multi-domain multitask learning: Tensor and neural network perspectives. In *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pp. 291–309. 2017. doi: 10.1007/978-3-319-58347-1_16.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Roman Zykov. Retailrocket recommender system dataset. Kaggle, https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset, 2017.

OVERVIEW: RESULTS, SCOPE AND NOTATION

Where is each result stated and proved, which assumptions does it use, and what does each symbol mean? Table 3 says where each result is stated and proved; Table 4 gives, for each assumption, the results that use it, what fails without it and where it is checked; and Table 5 fixes the notation, one meaning per symbol. The main text states Lemma 1, Proposition 1 and Theorem 1. Every other numbered result is stated here, among them the linear entry of the unread part (Lemma 2), what the past must supply (Proposition 10), the finite-label risk of a refit (Proposition 2), the error of fixed combinations (Proposition 4), the two-window identification of the factors (Lemma 3) and the cost of learning them (Theorem 2). Each appendix opens with the question it answers and names the main-text sections it supports: Appendices A, B, C and D prove the results of Section 3, Appendix E carries them to binary tasks, nonlinear coordinates and misspecified models, Appendix F specifies FLAG, Appendices G and H give the experiments of Section 5, including the drift statistics of Section 1, and Appendices I and J look at other datasets and compare with the closest precedents.

Table 3: Where each numbered result is stated and proved: first the results of the main text, then those stated in the appendix; remarks follow their result.

result	stated	proved	result	stated	proved
Stated in the main text					
Lemma 1 (what early labels identify)	3.2	A.1	Proposition 1 (fine-tuning's surrogate)	3.3	B
Theorem 1 (best safe update)	3.4	C.1			
Stated in the appendix					
Lemma 2 (linear entry)	C.1	C.1	Proposition 7 (θ_W not safe)	C.1	C.1
Propositions 8, 9 (weak reads)	C.1	C.1	Proposition 10 (what the past supplies)	C.1	C.1
Propositions 11, 12	C.1	C.1	split and bet, (9), (10)	C.2	C.2
Proposition 4 (fixed relations)	A.3	A.3	Lemma 3 (two windows)	D.1	D.2
Proposition 2 (finite-label risk)	A.1	A.1	Proposition 3 (refit optimal)	A.2	A.2
Proposition 5 (blind floor)	A.4	A.4	Proposition 6 (invisible change)	A.5	A.5
Theorem 2 (factor cost)	D.4	D.4	Lemma 4 (general form)	D.2	D.2
Lemma 5 (stability)	D.3	D.3	Lemma A (perturbation facts)	D.3	standard
Lemma 6 (matrix pencil)	D.4	D.4	Proposition 15 (pooled basis)	D.5	D.5
Lemma 7 (near-affine)	E.2	E.2	Proposition 16 (binary tasks)	E.1	E.1
Corollary 2 (what fine-tuning serves)	B	B	Propositions 13, 14 (information view)	C.3	C.3
Corollary 3 (no selection)	C.2	C.2	Proposition 17 (arrived slow labels)	F.2	F.2
Corollaries 1, 4	A.3, E.1	A.3, E.1			

Table 4: Scope of the assumptions: each assumption, the main-text and the appendix results that use it, what fails without it, and where it is checked.

	results that use it			
assumption	main text	appendix	what fails without it	where it is checked
correct base, factors and readouts	Lemma 1, Theorem 1	Proposition 4; Appendices A, C, D	the bias of Appendix E.3 enters, which the cluster scores do not estimate	the matured-window check (Section 4); the remainder of the split (Appendix H.4)
fixed heads and a gradient-span optimizer	Proposition 1	Corollary 2	Adam, decay toward zero and trained heads or towers break the exact form, which deeper layers keep only to first order (Appendix B)	simulated networks (Appendix B); the triage queue (Section 5.3)
common input law, known base, uncentred coordinates	—	Lemma 3, Theorem 2	without any one of the three, two windows do not fix the factors (counterexamples in Appendix D.2)	Tables 9 and 10, where they hold
two independent past windows and the pencil	—	Theorem 2, each of whose terms has its own lower-bound subfamily	FLAG's fitted constructions fall outside the rate, so the matured-window check of Section 4 judges the learned part	slopes in Table 10
linear coordinates, squared error of the mean or logit	Theorem 1, equation 9	Propositions 2, 3, 13	U, P_t and the predicted gain are working quantities at the pilot, with the remainders of Appendices E.1–E.2; under log loss a guarantee needs a bounded blind part (Appendix E.1)	deployments, whose tasks are mostly binary
the right number of factors	Lemma 1	Lemma 3	too few factors leave an approximation error, too many break uniqueness (Appendix E.3)	the width q is chosen on validation (Section 4)
the past's size on the read directions, $\mathbb{E}\theta_R\theta_R^\top = vI$	Theorem 1(i)–(ii)	Propositions 13, 14	the predicted gain is an average over changes of that size; a true size above $v/3$ keeps the expected gain on every direction read at $\lambda_j \geq 1$ (Appendix C.1)	predicted against realized gains (Section 5.2)
nothing trusted in the blind spot	Lemma 2, Theorem 1	Corollary 3	trusting the past's size there too keeps δ_t^* minimax exactly when $\|G_{t,RU}G_{t,UU}^{-1}\| \leq 1$ (Proposition 12)	computed before any slow label (Section 4)
no slow label at decision time, or arrival independent of the label	—	Proposition 17	arrived slow labels enter Z and shrink U; label-dependent arrival biases the logistic score	holds by construction: every slow label of a served window is withheld (Appendix H.1)
a shift in shared factors	model equation 1, and every result built on it		adaptation has less to gain; RetailRocket has the smallest shift of the four, its task correlations changing in strength while their axis stays within noise (Appendix H.6)	RetailRocket (Table 13)

A WHAT DO THE EARLY LABELS IDENTIFY, AND WHAT CAN FIXED COMBINATIONS REACH?

This appendix proves the results of Sections 3.2 and 3.3, bounds how fast new labels lower a covered task's risk for any number of factors, extends the floors to other comparators, and proves the blind floor of Section 2 and that a change only the slow task reads is invisible.

A.1 PROOF OF LEMMA 1, AND THE RISK OF A REFIT WITH FINITELY MANY LABELS

Proof of Lemma 1. For two current gains s, s' the base cancels, and the mean of fast task d changes by $z_d^\top(s - s')$. Summed over the fast tasks, its squared L^2 norm is $(s - s')^\top G_D(s - s')$, which vanishes exactly on $\ker G_D$ because $G_D \succeq 0$. A regression label is a location family in its mean, because its noise law does not depend on the gains, and the logistic link of a binary label is injective. Hence two gains give the same fast-label laws if and only if $s - s' \in \ker G_D$, which is (i). The same computation for task t° shows that the two slow means agree if and only if $s - s' \in \ker G_{t^\circ}$. The slow mean is therefore identified if $\ker G_D \subseteq \ker G_{t^\circ}$; conversely, for $\Delta \in \ker G_D \setminus \ker G_{t^\circ}$ the gains $s^\star$ and $s^\star + \Delta$ are indistinguishable yet predict different slow means. All tasks are identified if and only if $\ker G_D \subseteq \bigcap_t \ker G_t = \ker(\sum_t G_t)$, since a sum of positive semidefinite forms vanishes only if each does. □

With finitely many labels, the least-squares refit has two sources of error: label noise, and change on directions that no label informs. Stack the arrived labels into y and the base's predictions for them into o. Label i (fast task d_i, input x_i) gives row $z_{d_i}(x_i)^\top$ of the design Z; Q projects onto $\ker Z$, and Z^+ is the pseudo-inverse.

Table 5: Notation. Every symbol used outside a single result or proof is listed; the others are defined where they appear.

symbol	meaning	symbol	meaning
Symbols of the main text			
x	a unit's input	$e, m; \star$	a past window, of m; the current window
t, T	a task, of T	$D; t^{\circ}$	the fast tasks, whose labels have arrived; a slow task
f_t^0	base of task t	$\mu_{t,e}$	mean, or logit, of task t in window e
$k; \phi(x)$	number of factors; the factors, in $\mathbb{R}^k$	w_t	readout of task t
z_t	design of task t, $w_t \odot \phi = \partial \mu_{t,e}/\partial s_e$	$s_e; s^\star; \bar{s}$	gains of window e; current gains; mean past gains
$\delta_{t,e}$	residual $\mu_{t,e} - f_t^0 = z_t^\top s_e$	Δ	difference of gain vectors
G_D	$\sum_{d\in D} \mathbb{E}[z_d z_d^\top]$ on the arrived inputs	G_t	$\mathbb{E}[z_t z_t^\top]$ on the served inputs; $\|a\|^2_{G_t} = a^\top G_t a$
$U; B$	blind subspace, $\ker G_D$, $\ker Z$ or F's directions with $\lambda_j < 1$; a basis of it	P_t	G_t-orthogonal projection onto U, $B(B^\top G_t B)^+ B^\top G_t$
$y; o$	arrived labels; their base predictions	$Z; Q$	design, rows $z_{d_i}(x_i)^\top$; projection onto $\ker Z$
$\varepsilon; \sigma$	label noise; its s.d.	$K_\star$	second moment of $\mathrm{diag}(s^\star)\phi$
$h_t; H_D$	head of task t; the fast heads as columns	$V; \psi$	last shared layer; frozen representation (Proposition 1)
c_t	fine-tuning's weights on the fast tasks' changes, entries c_d	Δf_t	change of task t's prediction under fine-tuning
$L; \mathrm{dist}_G$	span of the past gains; distance in the metric of G	α_e	weights of a mixture of past windows
$r; u$	perturbation size; its direction, $s^\star = \bar{s} + ru$	R_t	risk on task t
$q; \ell_t(x; s)$	number of window coordinates; mean or logit at s	$\Sigma; \tau; v$	prior covariance; its scale; $\mathrm{tr}(\Sigma)/q$
s_0	pilot fit	Σ_ε	noise covariance of the arrived labels
$\theta; \hat{\theta}$	change $s^\star - \bar{s}$; its estimate	$\theta_R; \theta_U$	read part of θ; its part in U
$\theta_A; \theta_W$	full-prior update; isotropic update	$\delta_t^\star; \beta_t$	projected update $(I - P_t)\theta_W$; a bet in U
$\mathrm{Reg}_t; \Gamma_v$	regret against not adapting; class of laws of Theorem 1	$\rho; \hat{\rho}_t$	share of a bet the world delivers; its estimate on matured windows
$F; C$	cluster-adjusted information; $(I + vF)^{-1}$	$G_{t,RU}; G_{t,UU}$	blocks of G_t, read directions by U and U by U
$f_j, b_j; g_{tj}$	eigenpairs of F; $b_j^\top G_t b_j$	λ_j	information vf_j relative to the prior
$\odot$	entrywise product	$^+$	Moore–Penrose inverse
S	sample covariance of the past coordinates ($\widehat{\Sigma}_0$ in the appendix)		
Symbols used only in the appendix			
W	readouts as rows	S	past gains as rows, columns $S_{:j}$
Y_e	slice of window e (Appendices D.3, D.4 redefine it)	K_ϕ	$\mathbb{E}[\phi\phi^\top]$
$n_s; \sigma_s^2$	labels per task per past window; their noise variance	n	current labels per fast task
h	smallest gap between the two windows' gain ratios	$c_{\mathrm{in}}; c_{\mathrm{out}}$	$\|a_1\| + \|a_2\|; \|\omega\|_\infty$: new ratios in, off the past span
$\widehat{\Sigma}_0; v_0$	sample covariance of the past coordinates, the S of Section 4; $\mathrm{tr}(\widehat{\Sigma}_0)/q$	ζ_t	unread share, $\mathrm{tr}(QG_t)/\mathrm{tr}\, G_t$
$\pi_t; \nu_t$	θ_W's risk left to the past; its label noise, equation 11	$\oslash$	entrywise division
$\phi_\star$	current factor scores $\mathrm{diag}(s^\star)\phi$	$a_1, a_2; \omega$	$s^\star \oslash s_2 = a_1\mathbf{1} + a_2(s_1 \oslash s_2) + \omega$
$\kappa; \kappa_N$	Gram lower bounds, $G_D \succeq \|D\|\kappa I$; $A \succeq N\kappa_N P$	N	number of arrived labels
$A; P$	$Z^\top Z$; projection A^+A	M	factor map, $\phi(x) = Mx$, $x \in \mathbb{R}^p$
$\underline{\varrho}; \varrho^\star; \varrho_c$	gain ratios $s_1 \oslash s_2$; $s^\star \oslash s_2$; $a_1\mathbf{1} + a_2\hat{\varrho} + \omega$	Θ	change of factor basis
$\bar{h}$	$\min(h, 1)$	ϑ	sine of the gain-profile angle
$X_e; X_t$	input designs of past window e; of fast task t	$\mathcal{I}$	doubly centred task-by-window interaction
$\Upsilon; \tilde{w}_t$	full adapter's $k \times k$ matrix on pooled $\tilde{\phi}$; heads	$\Gamma; b_\Gamma$	gate matrix; gate offset
$\tilde{s}_e; \varsigma$	fitted past coordinates; their r.m.s. scale	$H; \mathcal{U}_0$	model-based information; score at s_0
$\ell_0; p_0$	arrived labels' predictions at s_0; $\mathrm{sig}(\ell_0)$	$\Omega; \widehat{\Omega}$	cluster-score covariance; its sandwich estimate
$\mathcal{U}_\varepsilon; \mathcal{U}_\iota$	score parts from label noise; from remainders	$\iota_t; \iota_D$	remainders beyond the model: task t; arrived labels
Ξ	fixed linear update, $s_\Xi = \bar{s} + \Xi\hat{\theta}$	ϵ_t	bound on the structural bias of task t
$\bar{\iota}_t; \bar{\iota}_D$	bounds on the sizes of ι_t and ι_D	$\xi_t; \xi_{t\circ}$	feature scales, $\|G_t\|_{\mathrm{op}} \le \xi_t^2$; $\|\hat{M}_{t\circ}\|_{\mathrm{op}} \le \xi_{t\circ}$
$J_D; J_t$	Jacobians of the shared parameters (Appendix B)	$\chi_D; \chi_t$	drift of the Jacobians over $\mathcal{K}$ (Lemma 7)
$\Lambda_t; \bar{\Lambda}_t$	$\lambda_{\max}(A^{+1/2} G_t A^{+1/2})$; population analogue	$\sigma_j(\cdot)$	jth singular value
sig	logistic function	Π	orthogonal projection
$\mathcal{O}; \mathfrak{a}$	arrived label pairs; arrival level of slow labels	p_{fail}	failure probability
Constants of Proposition 2; also C, C_i, c_i: positive constants, local to Appendices D and E.3			
ϵ_Y	largest slice error $\max_e \|\hat{Y}_e - Y_e\|_{\mathrm{op}}$	$\bar{\epsilon}$	its effect on the pencil, $C\epsilon_Y(\beta^{-1} + b_1\beta^{-2})$
$\beta; b_1$	$\sigma_k(Y_2)$; $\|Y_1\|_{\mathrm{op}}$	$\eta_W; \eta_M$	operator-norm errors of $\hat{W}$ and $\hat{M}$
η_{fac}	factor error, $\eta_W\|M\|_{\mathrm{op}} + (\|W\|_{\mathrm{op}} + \eta_W)\eta_M$	η_{row}	row-error bound $\epsilon_Y c_a + \|\omega\|_\infty \eta_{\mathrm{fac}}$
Δ_t	row t of $Y_\star - W\,\mathrm{diag}(\varrho_c)M$	κ_s	past design Gram at least $n_s\kappa_s I_p$
d_s	$T + p + \log(C/p_{\mathrm{fail}})$	c_a	$2\|a_1\| + C(1 + b_1/\beta)\|a_2\|$, of order c_{in}
C_{cond}	set by k and the bounds of Proposition 2; also c_{cond}		

Proposition 2 (Risk with finitely many early labels)**.** *Under the assumptions of Lemma 1, let the noise covariance be $\sigma^2 I$. The least-squares correction of the past mean, $\hat{s} = \bar{s} + Z^+(y - o - Z\bar{s})$, has risk on task t, averaged over label noise and a served input,*

$$R_t = \underbrace{\sigma^2\,\mathrm{tr}\big(G_t(Z^\top Z)^+\big)}_{\textit{label noise}} + \underbrace{\big\|G_t^{1/2} Q\,(s^\star - \bar{s})\big\|^2}_{\textit{unread change}}, \tag{6}$$

and when $QG_tQ = 0$ the correction gives the best linear unbiased estimate of task t's mean. If $QG_tQ \neq 0$ and the unread gains are unrestricted, every estimator that uses only these labels has unbounded worst-case risk on t.

Proof of Proposition 2. Write $y = o + Zs^\star + \varepsilon$. Since $Z^+ = A^+Z^\top$ and $Z^+Z = P$, $\hat{s} - s^\star = \bar{s} - s^\star + P(s^\star - \bar{s}) + A^+Z^\top\varepsilon = -Q(s^\star - \bar{s}) + A^+Z^\top\varepsilon$. The second term has mean zero and covariance $\sigma^2 A^+$. For a served input x independent of the labels the cross term vanishes on averaging over the noise, and $\mathbb{E}[\{z_t(x)^\top(\hat{s} - s^\star)\}^2]$ equals equation 6. When $QG_tQ = 0$, that is $\ker Z \subseteq \ker G_t$, the refit is the best linear unbiased estimate by Proposition 3 below; Remark 1 gives the unbounded risk when $QG_tQ \neq 0$. □

Remark 1 (An unread direction leaves the risk unbounded). If $\Delta \in \ker Z$ has $\Delta^\top G_t \Delta > 0$, every estimator from the arrived labels has unbounded worst-case risk on t.

The gains $s^\star$ and $s^\star + r\Delta$ give the arrived labels the same law, while the task means differ by $r^2\Delta^\top G_t\Delta$ in squared norm, so the larger of any estimator's two risks is at least $r^2\Delta^\top G_t\Delta/4$, for every $r > 0$. With a general noise covariance Σ_ε the first term of equation 6 becomes $\operatorname{tr}(G_tA^+Z^\top\Sigma_\varepsilon ZA^+)$: coverage concerns conditional means and does not need independent task noise. Evaluating on units that share noise with the arrived labels adds a cross term $-2z_t^\top A^+Z^\top \operatorname{Cov}(\varepsilon, \varepsilon_t)$ of either sign, which disjoint evaluation units remove (Runcie & Cheng, 2019).

Proof of the factor form of coverage in Lemma 1. Suppose, as the lemma's last sentence does, that $\mathbb{E}[\phi\phi^\top]$ is nonsingular on the arrived and on the served inputs; write K_ϕ for it on the arrived inputs. Let W_D hold the rows of W in D, so that $\|W_{D,:j}\|^2 = \sum_{d\in D} w_{dj}^2$ is the weight with which the fast tasks read factor j, and $G_D = K_\phi \odot (W_D^\top W_D)$. Then $\Delta^\top G_D\Delta = \sum_{d\in D}(w_d \odot \Delta)^\top K_\phi (w_d \odot \Delta)$ vanishes if and only if $w_{dj}\Delta_j = 0$ for all $d \in D$ and all j, so $\ker G_D = \{\Delta : \Delta_j = 0 \text{ whenever } W_{D,:j} \neq 0\}$. The same argument with the nonsingular second moment of the served inputs gives $\ker G_{t^\circ} = \{\Delta : \Delta_j = 0 \text{ whenever } w_{t^\circ j} \neq 0\}$. Task t° is covered if and only if every factor it reads is read by some fast task. □

Moreover $K_\phi \succeq \lambda_{\min}(K_\phi) I$ gives $\operatorname{diag}(w_d) K_\phi \operatorname{diag}(w_d) \succeq \lambda_{\min}(K_\phi) \operatorname{diag}(w_d \odot w_d)$, and summing, $G_D \succeq \lambda_{\min}(K_\phi)\, I \odot (W_D^\top W_D)$, the eigenvalue bound for entrywise products of Schur (1911, Satz VII). Coverage is a condition on unit-level inputs: one fast task that reads every factor covers them if its inputs vary in all k directions, whereas one aggregate mean per task gives the design $W_D \operatorname{diag}(\mathbb{E}\phi)$, of rank at most $|D|$. In the running example ($K_\phi = I$), click and short watch give $G_D = \operatorname{diag}(1.04, 1.09, 0)$ and purchase has $G_{t^\circ} = \operatorname{diag}(.16, 1, .64)$, so purchase is not covered; full watch alone gives $G_D = \operatorname{diag}(.09, .36, 1)$ and covers all three factors.

New-label rate for general k. Let each fast task have n labels, and let the arrived and the served inputs share K_ϕ. Under target coverage $\ker Z \subseteq \ker G_{t^\circ}$ the conditional least-squares target risk is $\sigma^2 \operatorname{tr}(G_{t^\circ}A^+)$. Under full factor coverage, $G_D \succ 0$, and on the event $A \succeq \frac{n}{2}G_D$, it is at most $2\sigma^2 \operatorname{tr}(G_{t^\circ}G_D^{-1})/n$, with $\operatorname{tr}(G_{t^\circ}G_D^{-1}) \leq \operatorname{cond}(K_\phi)\|w_{t^\circ}\|^2/\min_j \|W_{D,:j}\|^2$: the Loewner order reverses under inversion, $G_D^{-1} \preceq [\lambda_{\min}(K_\phi)\min_j \|W_{D,:j}\|^2]^{-1} I$ by the bound above, and $\operatorname{tr} G_{t^\circ} = \sum_j w_{t^\circ j}^2 (K_\phi)_{jj} \leq \lambda_{\max}(K_\phi)\|w_{t^\circ}\|^2$. This is $O(\sigma^2 k/(n|D|))$ when $\operatorname{cond}(K_\phi)$ is bounded, $\|w_{t^\circ}\|^2 = O(k)$ and $\min_j \|W_{D,:j}\|^2$ is at least a constant times $|D|$. If the whitened unit contributions $G_D^{-1/2}(\sum_{d\in D} z_{di}z_{di}^\top)G_D^{-1/2}$ are independent with eigenvalues in $[0, b]$, each has mean I, and the matrix Chernoff lower tail gives $\Pr\{A \not\succeq \frac{n}{2}G_D\} \leq k\, e^{-n/(8b)}$. For one full-support fast task with Gaussian z_d and $n > k + 1$, A is Wishart with scale G_D, $\mathbb{E}A^{-1} = G_D^{-1}/(n - k - 1)$, and the risk is exactly $\sigma^2 \operatorname{tr}(G_{t^\circ}G_D^{-1})/(n - k - 1)$. In the running example, with known factors, purchase's risk is asymptotically $5.20\,\sigma^2/n$ with full watch alone and $1.47\,\sigma^2/n$ with all three fast tasks.

A.2 The refit is the best linear unbiased estimate

Is the least-squares refit the best use of the early labels? For a covered task it is the best linear unbiased estimate, and with Gaussian noise it is also minimax. Condition on correct factors and

subtract $o + Z\bar{s}$ from the arrived labels, so that $\tilde{y} = y - o - Z\bar{s} = Z\theta + \varepsilon$ with $\theta = s^\star - \bar{s}$, $\mathbb{E}[\varepsilon \mid Z] = 0$ and $\mathrm{Cov}(\varepsilon \mid Z) = \sigma^2 I$. Let $G \succeq 0$ be a task metric G_t, or a fixed nonnegative combination of several.

Proposition 3 (Optimality of the refit)**.** *If* $\ker Z \subseteq \ker G$*, the least-squares estimate* $\hat{\theta} = Z^+\tilde{y}$ *gives the minimum-variance linear unbiased estimates of the corresponding task predictions, with conditional risk* $\sigma^2 \operatorname{tr}(GA^+)$*. If also* $\varepsilon \mid Z \sim \mathcal{N}(0, \sigma^2 I)$*, then* $\inf_{\tilde{\theta}} \sup_{\theta \in \mathbb{R}^k} \mathbb{E}_\theta \|G^{1/2}(\tilde{\theta} - \theta)\|^2 = \sigma^2 \operatorname{tr}(GA^+)$*. If* $\ker Z \not\subseteq \ker G$*, every estimator has infinite maximum risk.*

Proof. If $\ker Z \subseteq \ker G$, then $G^{1/2}P = G^{1/2}$, and the least-squares operator for $G^{1/2}\theta$ is $\Xi_0 = G^{1/2}Z^+$. Every linear unbiased estimator $\Xi\tilde{y}$ satisfies $\Xi Z = G^{1/2}$, so $\Xi = \Xi_0 + E$ with $EZ = 0$; as $Z^+ = A^+Z^\top$, $\Xi_0 E^\top = 0$ and $\mathrm{Cov}(\Xi\tilde{y}) - \mathrm{Cov}(\Xi_0\tilde{y}) = \sigma^2 EE^\top \succeq 0$, and taking traces gives the risk. Stacking the linear functions of all tasks gives the conclusion for all of them at once (Gauss–Markov). For the minimax value, the prior $\theta \sim \mathcal{N}(0, b^2 I_k)$ has posterior covariance $(b^{-2}I_k + \sigma^{-2}A)^{-1}$, so every estimator's maximum risk is at least $\operatorname{tr}\{G(b^{-2}I_k + \sigma^{-2}A)^{-1}\}$, which tends to $\sigma^2 \operatorname{tr}(GA^+)$ as $b \to \infty$ because G vanishes on $\ker A$; least squares has exactly this constant risk. If coverage fails, Remark 1 with G in place of G_t gives infinite maximum risk. □

Known unequal noise covariances lead to generalized least squares; estimated factors add the bias bounded by Theorem 2 (Appendix D); and for binary tasks Proposition 16 (Appendix E.1) takes the place of this result.

Remark 2 (One known direction). If $s^\star = \bar{s} + r\Delta$ for a known direction Δ, a task affected by Δ is identified exactly when $\sum_{d \in D} \mathbb{E}(z_d^\top \Delta)^2 > 0$, and its least-squares risk is $\sigma^2 \mathbb{E}(z_t^\top \Delta)^2 / \|Z\Delta\|^2$ with $\|Z\Delta\|^2 = \sum_i (z_{d_i}(x_i)^\top \Delta)^2 > 0$. This is Proposition 3 for the one-dimensional design $Z\Delta$; population coverage does not ensure that a finite design has the same rank.

A.3 Proof of Proposition 4, and other comparator classes

Let $\delta_{t,e} = \mu_{t,e} - f_t^0$ be task t's residual in window e and $L = \operatorname{span}\{s_1, \ldots, s_m\}$. Distances use the metric of a matrix G: $\operatorname{dist}_G(p, L) = \min_{b \in L} \|G^{1/2}(p - b)\|$. Even the best fixed mixture keeps an error, its floor.

Proposition 4 (Fixed relations have floors)**.** *Under equation 1, with the base and the past residuals known, coefficients that do not depend on* x*, and no added intercept, the best fixed mixture of past windows has error*

$$\min_{\alpha} \left\| \mu_{t^\circ,\star} - f_{t^\circ}^0 - \textstyle\sum_e \alpha_e \, \delta_{t^\circ,e} \right\|^2 = \operatorname{dist}_{G_{t^\circ}}(s^\star, L)^2, \tag{7}$$

and the best fixed mixture $\sum_{d \in D} c_d \, \delta_{d,\star}$ *of fast tasks has error* $\operatorname{dist}_{K_\star}(w_{t^\circ}, \operatorname{span}\{w_d : d \in D\})^2$*, where* $K_\star$ *is the second moment of the current factor scores* $\operatorname{diag}(s^\star)\phi$*, all on the served inputs. For a covered task, the refit of the gains errs by zero in the population.*

We prove it and ask whether complete predictions, an intercept or more windows change the floors.

Proof of Proposition 4. A combination of past windows predicts $f_{t^\circ}^0 + \sum_e \alpha_e \delta_{t^\circ,e} = f_{t^\circ}^0 + z_{t^\circ}^\top S^\top \alpha$, so its error is $z_{t^\circ}^\top (s^\star - S^\top \alpha)$, whose squared norm over the served inputs is $\|G_{t^\circ}^{1/2}(s^\star - S^\top \alpha)\|^2$; minimizing over α gives equation 7. Equivalently, the minimum is $s^{\star\top}\{G_{t^\circ} - G_{t^\circ} S^\top (S G_{t^\circ} S^\top)^+ S G_{t^\circ}\} s^\star$; the image of a finite-dimensional linear map is closed, so the minimum is attained even for singular G_{t°. In the current window $\delta_{d,\star} = w_d^\top \psi$ with $\psi = \operatorname{diag}(s^\star)\phi$, so a combination of fast tasks errs by $(w_{t^\circ} - W_D^\top c)^\top \psi$, and replacing $(s^\star, S, G_{t^\circ})$ by $(w_{t^\circ}, W_D, K_\star)$ gives the fast-task floor. □

Corollary 1 (Zero floors and the full adapter)**.** *The floors of Proposition 4 vanish exactly when* $s^\star \in L + \ker G_{t^\circ}$ *and when* $w_{t^\circ} \in \operatorname{span}\{w_d : d \in D\} + \ker K_\star$*, and a full adapter identifies* t° *from the fast labels exactly when* $\tilde{w}_{t^\circ}$ *lies in the row span of the fast heads* $\tilde{W}_D$*.*

A *full adapter* is $f_t^0 + \tilde{w}_t^\top \Upsilon \tilde{\phi}(x)$, with an unrestricted $k \times k$ matrix Υ on a pooled layer $\tilde{\phi}$ of nonsingular second moment. The floors are the distances of Proposition 4; multiplying the squared

Euclidean distances by the smallest eigenvalue of the metric gives lower bounds, since $\Delta^\top G\Delta \geq \lambda_{\min}(G)\|\Delta\|^2$. In the running example $G_{t^\circ} = \mathrm{diag}(.16, 1, .64)$, and the best combination of the two weeks errs by $.21\,r^2$, and by at least $.16\,r^2$; purchase is no multiple of full watch, so every combination of full watch errs by at least .861 at $r = 0$ and 1.52 at $r = 1$. For the adapter, the fast means identify $\tilde{W}_D\Upsilon$, and otherwise there is Δ with $\tilde{W}_D\Delta = 0$ and $\tilde{w}_{t^\circ}^\top\Delta \neq 0$, so that replacing Υ by $\Upsilon + \Delta b^\top$ with $b \neq 0$ changes the slow task and leaves every fast mean unchanged.

Mixtures of complete predictions and a fitted intercept. For complete means, $\mu_{t,\star} - \sum_e \alpha_e \mu_{t,e} = (1 - \mathbf{1}^\top\alpha) f_t^0 + z_t^\top(s^\star - S^\top\alpha)$. With $\mathbf{1}^\top\alpha = 1$ the reachable gains form the affine hull of the past gains, which lies in L, so the floor equation 7 still holds; if the past gains are centred, their affine hull equals L, because a coefficient vector can be shifted by a multiple of $\mathbf{1}$ without changing $S^\top\alpha$. Without the sum constraint the floor can fail: with independent unit-variance inputs x_1, x_2, $f_t^0 = x_2$, $z_t = (x_1, x_2)$, past gains $\pm\mathbf{e}_1$ and current gain $\mathbf{e}_2$, the residual floor is one, yet the sum of the two complete past means is exactly the current mean $2x_2$. Combinations of complete task predictions likewise involve $f_t^0 - \sum_d c_d f_d^0$. A free fitted intercept centres the features, which replaces $\mathbb{E}[zz^\top]$ by $\mathrm{Cov}(z)$ in equation 7 and $K_\star$ by $\mathrm{Cov}(\psi)$ in the fast-task floor. The past-window combination and the fast-to-slow regression of Section 5.1 combine complete predictions, the regression with an intercept; the known-truth experiments of Appendix G use the residual classes, where the floors apply exactly.

Remark 3 (Many past windows). The rank of S, not the number of past windows, decides whether a fixed mixture of them can follow a new window.

If $s^\star \in \mathrm{range}(S^\top) + \ker G_{t^\circ}$, a fixed mixture of past windows represents the slow task in the current window; when $\mathrm{rank}\, S = k$ this holds for every current gain. Repeating the regular and the promotion week adds samples but no directions, and the ranks of S and of the centred $S - \mathbf{1}\bar{s}^\top$ can differ by one. Once $\mathrm{rank}\, S = k$, fitting k gains instead of one weight per window is a reparametrization, and conditioning and sample size decide the comparison, which Theorem 2 quantifies for two windows.

A.4 The blind floor

What does a model that is not told the window lose, and why do per-task averages hide it? Such a model predicts, for each input, the relation averaged over windows; Proposition 5 prices this.

Proposition 5 (Blind floor)**.** *Under squared loss, the best predictor that does not see the window, $\bar{\mu}_t(x) = \mathbb{E}[\mu_{t,e}(x) \mid x]$, has risk above that of the best window-aware predictor by exactly $V = \sum_t \mathbb{E}_x \mathrm{Var}(\mu_{t,e}(x) \mid x)$, which under equation 1 equals $V = \mathbb{E}_x\,\mathrm{tr}\{W\,\mathrm{diag}(\phi)\mathrm{Cov}(s_e \mid x)\,\mathrm{diag}(\phi)W^\top\}$ and is zero when x determines e. At every x the residual $\mu_{t,e}(x) - \bar{\mu}_t(x)$ of every task averages to zero over windows. Under per-task log loss the gap is $\sum_t I(y_t; e \mid x)$.*

Proof. Conditional expectation is the orthogonal projection in L^2: for every predictor $\hat{\mu}$ that depends on x alone, $\mathbb{E}\|y - \hat{\mu}(x)\|^2 - \mathbb{E}\|y - \mu_e(x)\|^2 = \mathbb{E}\|\mu_e(x) - \hat{\mu}(x)\|^2$, where $\mu_e(x) = (\mu_{t,e}(x))_t$. The minimum is at $\hat{\mu} = \mathbb{E}[\mu_e(x) \mid x]$, with value $\sum_t \mathbb{E}_x \mathrm{Var}(\mu_{t,e}(x) \mid x)$, and $\mathbb{E}[\mu_{t,e}(x) - \bar{\mu}_t(x) \mid x] = 0$ is the averaging statement. Under equation 1, $\mu_{t,e}(x) - \bar{\mu}_t(x) = z_t(x)^\top(s_e - \mathbb{E}[s_e \mid x])$, which gives the trace form; if x determines e, $\mathrm{Cov}(s_e \mid x) = 0$. This is the squared-loss form of the λ term of Ben-David et al. (2010, Thm. 2), and the log-loss statement is the chain rule for conditional entropy applied task by task. □

Remark 4 (Averages cannot see the interaction). Reweighting windows only selects another blind predictor and cannot remove V. With $\mathcal{M}(x) = [\mu_{t,e}(x)]_{t,e}$ and the centring projections $\Pi_T = I_T - T^{-1}\mathbf{1}\mathbf{1}^\top$ and Π_m, the task-by-window interaction is $\Pi_T\mathcal{M}(x)\Pi_m = (\Pi_T W)\,\mathrm{diag}(\phi(x))(\Pi_m S)^\top$ (Fisher, 1925), of rank at most the number of factors whose gains vary. A doubly centred matrix $\mathcal{I}$, such as this interaction, has zero row and column sums, so adding it to a strictly positive task-by-window loss matrix, scaled to keep every entry nonnegative, changes no task or window average. Procedures that act only on these averages cannot see it; those that use task gradients or window derivatives, such as PCGrad (Yu et al., 2020) or IRM (Arjovsky et al., 2019), are not functions of them.

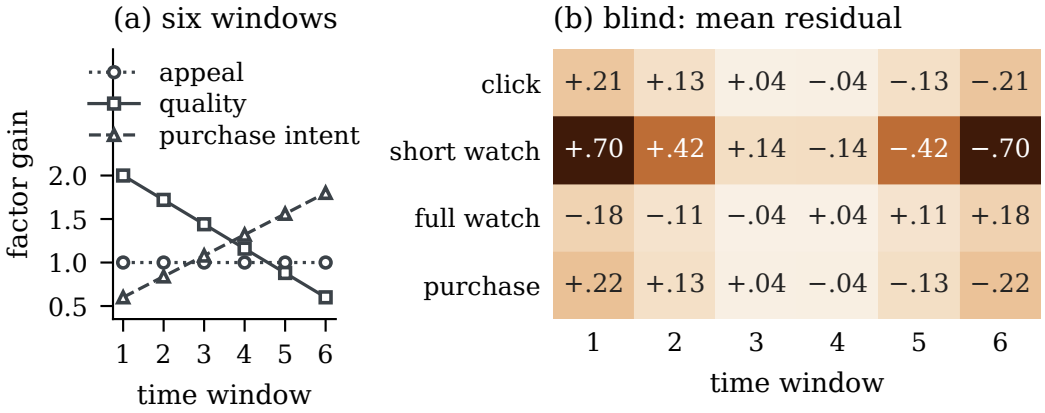


Figure 5: The blind floor in the running example over six windows: (a) gains; (b) mean residual of a blind multi-task model, each row averaging to zero, shaded by its size. The same model with per-window gains keeps every residual below .0033.

Illustration. Figure 5 uses the readouts of Table 7 over six consecutive windows: the appeal gain stays at 1, and in equal steps the quality gain falls from 2 to .6 while the purchase-intent gain rises from .6 to 1.8; $x \sim \mathcal{N}(\mathbf{1}, I_3)$ with $\mathbf{1} = (1, 1, 1)$, $\phi(x) = x$, $\sigma = 0.1$, and 4,000 training and 20,000 test units per window. A blind multi-task network and one with per-window gains share architecture and training (Adam, learning rate 10^{-2}, 3,000 steps; means over seeds). The blind residuals match $w_t^\top \operatorname{diag}(s_e - \bar{s})\mathbf{1}$ within .015 in all 24 cells and average to zero over windows for every task, so every per-task summary reports an unbiased model, yet single cells are off by up to .70 in the first and last windows, whose gains moved furthest from the average; per-window gains remove the pattern (Figure 5).

A.5 A CHANGE THAT ONLY THE SLOW TASK READS

Can the fast labels at least detect a change they cannot identify? Not if it is confined to the slow task.

Proposition 6 (Invisible changes)**.** *A change confined to the slow task leaves every fast-label law unchanged. Hence, between two worlds whose slow means are $\pm r$ and whose other laws agree, every estimator has worst-world squared error at least r^2, and every threshold decision has worst-world error probability at least $\frac{1}{2}$. Inside model equation 1, such worlds exist whenever coverage fails: take the gains $s^\star \pm r'\Delta$ with $\Delta \in \ker G_D \setminus \ker G_{t^\circ}$.*

Proof. The two worlds give every estimator or decision the same law. For any random estimate $\hat{f}$, $\frac{1}{2}\{\mathbb{E}(\hat{f} - r)^2 + \mathbb{E}(\hat{f} + r)^2\} = \mathbb{E}\hat{f}^2 + r^2 \geq r^2$, and the probabilities of the two opposite threshold mistakes sum to one, so the larger is at least $\frac{1}{2}$. For the gain construction, Lemma 1(i) makes the fast laws equal, while $\Delta \notin \ker G_{t^\circ}$ separates the slow means. □

A test on the fast labels therefore cannot flag such a change, however many labels arrive. With several fast tasks, a nested test of the shared-gain restriction (Sargan, 1958; Hansen, 1982) detects violations that the fast labels show, but it has no power against a change confined to the slow task, nor when the fast readouts are proportional: on a rotated factor basis, one pair of diagonal gains fits the fast readouts $(1, .5)$ and $(2, 1)$ exactly while the slow excess risk is 5 (the second instance of Remark 6, Appendix D.5).

B HOW DOES FINE-TUNING MOVE A TASK IT NEVER SEES? (SECTION 3.3)

This appendix proves Proposition 1, says which optimizers keep its exact form, derives what it implies for the running example and for multi-task optimizers, and checks it numerically.

Proof of the exact form. Let label i, of fast task d_i at input x_i, have loss $\ell(f_{d_i}(x_i), y_i)$ for any loss differentiable in the prediction, squared or logistic, and let ℓ_i' be that derivative. The gradient of this loss with respect to M is $\ell_i'\, h_{d_i}\psi(x_i)^\top$, and h_{d_i} is a column of H_D. A step that is a linear combination of per-example gradients, with coefficients that may depend on the data, on the iterate or on earlier steps, is therefore $\sum_i a_i h_{d_i}\psi(x_i)^\top$ for some scalars a_i, and so is any sum of such steps: momentum and Nesterov accumulate past gradients, task weights and minibatches rescale or drop terms, clipping by the global norm rescales a step, PCGrad (Yu et al., 2020) subtracts from one task's gradient its projection onto another's, and MGDA (Sener & Koltun, 2018) takes a convex combination of the per-task gradients, each of them a combination of per-example gradients. After any number of steps $M = M_0 + \sum_i a_i h_{d_i}\psi(x_i)^\top$, so for every input x, arrived or not, $\Delta M\psi(x) = H_D a(x)$ with $a(x) = \sum_i a_i\, \psi(x_i)^\top\psi(x)\, e_{d_i} \in \mathbb{R}^{|D|}$. The heads are fixed, hence $\Delta f_t(x) = h_t^\top H_D a(x)$ and $\Delta f_D(x) = H_D^\top H_D a(x)$. Let $\Pi = (H_D^\top H_D)^+ H_D^\top H_D$, the orthogonal projection onto $\mathrm{range}(H_D^\top H_D) = \mathrm{range}(H_D^\top)$; then $H_D(I - \Pi) = 0$, because $I - \Pi$ projects onto $\ker(H_D^\top H_D) = \ker H_D$. So

$$c_t^\top \Delta f_D(x) = h_t^\top H_D (H_D^\top H_D)^+ H_D^\top H_D\, a(x) = h_t^\top H_D \Pi a(x) = h_t^\top H_D a(x) = \Delta f_t(x),$$

which is equation 3, at every step and every input. A bias of the shared layer is a constant coordinate of ψ, and fixed task biases b_t do not move.

What keeps the exact form and what breaks it. Weight decay toward the pretrained weights adds a multiple of $-\Delta M$ to a step and keeps the form above; decay toward zero adds a multiple of $-M_0$, whose columns need not lie in the range of H_D, and Adam scales each coordinate of the gradient by its own factor, which leaves that range too. If the fast heads are trained as well, Δf_D includes the heads' own moves and H_D changes along the path, and with it c_t; the theorem therefore fixes the heads. Anything between M and the heads that depends on the input, such as a normalization layer or a nonlinear tower, replaces h_t by an input-dependent Jacobian and breaks the pointwise form; the first-order statement covers it.

All shared layers, to first order. Let ϑ collect the shared parameters, let row i of J_D be the gradient of $f_{d_i}(x_i)$ with respect to ϑ at the pretrained weights and $J_t(x)$ that of $f_t(x)$, and let $K_{DD} = J_D J_D^\top$ and $K_t(x, X_D) = J_t(x) J_D^\top$ be the cross-task tangent kernel (Jacot et al., 2018) on the arrived inputs X_D. A gradient-span optimizer moves ϑ by $J_D^\top a$ for some $a \in \mathbb{R}^N$, up to terms of second order in the move, since the per-example gradients change only by that order along the path. To first order, then, $\Delta f_D(X_D) = K_{DD} a$ and $\Delta f_t(x) = K_t(x, X_D) a$. If $K_{DD} c' = 0$ then $\|J_D^\top c'\|^2 = c'^\top K_{DD} c' = 0$, so $J_D^\top (I - K_{DD}^+ K_{DD}) = 0$ and

$$\Delta f_t(x) = K_t(x, X_D)\, K_{DD}^+\, \Delta f_D(X_D),$$

a linear map of the fast tasks' changes on the arrived inputs, fixed at the pretrained weights and fitted to no slow label. For the last shared layer the kernel is the product of a task part $h_t^\top h_d$ and an input part $\psi(x)^\top\psi(x')$, and the map reduces to equation 3; through deeper layers it is set by how the parametrization shares parameters between tasks.

Corollary 2 (What fine-tuning serves)**.** *Under the conditions of Proposition 1: (a) a fast task d has $c_d = \Pi e_d$ and $c_d^\top \Delta f_D = \Delta f_d$, so fine-tuning moves it by its own change alone, and $c_d = e_d$ when the fast heads are linearly independent; (b) under equation 1, if the network before fine-tuning serves the mean past gains and fine-tuning fits the fast tasks' changes $z_d^\top\theta$ exactly, then it errs on t° by at least $\mathrm{dist}_{K_\theta}(w_{t^\circ}, \mathrm{span}\{w_d\}_{d\in D})^2$, the fast-task floor of Proposition 4, for the change from the mean past gains, with K_θ the second moment of $\mathrm{diag}(\theta)\phi$ on the served inputs, however many labels arrive; (c) under (b), if $\mathbb{E}\phi\phi^\top \succ 0$ on the arrived and the served inputs, then for every $u \in U$ fine-tuning's improvement on t° satisfies $\mathcal{I}_{t^\circ}(\theta + u) = \mathcal{I}_{t^\circ}(\theta) + 2\theta^\top \sum_{d\in D} c_{t^\circ d} G_{dt^\circ} u$, with $G_{dt^\circ} = \mathbb{E} z_d z_{t^\circ}^\top$ on the served inputs, so its overlap with U is zero for every change if and only if $\sum_d c_{t^\circ d} G_{dt^\circ} B = 0$. This holds whenever $\mathbb{E}\phi\phi^\top$ is diagonal on the served inputs, whatever the heads.*

Proof. (a) $h_d = H_D e_d$, so $c_d = (H_D^\top H_D)^+ H_D^\top H_D e_d = \Pi e_d$ and $c_d^\top \Delta f_D(x) = e_d^\top \Pi H_D^\top H_D a(x) = \Delta f_d(x)$ because Π fixes $\mathrm{range}(H_D^\top)$; $\Pi = I$ when H_D has full column rank. (b) By equation 3 the served change of t° is $c_{t^\circ}^\top \Delta f_D = \sum_d c_d\, w_d^\top\, \mathrm{diag}(\theta)\phi$, while its

change is $w_{t^\circ}^\top \operatorname{diag}(\theta)\phi$; the weights c_d do not depend on x, so the error is at least the smallest over all such weights, which is the distance in K_θ as in the proof of Proposition 4(b). (c) With $\mathbb{E}\phi\phi^\top \succ 0$, U is spanned by the factors no fast task reads (Lemma 1), so $z_d(x)^\top u = 0$ for $d \in D$ and every input; a shift by u leaves the early labels, and so the fit and the served change $a = \sum_d c_{t^\circ d} z_d^\top \theta$, unchanged, and moves the change of t° from $b = z_{t^\circ}^\top \theta$ to $b + e$ with $e = z_{t^\circ}^\top u$. Then $\mathbb{E}(b+e)^2 - \mathbb{E}(a-b-e)^2 = \mathbb{E}b^2 - \mathbb{E}(a-b)^2 + 2\mathbb{E}[ae]$, and $\mathbb{E}[ae] = \theta^\top \sum_d c_{t^\circ d} G_{dt^\circ} u$. With a diagonal second moment, $G_{dt^\circ} u = \operatorname{diag}(w_d)\mathbb{E}[\phi\phi^\top]\operatorname{diag}(w_{t^\circ})u$ is zero, because w_d vanishes on the factors that carry u. □

Overlap checks. Over 2,000 random instances with four factors, three fast tasks that do not read the fourth and a slow task that reads all four, fine-tuning the last shared layer on the fast tasks' population loss follows equation 3 to 2×10^{-12} after 200 gradient steps and at convergence, where it fits the fast changes exactly; the slope in (c) matches the improvement along U to 3×10^{-11}; the overlap $\sum_d c_{t^\circ d} G_{dt^\circ} B$ is exactly zero for every diagonal $\mathbb{E}\phi\phi^\top$ and nonzero in all 2,000 instances with a random correlated one (median norm .29).

The running example. Fine-tune the last shared layer of a network with the whitened factors as ψ, the pretrained layer $M_0 = I$ and heads $h_t = \operatorname{diag}(\bar{s})w_t$, which serves the mean past gains, in the population and until the fast tasks are fitted exactly. Purchase is then served the change $\{(W_D^\top c_{t^\circ}) \odot \theta\}^\top \phi$, as if it read the factors with the readout $W_D^\top c_{t^\circ}$, where W_D holds the fast readouts as rows. With click and short watch, $c_{t^\circ} = (-.638, 1.191)$ reproduces purchase's readout on appeal and quality and leaves intent at its past value, and fine-tuning errs .124, as the refit does. With short and full watch, which cover purchase, $c_{t^\circ} = (.47, .68)$ gives the readout $(.30, .88, .68)$ instead of $(-.4, 1, .8)$, and fine-tuning errs .342, where the refit errs zero and not adapting .350; for a campaign that moves only appeal, it errs .490 against .160 for not adapting. With short watch alone it errs .358, above not adapting, and the refit .124. Gradient descent on the population losses reproduces these errors exactly; Adam, which the theorem excludes, reaches .327, .398 and .327 in the last three cases, still far from the refit.

Negative transfer without conflict. Under (A0) the expected loss of every task in the current window, squared or logistic, is minimized at the gains $s^\star$, since each task's mean there is $f_t^0 + z_t^\top s^\star$: the tasks share a minimizer, and none pulls against another. In the running example the last shared layer can reach it, since $M = \operatorname{diag}(s^\star \oslash \bar{s})$ serves every task its current mean and minimizes all four losses at once. In the population, every iterate of an optimizer covered by Proposition 1 lies in $M_0 + \{H_D N : N \text{ any matrix}\}$, and if H_D has full column rank a single layer there fits the fast tasks exactly; so every such optimizer that fits them, SGD, PCGrad and MGDA alike, ends at that layer, the minimizer of the fast losses nearest to M_0, and not at the common minimizer. Conflict-averse optimizers change the path, not the end point.

Numerical checks. A network with frozen random features ($\psi \in \mathbb{R}^8$), a last shared layer of width 5 and four tasks, two of them fast with 200 labels each on a target outside the model's form, is fine-tuned for 300 minibatch steps. On 500 held-out inputs the largest deviation from equation 3, relative to the slow tasks' largest change, stays of order 10^{-14} at every step under SGD, momentum, drifting task weights, PCGrad, which projected in 140 of the 300 steps, and MGDA, and is 1.05 under Adam, where the weights c_t explain .95 and .35 of the variance of the two slow tasks' changes. Through all layers, a network with two tanh layers and a last shared layer is pretrained on the regular week of the running example, with nonlinear factors, and fine-tuned on 300 arrived inputs of the campaign's fast labels. With the tangent kernel at the pretrained weights, and a ridge of 10^{-6} times its mean diagonal in place of the pseudoinverse, the first-order prediction explains more than .999 of the variance of both slow tasks' changes on 1,000 held-out inputs under SGD at width 256, and .35 and .92 under Adam; at width 32, where the weights move further, it explains .91 and .94 under SGD and $-.13$ and .60 under Adam.

C WHICH UPDATE IS SAFE? (SECTION 3.4)

This appendix states and proves Lemma 2 and Proposition 7, proves Theorem 1, splits any update into the projected update, a transport error and a bet equation 9, shows that no statistic of the early

labels can choose between updates that differ in the unread part, works out the examples of the main text, checks the results numerically, and restates, as supporting material, the split of the isotropic update's risk behind the threshold that defines U in Section 4.

C.1 Proofs of Lemma 2, Proposition 7 and Theorem 1

In proofs we use the *regret* against not adapting, the negative of the improvement equation 2, as in minimax-regret treatment choice (Manski, 2004):

$$\mathrm{Reg}_t(\delta;\theta) = -\mathcal{I}_t(\delta;\theta) = \mathbb{E}\|\delta-\theta\|^2_{G_t} - \|\theta\|^2_{G_t}. \tag{8}$$

Let R have orthonormal columns spanning the read directions, the orthogonal complement of U, and write $G_{t,UR} = B^\top G_t R$, and $G_{t,UU}$, $G_{t,RR}$ likewise. The *overlap* of an update is $B^\top G_t \mathbb{E}\delta$.

Lemma 2 (The unread part enters linearly)**.** *For every update δ computed from the early labels, every θ and every $u \in U$, $\mathcal{I}_t(\delta;\theta+u) = \mathcal{I}_t(\delta;\theta) + 2\langle \mathbb{E}\delta, u\rangle_{G_t}$. So the improvement is bounded below over the shifts $\theta+U$ if and only if the overlap $B^\top G_t\mathbb{E}\delta$ is zero, and then it is the same for all of them.*

Proposition 7 (Leaving the unread directions alone is not safe)**.** *If $G_{t,UR} \neq 0$, the overlap of θ_W is nonzero at some θ, and its worst-case improvement over $\theta+U$, and over Γ_v, is unbounded below.*

Setting. Subtract from the early labels the predictions of the base at the mean past gains, $\tilde{y} = Z\theta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,\sigma^2 I)$ independent of θ; a known noise covariance Σ_ε only replaces F by $Z^\top\Sigma_\varepsilon^{-1}Z$. Split $\theta = \theta_R + \theta_U$ into its Euclidean projections onto the read directions, the range of F, and onto $U = \ker Z$. Since $Z\theta_U = 0$, the law of $\tilde{y}$, and of every update δ computed from it, depends on θ only through θ_R; given θ_R, δ is independent of θ_U. In Appendices C.1 and C.2, inner products and norms are those of G_t, and vectors that differ by an element of $\ker G_t$, which no error of task t sees, are identified. Let B have orthonormal columns spanning U and $G_{t,UU} = B^\top G_t B$.

Three facts about P_t. (a) $P_tP_t = P_t$ and $G_tP_t = P_t^\top G_t$, so P_t is idempotent and self-adjoint in t's geometry, and $\langle (I-P_t)a, P_tb\rangle = 0$ for all a, b. (b) For $u = Bc \in U$, $P_tu - u = B(G^+_{t,UU}G_{t,UU} - I)c$, whose squared norm $c^\top(G^+_{t,UU}G_{t,UU}-I)G_{t,UU}(G^+_{t,UU}G_{t,UU}-I)c$ is zero: P_t fixes U, and $(I-P_t)u = 0$. (c) Hence, for $u \in U$, $\langle a, u\rangle = \langle P_ta, u\rangle$ and $\langle (I-P_t)a, u\rangle = 0$. Write $a_V = (I-P_t)a$ and $a_B = P_ta$; by (a) $\|a\|^2 = \|a_V\|^2 + \|a_B\|^2$, and by (b) $\theta_V = (I-P_t)\theta_R$ is a function of the read part, while $\theta_B = P_t\theta$ is the point of U nearest to θ in t's geometry.

Lemma 2, and the worst case over Γ_v. For $u \in U$ the labels have the same law at θ and at $\theta+u$, so $\mathbb{E}_{\theta+u}\delta = \mathbb{E}_\theta\delta$, and expanding equation 8 as $\mathbb{E}\|\delta\|^2 - 2\langle \mathbb{E}\delta, \theta\rangle$ gives $\mathrm{Reg}_t(\delta;\theta+u) = \mathrm{Reg}_t(\delta;\theta) - 2\langle \mathbb{E}_\theta\delta, u\rangle$, with $\langle \mathbb{E}_\theta\delta, u\rangle = \langle P_t\mathbb{E}_\theta\delta, u\rangle$ by (c). Under a law in Γ_v let $m(\theta_R) = \mathbb{E}[P_t\delta \mid \theta_R]$, which lies in U. Because δ is independent of θ_U given θ_R, the Bayes regret is

$$\mathbb{E}\ \mathrm{Reg}_t(\delta;\theta) = \mathbb{E}\|\delta\|^2 - 2\mathbb{E}\langle \delta, \theta_R\rangle - 2\mathbb{E}\langle m(\theta_R), \theta_U\rangle,$$

and its first two terms do not depend on the conditional law of θ_U given θ_R. The point mass at $\theta_U = -c\,m(\theta_R)$ is such a law for every $c > 0$, and it makes the last term $2c\,\mathbb{E}\|m(\theta_R)\|^2$, unbounded in c unless $\|m(\theta_R)\| = 0$ almost surely. An update with $\mathbb{E}\|\delta\|^2 = \infty$ has infinite regret already.

Proposition 7. θ_W lies in the read directions, the range of F, and $\mathbb{E}_\theta\theta_W = (I-C)\theta_R$ with $C = (I+vF)^{-1}$, which is invertible there because $f_j > 0$ on the read directions. Its overlap at θ is $B^\top G_t(I-C)\theta_R = G_{t,UR}R^\top(I-C)\theta_R$, nonzero for some θ_R exactly when $G_{t,UR} \neq 0$; by Lemma 2 its regret over $\theta+U$ is then unbounded. Over Γ_v, $m(\theta_R) = P_t(I-C)\theta_R$ is nonzero with positive probability under $\mathcal{N}(0, vI)$, so the worst case above is unbounded as well. The running example with $\mathbb{E}\phi\phi^\top = I$ has $G_{t,RU} = 0$, which is why there $\delta_t^* = \theta_W$.

Theorem 1(i). Let $\|m(\theta_R)\| = 0$ almost surely. Then $\mathbb{E}\langle \delta_B, \theta_B\rangle = \mathbb{E}\langle m(\theta_R), \theta_B\rangle = 0$, and by (a)

$$\mathbb{E}\ \mathrm{Reg}_t(\delta;\theta) = \mathbb{E}\big[\|\delta_V - \theta_V\|^2 - \|\theta_V\|^2\big] + \mathbb{E}\|\delta_B\|^2.$$

The right side depends on the law of θ only through the joint law of $(\theta_R, \tilde{y})$, which Γ_v fixes: $\theta_R \sim \mathcal{N}(0, vI)$ on the read directions and $\tilde{y} \mid \theta_R \sim \mathcal{N}(Z\theta_R, \sigma^2 I)$. Every such update therefore has

the same Bayes regret under every law in Γ_v, and by the previous paragraph every other update has infinite worst regret. The first term is smallest at $\delta_V = \mathbb{E}[\theta_V \mid \tilde{y}] = (I - P_t)\mathbb{E}[\theta_R \mid \tilde{y}]$, uniquely up to a null set, and the second at $\delta_B = 0$. By Gaussian conjugacy the posterior of θ_R has precision $v^{-1}I + F$ on the read directions and mean $(v^{-1}I+F)^{-1}Z^\top\tilde{y}/\sigma^2 = (v^{-1}I+F)^{-1}F\hat{\theta} = \theta_W$, because $Z^\top\tilde{y}/\sigma^2 = F\hat{\theta}$ for the least-squares estimate $\hat{\theta} = F^+Z^\top\tilde{y}/\sigma^2$. Hence the unique minimizer is $(I - P_t)\theta_W = \delta_t^*$, which meets the condition of the previous paragraph because $P_t(I - P_t) = 0$.

Theorem 1(ii). By (c), $\langle\delta_t^*, u\rangle = 0$ for $u \in U$, so by Lemma 2 $\mathrm{Reg}_t(\delta_t^*; \theta + u) = \mathrm{Reg}_t(\delta_t^*; \theta)$ for every θ and $u \in U$. Its Bayes regret is $\mathbb{E}\|\delta_t^* - \theta_V\|^2 - \mathbb{E}\|\theta_V\|^2 = -\mathbb{E}\|\delta_t^*\|^2$, because a posterior mean is uncorrelated with its error. Under the prior, $F\hat{\theta} = F\theta_R + Z^\top\varepsilon/\sigma^2$ has second moment $vF^2 + F$, so with $C = (I + vF)^{-1}$ and $vCF = I - C$, $\mathbb{E}[\theta_W\theta_W^\top] = v^2C(vF^2 + F)C = v(I - C)$, and the predicted gain is

$$\mathbb{E}\|\delta_t^*\|_{G_t}^2 = v\,\mathrm{tr}\left\{(I - P_t)^\top G_t(I - P_t)(I - C)\right\},$$

computable from F, G_t and v before any slow label. *The hedge and the Schur form.* With $a = R^\top\theta_W$, $\theta_W = Ra$ and $B^\top G_t\theta_W = G_{t,UR}a$, so $\delta_t^* = Ra - BG_{t,UU}^+G_{t,UR}a$ and $B^\top\delta_t^* = -G_{t,UU}^+G_{t,UR}a$, which is nonzero whenever $G_{t,UR}a \neq 0$, since $G_{t,UR}a$ lies in the range of $G_{t,UU}$ for a positive semidefinite G_t. Its overlap is $B^\top G_t\delta_t^* = G_{t,UR}a - G_{t,UU}G_{t,UU}^+G_{t,UR}a = 0$, and $\|\delta_t^*\|_{G_t}^2 = a^\top G_{t,RR}a - 2a^\top G_{t,RU}G_{t,UU}^+G_{t,UR}a + a^\top G_{t,RU}G_{t,UU}^+G_{t,UU}G_{t,UU}^+G_{t,UR}a = a^\top(G_t/G_{t,UU})a$. In prediction space, $B^\top\delta_t^*$ is minus the least-squares coefficient of $z_t^\top Ra$ on the unread-part features $z_t^\top B$ over the served inputs, and $\|\delta_t^*\|_{G_t}^2$ is the residual second moment of that regression.

Proposition 8 (Weakly read directions)**.** *Let U be spanned by eigenvectors of F, such as those with $\lambda_j < 1$ in Section 4, so that the early labels may read U weakly, and let $\delta_t^* = (I - P_t)\theta_W$. Then $\mathrm{Reg}_t(\delta_t^*; \theta + u) = \mathrm{Reg}_t(\delta_t^*; \theta)$ for all θ and $u \in U$, and under every law whose part orthogonal to U is $\mathcal{N}(0, vI)$ on the directions F informs, with any conditional law of its part in U, the average regret of δ_t^* is $-v\,\mathrm{tr}\{(I - P_t)^\top G_t(I - P_t)(I - C)\}$.*

Proof. Write $A = (v^{-1}I + F)^{-1}$ and $F\hat{\theta} = F\theta + \xi$ with $\xi \sim \mathcal{N}(0, F)$. Because U and its orthogonal complement $U^\perp$ are invariant under F, A and AF map each into itself, and ξ splits into independent parts ξ_U and $\xi_\perp$. Since $I - P_t$ annihilates U (fact (b)), $\delta_t^* = (I - P_t)A(F\theta_\perp + \xi_\perp)$, with $\theta_\perp$ the part of θ in $U^\perp$: its law depends on θ only through $\theta_\perp$, which a shift by $u \in U$ leaves unchanged, and by (c) $\langle\mathbb{E}\delta_t^*, u\rangle = 0$, so the regret is the same at θ and $\theta + u$. Its average is $\mathbb{E}\|\delta_t^*\|^2 - 2\mathbb{E}\langle\delta_t^*, \theta_V\rangle$ with $\theta_V = (I - P_t)\theta_\perp$. Under $\theta_\perp \sim \mathcal{N}(0, vI)$, $m = A(F\theta_\perp + \xi_\perp)$ is the posterior mean of $\theta_\perp$ given $F\theta_\perp + \xi_\perp$, so $\mathbb{E}[m^\top K\theta_\perp] = \mathbb{E}[m^\top Km]$ for $K = (I - P_t)^\top G_t(I - P_t)$, whatever the part in U does; the average regret is $-\mathbb{E}\|\delta_t^*\|^2$, and the second-moment computation above, restricted to $U^\perp$, gives its value, since $I - C$ maps U into itself and $I - P_t$ annihilates it. The labels' weak information on U could in principle be used by another update; Proposition 9 bounds what it can gain.

Proposition 9 (Near-minimax on a thresholded unread set)**.** *Under Proposition 8, with R the eigendirections of F outside U, F_R and F_U its blocks, $K = (I - P_t)^\top G_t(I - P_t)$ restricted to R, and H the map $\theta_R \mapsto (P_t\theta_R)_U$ in U's eigenbasis, let*

$$\Delta_t = \mathrm{tr}\left\{K\left[(v^{-1}I + F_R)^{-1} - (v^{-1}I + F_R + H^\top F_UH)^{-1}\right]\right\} \geq 0.$$

Over Γ_v (read part $\mathcal{N}(0, vI)$ on R, any conditional law on U), δ_t^*'s worst-case regret is $-g_t$, $g_t = \mathbb{E}\|\delta_t^*\|_{G_t}^2$, and no update's is below $-g_t - \Delta_t$. $\Delta_t = 0$ when $F_UH = 0$, in particular when $B^\top G_t(I - BB^\top) = 0$ or $U = \ker Z$, which gives Theorem 1(i)'s value.

Proof. The upper value is Proposition 8. For the lower one, take the law $\pi_0 \in \Gamma_v$ with $\theta_U = -(P_t\theta_R)_U$, so that the unread part $P_t\theta$ is zero and $\theta = \theta_V = (I - P_t)\theta_R$. Every update's worst-case regret is at least its Bayes regret under π_0, and the least Bayes regret is $-\mathbb{E}\|(I - P_t)\mathbb{E}[\theta_R \mid y]\|_{G_t}^2$. Under π_0 the early labels observe θ_R on R with precision F_R and $-H\theta_R$ on U with precision F_U, independently, so the posterior covariance of θ_R is $(v^{-1}I + F_R + H^\top F_UH)^{-1}$, against $(v^{-1}I + F_R)^{-1}$ for δ_t^*, which uses R alone; for nested Gaussian information the two squared norms differ

by the trace of K times the difference of the posterior covariances, which is Δ_t. If $F_U H = 0$ the U-labels carry no information about θ_R under π_0. Over 386 random instances with a coupled G_t, Δ_t is a median 0.15% of g_t (1.9% at the 90th percentile, 20% at most) and exactly zero in the 77 with uncorrelated blocks; Monte Carlo runs of 20,000 draws confirm δ_t^*'s regret $-g_t$ under π_0 and under a heavy-tailed unread part, and the Bayes regret $-g_t - \Delta_t$ under π_0, within 2.4 standard errors in all 58 instances checked.

What the past must supply. Section 3.4 takes from the past only a lower bound v on the size of a change. Let $K = (I - P_t)^\top G_t (I - P_t)$ on the read directions, $C = (I + vF)^{-1}$, and ζ_t^2 the largest eigenvalue of CKC relative to K; $\zeta_t \le 1$ whenever K commutes with F.

Proposition 10 (What the past must supply)**.** *(i) If the read part's true second moment is $v'I$ with $v' \ge v$, or is any $\Sigma' \succeq vI$ and $\zeta_t \le 1$, then δ_t^*'s average regret is at most $-\mathbb{E}\|\delta_t^*\|^2_{G_t}$. If $\zeta_t > 1$, the size-robust update $(I - P_t)\{\theta_W + (1 - \zeta_t^{-1})(\hat\theta - \theta_W)\}$, which shrinks less, has average regret at most minus its own predicted gain for every $\Sigma' \succeq vI$; that gain is never negative when every read direction has $\lambda_j = v f_j \ge 1$. (ii) If the read part's law is unrestricted, not adapting is the only update whose worst-case regret is not positive.*

The past must therefore supply a lower bound on the size of a change, and this floor shapes the update three ways: it leaves the unread part alone; on each read direction the labels' noise must lie below it, $\lambda_j \ge 1$, so that even the unshrunk estimate is safe there; and it borrows no shrinkage between directions that t weighs unequally, which ζ_t measures.

Proposition 10(i). Let the read part θ_R have second moment $\Sigma' \succeq vI$ and the part in U any conditional law; by (ii) and fact (c) only the read part enters the average regret. On the read directions $\hat\theta = \theta_R + \eta$ with $\eta \sim \mathcal{N}(0, N)$, $N = F^{-1}$, and $\theta_W = (I - C)\hat\theta$. For an update $(I - P_t)L\hat\theta$ with a fixed matrix L, since $I - P_t$ is the G_t-orthogonal projection along U (fact (b)), with $K = (I - P_t)^\top G_t (I - P_t)$,

$$\mathbb{E}\ \mathrm{Reg}_t = \mathrm{tr}\{[(L - I)^\top K (L - I) - K]\,\Sigma'\} + \mathrm{tr}(L^\top K L N),$$

which is affine in Σ': its increase from vI to Σ' is $\mathrm{tr}\{[(L-I)^\top K(L-I) - K](\Sigma' - vI)\}$. It is non-positive for every $\Sigma' \succeq vI$ if and only if $(L-I)^\top K(L-I) \preceq K$, and for every $\Sigma' = v'I$ with $v' \ge v$ if and only if the trace of that difference is non-positive. For δ_t^*, $L = I - C$: the trace $\mathrm{tr}\{K(C^2 - I)\}$ is non-positive because $0 \prec C \preceq I$, and the matrix condition is $CKC \preceq K$, that is $\zeta_t \le 1$, which holds when K commutes with F, hence with C, since then $CKC = K^{1/2} C^2 K^{1/2} \preceq K$. At $\Sigma' = vI$ the average regret is $-\mathbb{E}_v\|\delta_t^*\|^2$ by (ii). If $\zeta_t > 1$, a direction u with $u^\top(CKC - K)u > 0$ and $\Sigma' = vI + s\,uu^\top$ make δ_t^*'s average regret grow linearly in s, and eventually positive. The size-robust update is $L_\gamma = I - \gamma C$ with $\gamma = \zeta_t^{-1}$, so $\gamma^2 CKC \preceq K$ by the definition of ζ_t. Its average regret at vI, $R(\gamma)$, is a convex quadratic in γ minimized at $\gamma = 1$, because θ_W is the posterior mean under $\mathcal{N}(0, vI)$ and so has the smallest average regret of all updates; hence $R(\gamma) \le R(0) = \mathrm{tr}(KN) - v\,\mathrm{tr}\,K$, which is at most zero when $N \preceq vI$, that is, when every read direction has $\lambda_j = v f_j \ge 1$. If K is singular on the read directions and no finite ζ_t exists, $\gamma = 0$ serves $(I - P_t)\hat\theta$, with the same guarantee. A true size below v can make each of these updates lose, which is why FLAG takes a lower bound for v. Over 1,972 random instances with a coupled G_t and U the eigendirections with $\lambda_j < 1$, $\zeta_t \le 1$ in 96%; where $\zeta_t > 1$, δ_t^* loses to not adapting only once the second moment along the worst direction exceeds v by a median factor of 312, and the size-robust update keeps a median 99.8% of δ_t^*'s predicted gain. Over 200 draws of $\Sigma' \succeq vI$ per instance, neither the size-robust update nor, where $\zeta_t \le 1$, δ_t^* ever exceeds its average regret at vI, and forty Monte Carlo runs of $40{,}000$ draws each match the closed form within three standard errors.

Proposition 10(ii). If the read part may take any value, $\theta = 0$ is a possible world, and there $\mathrm{Reg}_t(\delta; 0) = \mathbb{E}_0\|\delta\|^2$. It is positive unless $\delta = 0$ almost surely under $\theta = 0$, and then almost surely under every θ, because Gaussian label laws for different θ are mutually absolutely continuous; so every update but not adapting has positive regret in that world.

When (A0) holds only approximately. Let the early labels have means $Z\theta + r_D$ and the slow task's change on the served inputs be $z_t^\top \theta + r_t$, with residuals r_D and r_t outside the model, and let $\langle a, b\rangle$ and $\|\cdot\|$ be the mean product and norm over the served inputs.

Proposition 11 (Approximate (A0)). *For a linear update $\delta = L\tilde{y}$, such as δ_t^* with $L = (I - P_t)(v^{-1}I + F)^{-1}Z^\top\Sigma_\varepsilon^{-1}$, let $m = LZ\theta$ be its mean under the model and $b = Lr_D$ its shift. Its regret against not adapting is*

$$\mathrm{Reg}_t^{\mathrm{true}}(\delta) = \mathrm{Reg}_t(\delta;\theta) + \|b\|_{G_t}^2 + 2\langle m-\theta, b\rangle_{G_t} - 2\langle z_t^\top(m+b), r_t\rangle,$$

so $|\,\mathrm{Reg}_t^{\mathrm{true}} - \mathrm{Reg}_t\,| \le \|b\|_{G_t}^2 + 2\|m-\theta\|_{G_t}\|b\|_{G_t} + 2\|m+b\|_{G_t}\|r_t\|$. If $r_D = 0$ and r_t is independent of θ, the last term averages to zero over any law of θ with mean zero, and δ_t^* keeps its average regret $-\mathbb{E}\|\delta_t^*\|_{G_t}^2$.

Proof. The true regret is $\mathbb{E}\|z_t^\top\delta - z_t^\top\theta - r_t\|^2 - \|z_t^\top\theta + r_t\|^2 = \mathbb{E}\|\delta\|_{G_t}^2 - 2\langle\mathbb{E}\delta,\theta\rangle_{G_t} - 2\langle z_t^\top\mathbb{E}\delta, r_t\rangle$; with $\mathbb{E}\delta = m + b$ and $\mathbb{E}\|\delta\|_{G_t}^2 = \mathbb{E}_{\mathrm{model}}\|\delta\|_{G_t}^2 + 2\langle m,b\rangle_{G_t} + \|b\|_{G_t}^2$ this is the identity, Cauchy–Schwarz gives the bound, and $\mathbb{E}_\theta m = 0$ gives the average. Over 200 random instances with residuals on both sides, a Monte Carlo of 4,000 label draws matches the identity within three standard errors in all 200.

The population limit and the uncorrelated case. As the information grows on every read direction, $\theta_W \to \theta_R$ in mean square, so $\delta_t^* \to \theta_V$, whose gain $\|\theta\|^2 - \|\theta-\theta_V\|^2 = \|\theta\|^2 - \|\theta_B\|^2 = \|\theta_V\|^2$ is $\mathrm{dist}_{G_t}(\theta, U)^2$ by (c). If $B^\top G_t(I - BB^\top) = 0$, so that t's features on U are uncorrelated with its others, then $P_t = BG_{t,UU}^+ G_{t,UU}B^\top$, which equals the orthogonal projection $Q = BB^\top$ up to $\ker G_t$, and $\delta_t^* = (I-Q)\theta_W = \theta_W$ because θ_W lies in the range of F.

Proposition 12 (Trusting the past's size in the unread part). *In the population, among updates that serve the read part and fill the unread part linearly, and over laws with $\mathbb{E}\theta_R\theta_R^\top = vI$, $\mathbb{E}\theta_U\theta_U^\top \preceq v_U I$ and any cross moment, δ_t^* is minimax if and only if $\|G_{t,RU}G_{t,UU}^{-1}\| \le \sqrt{v_U/v}$; with $v_U = v$, if and only if $\|G_{t,RU}G_{t,UU}^{-1}\| \le 1$.*

Proof. In the population the read part is known. Consider the updates that serve it and fill the unread part linearly, $\delta = \theta_R + D\theta_R$ with D mapping the read directions into U, and write G_t in blocks $G_{t,RR}$, $G_{t,RU}$, $G_{t,UU}$ in orthonormal bases of the read directions and of U, with $G_{t,UU}$ invertible. Let the laws have $\mathbb{E}\theta_R\theta_R^\top = vI$, $\mathbb{E}\theta_U\theta_U^\top \preceq v_U I$ and cross moment $\Sigma_{UR} = \mathbb{E}\theta_U\theta_R^\top$. Then $\delta - \theta = D\theta_R - \theta_U$ lies in U, and with $Y = D^\top G_{t,UU} + G_{t,RU}$, the map from the read part to the update's leverage on U,

$$\mathrm{Reg}_t(\delta;\theta) = \theta_R^\top(D^\top G_{t,UU}D - G_{t,RR})\theta_R - 2\theta_R^\top Y\theta_U,$$
$$\mathbb{E}\ \mathrm{Reg}_t(\delta;\theta) = v\,\mathrm{tr}(D^\top G_{t,UU}D) - v\,\mathrm{tr}\,G_{t,RR} - 2\,\mathrm{tr}(Y\Sigma_{UR}).$$

A second-moment matrix with these blocks exists if and only if $\|\Sigma_{UR}\| \le \sqrt{vv_U}$ in operator norm: if $\mathbb{E}\theta_U\theta_U^\top = v_U I$ it is positive semidefinite exactly when $\Sigma_{UR}\Sigma_{UR}^\top \preceq vv_U I$, and positive semidefiniteness always gives $\Sigma_{UR}\Sigma_{UR}^\top/v \preceq \mathbb{E}\theta_U\theta_U^\top \preceq v_U I$. The worst Bayes regret is therefore

$$\bar{r}(D) = v\left\|(Y - G_{t,RU})G_{t,UU}^{-1/2}\right\|_F^2 - v\,\mathrm{tr}\,G_{t,RR} + 2\sqrt{vv_U}\,\|Y\|_*,$$

with $\|\cdot\|_*$ the nuclear norm, dual to the operator norm, and $D^\top = (Y - G_{t,RU})G_{t,UU}^{-1}$ substituted in the first term, which makes $\bar{r}$ strictly convex in Y. The projected update fills the unread part with $D = -G_{t,UU}^{-1}G_{t,UR}$, since $(I-P_t)\theta = \theta_R - G_{t,UU}^{-1}G_{t,UR}\theta_R$, and this is $Y = 0$. It minimizes $\bar{r}$ if and only if zero lies in the subdifferential at $Y = 0$: the first term has gradient $-2vG_{t,RU}G_{t,UU}^{-1}$ there, and the subdifferential of the nuclear norm at zero is the unit ball of the operator norm, so the condition is $\|G_{t,RU}G_{t,UU}^{-1}\| \le \sqrt{v_U/v}$, which at $v_U = v$ is the condition of (v). When it fails, the minimax fill keeps part of the leverage of leaving the unread part at the past value ($D = 0$, $Y = G_{t,RU}$): with one blind direction, $G_{t,UU} = g$ is a number, $\|Y\|_*$ is the Euclidean norm of the column Y, and the minimizer is the soft threshold $Y = \left(1 - g\sqrt{v_U/v}/\|G_{t,RU}\|\right)_+ G_{t,RU}$.

The theorem concerns the squared error of the mean, or of the logit for a binary task, which is the problem's criterion (Section 3.1); how the guarantee carries to log loss is discussed in Appendix E.1.

The size condition. Along an eigendirection b of F with eigenvalue f and $\lambda = vf$, θ_W's coordinate is $a\hat{\theta}_b$ with $a = \lambda/(1+\lambda)$ and $\hat{\theta}_b \sim \mathcal{N}(\theta_b, 1/f)$, independently across eigendirections, so

$$\mathbb{E}(a\hat{\theta}_b - \theta_b)^2 - \theta_b^2 = a^2/f + (a^2 - 2a)\,\theta_b^2,$$

which is negative if and only if $\theta_b^2 > a/\{f(2-a)\} = v/(2+\lambda)$. For a task with one read direction, or whose matrix $(I-P_t)^\top G_t(I-P_t)$ is diagonal in F's eigenbasis, δ_t^*'s regret is the sum of these terms weighted by that diagonal. Averaged over a true size v' of the read part, $\mathbb{E}\theta_R\theta_R^\top = v'I$, the regret of δ_t^* is $\sum_j a_{tj}\lambda_j\{v-(2+\lambda_j)v'\}/(1+\lambda_j)^2$ over the read eigendirections, with $a_{tj} \geq 0$ the diagonal of $(I-P_t)^\top G_t(I-P_t)$ in F's eigenbasis, for every task: the expected gain survives a wrong size as long as $v' > v/(2+\lambda_j)$ on the directions t depends on, and on the directions that FLAG counts as read, $\lambda_j \geq 1$, a true size above $v/3$ suffices.

C.2 THE SPLIT, THE BET AND WHAT NO EARLY LABEL CAN PRICE

Split any update likewise, $\delta = \delta_V + \delta_B$ with $\delta_V = (I-P_t)\delta$ and $\delta_B = P_t\delta$, and the change as $\theta_V = (I-P_t)\theta$ and $\theta_B = P_t\theta$, the unread part. Then

$$\mathrm{Reg}_t(\delta;\theta) = \mathrm{Reg}_t(\delta_t^*;\theta) + \underbrace{\mathbb{E}\|\delta_V-\theta_V\|_{G_t}^2 - \mathbb{E}\|\delta_t^*-\theta_V\|_{G_t}^2}_{\text{transport error}} + \underbrace{\mathbb{E}\|\delta_B\|_{G_t}^2 - 2\langle\mathbb{E}\delta_B,\theta_B\rangle_{G_t}}_{\text{bet}}, \quad (9)$$

and for every draw of the labels the bet is

$$\|\delta_B\|_{G_t}^2 - 2\langle\delta_B,\theta_B\rangle_{G_t} = \|\delta_B\|_{G_t}^2(1-2\rho), \qquad \rho = \langle\delta_B,\theta_B\rangle_{G_t}/\|\delta_B\|_{G_t}^2. \quad (10)$$

The transport error depends only on what the early labels read and is non-negative on average over Γ_v; it is the harm of Proposition 1. The bet pays exactly when $\rho > \frac{1}{2}$, and ρ is revealed only when the task's own labels mature.

Proof of equation 9 and equation 10. By fact (a), for every draw, $\|\delta-\theta\|^2 = \|\delta_V-\theta_V\|^2 + \|\delta_B-\theta_B\|^2$ and $\|\theta\|^2 = \|\theta_V\|^2 + \|\theta_B\|^2$, so

$$\mathrm{Reg}_t(\delta;\theta) = \mathbb{E}\|\delta_V-\theta_V\|^2 - \|\theta_V\|^2 + \mathbb{E}\|\delta_B\|^2 - 2\langle\mathbb{E}\delta_B,\theta_B\rangle.$$

For δ_t^* the unread part is zero, so $\mathrm{Reg}_t(\delta_t^*;\theta) = \mathbb{E}\|\delta_t^*-\theta_V\|^2 - \|\theta_V\|^2$, and subtracting gives equation 9, with $\theta_V = (I-P_t)\theta$ and $\theta_B = P_t\theta$. The bet of one draw is $\|\delta_B\|^2 - 2\langle\delta_B, P_t\theta\rangle = \|\delta_B\|^2(1-2\rho)$, which is equation 10. The transport error depends on θ only through the read part, since $\theta_V = (I-P_t)\theta_R$ and the law of δ_V does. For an update $\delta_t^* + \beta$ with β in U the transport error vanishes, and $\|\delta_t^*+\beta-\theta\|^2 - \|\delta_t^*-\theta\|^2 = \|\beta\|^2(1-2\rho)$ for every draw and every θ. For a covered task $G_tB = 0$, so $P_t = 0$ and no update has a bet; in the population, where $\delta_t^* = \theta$, the transport error is the whole error, as for fine-tuning with short and full watch (Appendix B).

What no early label can price. By equation 9, every update's regret is its projected part, a transport error and a bet. The first two depend on θ only through the read part. The bet's sign depends on the unread part, and the early labels have the same law whatever the unread part, so no statistic of them, validation on the early labels included, carries information about the sign of any bet.

Corollary 3 (No selection from the early labels)**.** *In the population, let two updates differ only in the unread part, $\delta_i = \delta_t^* + \beta_i$ with β_i in U a function of the read part, $i = 1, 2$, and $\|\beta_1 - \beta_2\|_{G_t} > 0$. No update computed from the early labels errs within any constant factor of the better of the two in every world.*

Proof. Given the read part, let world i have $\theta_U = \beta_i - P_t\theta_R$, which lies in U; then $P_t\theta = \beta_i$ by fact (b), and since $\delta_t^* = \theta_V$ in the population, δ_i errs by $\|\theta_V + \beta_i - \delta_t^* - \beta_i\|^2 = 0$ there. The two worlds share the read part and so give the early labels one law, and every update δ has $\|\beta_1-\beta_2\|^2 \leq 2\mathbb{E}\|\delta-\theta^{(1)}\|^2 + 2\mathbb{E}\|\delta-\theta^{(2)}\|^2$, so it errs by at least $\frac{1}{4}\|\beta_1-\beta_2\|^2 > 0$ in one world, where the better of the two errs by zero. □

Choosing between the two by validation on the early labels is one such rule.

Related criteria. The criterion of Theorem 1 is a Γ-minimax criterion (Berger, 1985), with regret against the status quo in place of risk as in minimax-regret treatment choice (Manski, 2004; Tetenov, 2012), over the multiple-prior class that fixes the law of what the data identify and leaves the rest free (Giacomini & Kitagawa, 2021); not losing to a baseline policy is also what safe policy improvement asks (Thomas et al., 2015; Laroche et al., 2019).

The examples of the main text. *Two worlds* (Section 3.4; Figure 2c). With click and short-watch labels, in the population, purchase's unread part is intent, and G_t is diagonal, so $\delta_t^* = \theta_W$. The full prior of Appendix F.1 fills intent with its completion of the read part, $+.328\,r$ in the gains of Section 3.1; the actual campaign has $-.441\,r$. In the world that follows the co-movement, not adapting errs by $.294\,r^2$, θ_W by $.069\,r^2$ and θ_A by zero, regrets $-.226\,r^2$ and $-.294\,r^2$; in the actual one, not adapting errs by $.350\,r^2$, θ_W by $.124\,r^2$ and θ_A by $.378\,r^2$, regrets $-.226\,r^2$ and $+.028\,r^2$. θ_A's bet has stake $.069\,r^2$ and $\rho = -1.35$ in the actual world, and by equation 10 it costs $.254\,r^2$ there, exactly θ_A's excess over θ_W.

Why the projection is in the task's geometry. In two dimensions let the first coordinate be read and the second not, $\theta = (1, 1)$ and $G_t = \left(\begin{smallmatrix} 1 & -.9 \\ -.9 & 1 \end{smallmatrix}\right)$. Not adapting errs by .2. The population limit of θ_W, exact on the read coordinate and at the past value on the other, errs by 1: it undoes a cancellation that the past value enjoys, so being right on every read direction does not make an update better than not adapting. The projected update $(I - P_t)\theta = (1, .9)$ errs by .01, and the condition of Proposition 12 holds.

The projection is not free. Let the running example's factors be non-negative, with $\mathbb{E}\phi\phi^\top = .51\,I + .49\,\mathbf{1}\mathbf{1}^\top$ (mean .7 and unit second moment), so that purchase's features couple intent to appeal and quality. With click and short-watch labels, in the population, δ_t^* fills intent from the read part, and on the actual campaign it errs by .465, against .124 for leaving intent at the past value and .693 for not adapting. Leaving intent at the past value is then itself a bet, with stake .108 and $\rho = 2.07$, and this campaign pays it. Over all changes of unit norm, however, leaving intent at the past value loses to not adapting by up to .161, while δ_t^* never loses, its worst regret being zero, reached by changes in U; the condition of Proposition 12 holds here too.

Numerical checks. Two hundred random instances have six coordinates, a two-dimensional $U = \ker Z$, information from $\lambda = .1$ to 100 on the read directions with $v = 1$, a strongly coupled G_t, with median $\|B^\top G_t R\|_F / (\|B^\top G_t B\|_F \|R^\top G_t R\|_F)^{1/2}$ of .73 for an orthonormal basis R of the read directions, and an anisotropic Σ. In closed form, δ_t^*'s regret varies along U by rounding error of order 10^{-14} times $\|\theta\|_{G_t}^2$ for unread parts from a tenth to a hundred times the read part, while the regrets of θ_W and θ_A change along the worst direction of U at median rates of 2.86 and 3.35 per unit of the unread part's norm; with an adversarial unread part they lose to not adapting in 85% and 88% of the worlds, and with a random one in 33% and 39%, against 15% for δ_t^*, whose losses are label noise. δ_t^*'s Bayes regret equals $-\mathbb{E}\|\delta_t^*\|_{G_t}^2$ up to rounding of the same order; 1,000 perturbations that keep the condition of (i) all raise it, each by exactly its own expected squared norm, and every perturbation that leaves the condition has infinite worst regret. The population gain equals $\mathrm{dist}_{G_t}(\theta, U)^2$ to 10^{-14}, and scaling the information by a hundred brings the finite-label gain within 1.5% of it in the median instance. A Monte Carlo run of two hundred thousand label draws matches the closed-form regrets of θ_W, θ_A and δ_t^* within 1.4 standard errors, and over 500 instances equation 10 holds up to rounding of order 10^{-14} for both candidate bets, $P_t\theta_A$ and $P_t\theta_W$, with ρ from -43 to 87. Over 2,000 further instances with six coordinates, a two-dimensional U and a random G_t, the identity of Lemma 2 holds to 3×10^{-14} relative error; along the worst unit direction of U the regret of θ_W has slope a median .91 times $\|\theta_W\|_{G_t}$ (from .03 to 2.8) and that of δ_t^* at most 3×10^{-15}; the Schur form of the gain holds to 3×10^{-15}; δ_t^* moves the unread coordinates by a median $.48\|\theta_W\|$; and with $G_{t,RU} = 0$, $\delta_t^* = \theta_W$ to 4×10^{-14}.

C.3 The information view: how much of a task's risk is left to the past

How much of the isotropic update's risk comes from what the early labels leave to the past, and how much from their noise? This subsection restates, as supporting material, the split of θ_W's risk into these two parts, which explains the threshold $\lambda_j < 1$ that defines U, bounds what the full prior's shape can change, and gives the limit in which the projected update's blind spot becomes $\ker Z$. FLAG does not compare the two parts with each other: it serves the projected update, and the split only motivates the threshold. With $C = (I + vF)^{-1}$, let

$$\pi_t = v\,\mathrm{tr}(G_t C^2) \ \text{(left to the past)}, \qquad \nu_t = v\,\mathrm{tr}\{G_t(C - C^2)\} \ \text{(label noise)}. \tag{11}$$

Proposition 13 (The isotropic update's risk, and what the prior's shape can change)**.** *Let $\tilde{y} = Z\theta + \varepsilon$ with noise covariance Σ_ε and $F = Z^\top \Sigma_\varepsilon^{-1} Z$. (i) If $\mathbb{E}\theta = 0$ and $\mathbb{E}\theta\theta^\top = vI$, θ_W is the best affine estimate of θ, with risk $\pi_t + \nu_t$ on task t. (ii) If $\Sigma \succeq \frac{v}{2} I$, then for every draw of the labels, whatever θ, $\|\theta_A - \theta_W\|_{G_t}^2 \le (\pi_t / v)\,\|\theta_A\|^2$.*

Proposition 14 (Closed form, and coverage as the limit)**.** *Under Proposition 13(i), let F have eigenpairs (f_j, b_j), $\lambda_j = v f_j$ and $g_{tj} = b_j^\top G_t b_j$. (i) $\pi_t/v = \operatorname{tr}(QG_t) + \sum_{j:\, f_j>0} g_{tj}/(1+\lambda_j)^2$ and $\nu_t/v = \sum_{j:\, f_j>0} g_{tj}\lambda_j/(1+\lambda_j)^2$, so along each eigendirection the part left to the past exceeds the label noise exactly when $\lambda_j < 1$. (ii) As the information grows on every read direction, $\pi_t \to v \operatorname{tr}(QG_t)$ and $\nu_t \to 0$, and the directions with $\lambda_j < 1$ become $\ker Z$: the projected update's blind spot becomes $\ker Z$, which carries none of t's weight exactly when $\ker Z \subseteq \ker G_t$, that is, when the arrived design covers t. (iii) With $F = nF_1$ for n labels per fast task, θ_W's risk $R_{W,t} = \pi_t + \nu_t = v \operatorname{tr}(G_t C)$ has $\nu_t = -\partial R_{W,t}/\partial \log n$ and $\pi_t = \partial R_{W,t}/\partial \log v$.*

Proof. In the affine experiment $\hat\theta = F^+ Z^\top \Sigma_\varepsilon^{-1} \tilde y$, and since $\operatorname{range}(Z^\top \Sigma_\varepsilon^{-1}) = \operatorname{range}(F)$, $\theta_W = (v^{-1} I + F)^{-1} F \hat\theta = v C Z^\top \Sigma_\varepsilon^{-1} \tilde y$.

Proposition 13(i). The best affine estimate of θ is its mean plus the cross covariance with $\tilde y$ times the inverse covariance of $\tilde y$, $vZ^\top (vZZ^\top + \Sigma_\varepsilon)^{-1} \tilde y$, which the push-through identity turns into θ_W. Its error is uncorrelated with $\tilde y$, which gives covariance minimality among affine estimators and so the smallest G_t-weighted risk for every task. Since $vCF = I - C$, the error is $\theta_W - \theta = -C\theta + vCZ^\top \Sigma_\varepsilon^{-1} \varepsilon$, with covariance terms vC^2 and $v^2 CFC = v(C - C^2)$ and no cross term, so the risk on task t is $\pi_t + \nu_t$. If the change's second moment is $v' I$ instead, θ_W's risk is $(v'/v)\pi_t + \nu_t$.

Proposition 13(ii). With $K_X = (X^{-1} + F)^{-1}$, $\theta_A = K_\Sigma F \hat\theta$ and $\theta_W = K_{vI} F \hat\theta$ with $K_{vI} = vC$, and $A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}$ for $A = \Sigma^{-1} + F$ and $B = v^{-1} I + F$ gives $\theta_A - \theta_W = C(I - v\Sigma^{-1})\theta_A$. If $\Sigma \succeq \frac{v}{2} I$, the eigenvalues of $v\Sigma^{-1}$ lie in $(0, 2]$, so $\|I - v\Sigma^{-1}\| \le 1$, and $\|G_t^{1/2} C\|^2 \le \operatorname{tr}(CG_tC) = \pi_t/v$. The prior of Appendix F.1, $\Sigma = \tau[\frac{1}{2}\widehat\Sigma_0 + (\frac{1}{2} v_0 + \epsilon) I]$ with $\epsilon = 10^{-3} \max(1, v_0)$, has $v = \tau(v_0 + \epsilon)$ and $\lambda_{\min}(\Sigma) \ge \tau(\frac{1}{2} v_0 + \epsilon) \ge v/2$. Without the identity part, $v\Sigma^{-1}$ is unbounded on the directions the past windows did not move, and no such bound holds.

Proposition 14. In F's eigenbasis C has eigenvalues $(1+\lambda_j)^{-1}$, so C^2 and $C - C^2$ have $(1+\lambda_j)^{-2}$ and $\lambda_j (1 + \lambda_j)^{-2}$, with $\lambda_j = 0$ on $\ker Z$, where $C^2 = I$; summing against g_{tj} gives (i), and the two coefficients compare as 1 and λ_j. For (ii), every coefficient with $f_j > 0$ tends to zero, and once the smallest λ_j on the range of F reaches one, the directions below the threshold are those of $\ker Z$; $\operatorname{tr}(QG_t) = 0$ exactly when $G_t Q = 0$. For (iii), $dC/dn = -vCF_1C$, so $n\, dR_{W,t}/dn = -v \operatorname{tr}\{G_t C(vF)C\} = -v \operatorname{tr}\{G_t(C - C^2)\} = -\nu_t$ because $vFC = I - C$; and $d(vC)/dv = d(v^{-1} I + F)^{-1}/dv = C^2$, so $v\, dR_{W,t}/dv = \pi_t$. □

Readings. By Proposition 14(iii), ν_t is the elasticity of θ_W's risk to the number of early labels and π_t its elasticity to the prior's scale. Read with equation 9, Proposition 13(ii) bounds by $(\pi_t/v)\|\theta_A\|^2$ the stake of the bet $P_t(\theta_A - \theta_W)$ that separates the full-prior update from θ_W, and the transport part of their difference as well, since both are parts of $\theta_A - \theta_W$ in t's geometry. For a covered task $G_t Q = 0$, so the two updates differ only in how they shrink the read directions. Each is the posterior mean, hence the Bayes estimate, under its own prior, $\mathcal{N}(0, vI)$ or $\mathcal{N}(0, \Sigma)$, so ranking them requires choosing one of the two; the waiting benchmark trusts only the size v, which they share.

From the affine model to the fitted procedure. If the noise $\hat\theta - P\theta$ has mean zero, is independent of θ and has covariance at most F^+ on $(\ker Z)^\perp$, then θ_W has risk at most $\pi_t + \nu_t$, because its noise covariance is at most $(I - C)F^+(I - C) = v(C - C^2)$; for a generalized least-squares estimate with score covariance Ω, $\operatorname{Cov}(\hat\theta - P\theta) = H^+ \Omega H^+$, and the construction of F gives $F^+ \succeq H^+ \widehat\Omega H^+$ and $F \preceq H$. Remainders of the fast or slow means beyond the model add the bias bounded in Appendix E.3.

The running example. In the coordinates of Appendix F.1, each factor's gain divided by its root-mean-square over the two weeks, with the prior of the two weeks at $\tau = 1$, so that $v = .301$, $\sigma = .1$, known factors and 30 labels per fast task with the expected information: with click and short watch, $\pi_t/v = 1.117$, nearly all of it on the unread intent, and $\nu_t/v = .0012$; with the weak reader of Section 4 added, intent carries $\lambda = .63$, $\pi_t/v = .420$ and $\nu_t/v = .266$, and intent leaves U at about 48 labels per fast task, where λ passes one; with full watch instead, $\pi_t/v = 9.0 \times 10^{-7}$ and $\nu_t/v = .0016$.

D WHEN DO PAST WINDOWS FIX THE FACTORS, AND AT WHAT COST?

This appendix proves the results of Section 3.2: when past windows identify the factors (Lemma 3), what estimating them costs a slow task (Theorem 2), and why diagonal gains on a pooled basis cannot represent a new window (Proposition 15); it also bounds how stably two windows fix the factors. Throughout, C, C_0, C_1, C_2 and c_0, c_1 denote positive constants that may change from line to line. They do not depend on n, n_s, σ, σ_s, p_{fail}, h or ϑ, but may depend on k and on the scale bounds of each statement; C_{cond} and c_{cond} depend also on the condition numbers listed in Theorem 2.

D.1 THE TWO RESULTS

Section 3.2 states the next lemma in words; this subsection states it, gives its estimator, and relates the cost of estimated factors, Theorem 2 (Appendix D.4), to the early labels.

Lemma 3 (Two windows identify the factors)**.** *Suppose the base is known, $W = [w_1, \dots, w_T]^\top$ has full column rank, and the windows share an input law with $\mathbb{E}[\phi\phi^\top] \succ 0$. Two past windows with nonzero gains identify the readouts and factors, up to order and scale, if and only if their gain ratios s_{1j}/s_{2j}, $j = 1, \dots, k$, are distinct.*

The estimator of Lemma 3. Let Y_e be the $T \times k$ matrix of window e's residuals $\delta_{t,e}$ at k inputs whose factor values are linearly independent, the same inputs in both windows. Multiplying Y_1 by Y_2^+ cancels the factors:

$$Y_1 Y_2^+ = W \operatorname{diag}(s_1 \oslash s_2)\, W^+, \tag{12}$$

with $\oslash$ entrywise division, so the eigenvectors of $Y_1 Y_2^+$ for its nonzero eigenvalues are the columns of W (each factor's weights across tasks), and the eigenvalues are the gain ratios. Appendix D.4 applies the same pencil to least-squares fits of the two windows.

The bound. Write $\varrho = s_1 \oslash s_2$ for the past ratios and split the current ratios as $\varrho^\star = s^\star \oslash s_2 = a_1 \mathbf{1} + a_2 \varrho + \omega$, where a_1, a_2 give the part inside the span of $\mathbf{1}$ and ϱ, which is the span of the past gains divided entrywise by s_2, and ω is the rest; the sizes in Theorem 2 are $c_{\text{in}} = |a_1| + |a_2|$ and $c_{\text{out}} = \|\omega\|_\infty$. With $\bar{h} = \min(h, 1)$ and $\kappa > 0$ the smallest eigenvalue of $G_D/|D|$, Theorem 2 gives

$$R_{t^\circ} \;\lesssim\; \frac{\sigma_s^2}{n_s}\, c_{\text{in}}^2 + \frac{\sigma_s^2}{n_s \bar{h}^2}\, c_{\text{out}}^2 + \frac{\sigma^2 k}{n\,|D|\,\kappa}, \tag{13}$$

the three terms of Theorem 2 in order: past-window noise, the error from the ratio gap and early-label noise. When the ratios crowd together (h small), the estimated factor axes are less precise. The part of a new window inside the span of the past gains adds no error from this imprecision, although large coefficients there amplify the past windows' noise; the part outside the span is multiplied by it. In the running example ($h = 5/9$) the campaign incurs the middle term through its off-plane component.

What $\lesssim$ absorbs. Relative to Theorem 2, equation 13 absorbs four quantities of equation 14: (i) the factor d_s/κ_s of the first two terms, where $d_s = T + p + \log(C/p_{\text{fail}})$ collects the dimensions and the failure probability and κ_s is the lower bound $n_s \kappa_s I_p$ on each past design Gram matrix; (ii) the factor $(1 + \xi_{t^\circ}^2/\kappa)$ of the first two terms and the factor $4\xi_{t^\circ}^2$ of the last, where ξ_{t° bounds the estimated slow feature map $\hat{M}_{t^\circ}$ and is at most $\|M_{t^\circ}\|_{\text{op}} + \sqrt{\kappa}/4$ on the event of the theorem; (iii) the constant C_{cond}, which depends only on k and on the bounds on $\operatorname{cond}(W)$, $\sigma_{\min}(W)$, β, b_1, $\lambda_{\min}(K_\phi)$, $\lambda_{\max}(K_\phi)$, $\|M\|_{\text{op}}$ and the past gains listed in Theorem 2; and (iv) the ratio of $c_a = 2|a_1| + C(1 + b_1/\beta)|a_2|$ to c_{in}, at most $\max\{2, C(1 + b_1/\beta)\}$, which the same bounds control. Theorem 2 writes the last term as σ^2/n and so absorbs also its factor $k/(|D|\kappa)$. Its "with high probability" means with probability at least $1 - p_{\text{fail}}$ once n_s and n reach the sample sizes of Theorem 2, and "the fast tasks read every factor" is the full fast coverage $G_D \succeq |D|\kappa I_k$ with $\kappa > 0$.

D.2 LEMMA 3 IN GENERAL FORM

This subsection proves Lemma 3 for any number of past windows and says what remains identified when its condition fails. Throughout, the base is known and subtracted, so the residuals $\delta_{t,e}(x) =$

$w_t^\top \operatorname{diag}(s_e)\phi(x)$ form a task-by-window-by-input array of canonical polyadic form $\sum_j W_{:j} \otimes S_{:j} \otimes \phi_j$ (Harshman, 1970; Leurgans et al., 1993). Two conditions are used: **(A1)** W has full column rank; **(A2)** a common input law $\mathbb{P}_0$, dominated by every window's input law, has $\mathbb{E}_{\mathbb{P}_0}[\phi\phi^\top] \succ 0$. When recovered factors serve a new window, its input law must also be dominated by $\mathbb{P}_0$.

Two decompositions are *essentially equivalent* if a common permutation of their k terms, and nonzero rescalings of the three factors of each term with product one, make them equal $\mathbb{P}_0$-almost everywhere; a rank-k decomposition is any representation with at most k rank-one terms.

Lemma 4 (General form of Lemma 3)**.** *Assume equation 1 with a known base, (A1), (A2), $k \geq 2$ factors and $m \geq 2$ past windows. (i)* Pooling. *For every invertible Θ, the pooled mean $\mathbb{E}[y \mid x] = W\bar{\psi}(x)$, with the pooled scores $\bar{\psi} = \mathbb{E}[s_e \mid x] \odot \phi$, is reproduced by $(W\Theta^{-1}, \Theta\bar{\psi})$, and one task's profile across windows, $S\operatorname{diag}(w_t)\phi(x)$, by $(S\operatorname{diag}(w_t)\Theta^{-1}, \Theta\phi)$. (ii)* Uniqueness. *The array has rank k and an essentially unique rank-k decomposition if and only if every gain profile is nonzero and no two are proportional, that is, $\operatorname{krank}(S) \geq 2$. For two windows whose gains are all nonzero, this holds if and only if the ratios s_{1j}/s_{2j} are distinct. (iii)* Proportional profiles. *For each maximal class of proportional nonzero profiles, the class profile and its task and input subspaces are identified; within a class of more than one profile the class's factors admit any invertible change of basis, and a factor with a zero profile is not observed.*

Proof. Outline. Pooling is a change of variables. For uniqueness, mix the window slices with two generic weight vectors: the pencil of the two mixtures has the task-factor lines as eigenvectors, with distinct eigenvalues exactly when no two gain profiles are proportional.

Step 1: pooling, part (i). $\mathbb{E}[y \mid x] = \sum_e p(e \mid x) W \operatorname{diag}(s_e)\phi(x) = W\bar{\psi}(x)$ and $(W\Theta^{-1})(\Theta\bar{\psi}) = W\bar{\psi}$; the one-task statement is the same identity. If e is independent of x, $\bar{\psi} = \bar{s} \odot \phi$.

Step 2: the array has rank k. Since $\mathbb{E}_{\mathbb{P}_0}[\phi\phi^\top] \succ 0$, the essential range of ϕ spans $\mathbb{R}^k$; choose $x_1, \ldots, x_k$ in a common full-measure set with $\Phi = [\phi(x_1), \ldots, \phi(x_k)]$ invertible. For window weights $\alpha \in \mathbb{R}^m$ the residual slices then satisfy

$$Y_e = W \operatorname{diag}(s_e)\Phi, \qquad Y_\alpha = \textstyle\sum_e \alpha_e Y_e = W \operatorname{diag}(S^\top\alpha)\Phi.$$

Suppose every $S_{:j} \neq 0$. Outside finitely many proper hyperplanes, $\alpha^\top S_{:j} \neq 0$ for all j, so Y_α has rank k; hence the array has rank at least k, the displayed representation gives at most k, and in any competing k-term decomposition $Y_e = \tilde{W}\tilde{D}_e\tilde{\Phi}$ the matrices $\tilde{W}, \tilde{\Phi}, \tilde{D}_\alpha$ have full rank.

Step 3: the pencil fixes the task-factor lines. Suppose also that the profiles are pairwise nonproportional. For fixed generic α, choose α' outside the finitely many hyperplanes $(\alpha'^\top S_{:j})(\alpha^\top S_{:l}) = (\alpha'^\top S_{:l})(\alpha^\top S_{:j})$, $j \neq l$, which are proper because $S_{:j}$ and $S_{:l}$ are not proportional. The ratios $\varrho_j = (\alpha'^\top S_{:j})/(\alpha^\top S_{:j})$ are then distinct, and the reverse-order rule for Moore–Penrose inverses of full-column and full-row factors gives

$$Y_{\alpha'} Y_\alpha^+ = W \operatorname{diag}(S^\top\alpha')\Phi \left(W \operatorname{diag}(S^\top\alpha)\Phi\right)^+ = W \operatorname{diag}(\varrho)\, W^+.$$

On $\operatorname{col}(Y_\alpha) = \operatorname{col}(W)$ this operator has the distinct eigenvalues ϱ_j with one-dimensional eigenspaces $\operatorname{span}(W_{:j})$, so every competing decomposition has the same task-factor lines up to a common permutation.

Step 4: input factors and gains. The same argument for $Y_{\alpha'}^\top (Y_\alpha^\top)^+$ gives the input-factor lines on $x_1, \ldots, x_k$; once W is identified, left-inverting each Y_e identifies $\phi_j(x)$ at every x of the common support, and each rank-one term then identifies $S_{:j}$ up to the two compensating scales. For two windows, $\alpha' = \mathbf{e}_1$ and $\alpha = \mathbf{e}_2$ give the ratios s_{1j}/s_{2j} and equation 12.

Step 5: necessity, and part (iii). If $S_{:j} = 0$ the term never appears, and proportional profiles break uniqueness by Remark 5. For (iii), a generic α' makes the ratios equal exactly within proportional classes, so the eigenspaces of $Y_{\alpha'} Y_\alpha^+$ are the classes' task subspaces and, after transposition, their input subspaces; restricting to a class recovers its profile up to scale, and within a class the construction of Remark 5 is the remaining ambiguity. □

Remark 5 (Proportional profiles). Factors with nonzero proportional gain profiles are fixed only up to a change of basis among them.

If $S_{:l}$ is a nonzero multiple of $S_{:j}$, so that the ratio s_{el}/s_{ej} is the same in every window e, the two terms contribute $s_{ej}\,[W_{:j}\ \ (s_{el}/s_{ej})W_{:l}]\,[\phi_j\ \ \phi_l]^\top$; for any invertible non-monomial 2×2 matrix Θ, the task matrix $[W_{:j}\ \ (s_{el}/s_{ej})W_{:l}]\Theta$ and input matrix $[\phi_j\ \ \phi_l]\Theta^{-\top}$, both with profile $S_{:j}$, leave every slice unchanged and generally change the factor lines; the same construction applies to every proportional class of size above one.

Proof of Lemma 3. Its assumptions are (A1) and (A2) with a common input law. For $k \geq 2$, part (ii) of Lemma 4 with $m = 2$ gives the two-window statement and part (i) the pooled one; for $k = 1$ a rank-one array fixes its factors up to scale. □

Each premise of Lemma 3 is needed. *Common input law:* with $k = T = 1$, $m = 2$ and inputs $\{x', x''\}$, if window 1 observes only x' and window 2 only x'', then $(w, s, \phi) = (1, (2, 3), (1, 1))$ and $(1, (2, 6), (1, \frac{1}{2}))$ give the same observed means. *Known base:* the complete slices $\mathrm{diag}(1, 2, 4)$ and $\mathrm{diag}(2, 3, 5)$ are reproduced, for any invertible Θ, by the readouts Θ, the factor map Θ^{-1}, the same gains and the base $\mathrm{diag}(1, 2, 4) - \Theta\,\mathrm{diag}(1, 2, 4)\Theta^{-1}$, although the ratios are distinct. *Uncentred coordinates:* exactly two centred gains have $s_2 = -s_1$, so every ratio is -1, and two centred windows separate no pair of factors when $k > 1$, at any sample size. Kruskal's condition $\mathrm{krank}(W) + \mathrm{krank}(S) + \mathrm{krank}(\Phi) \geq 2k + 2$ (Kruskal, 1977) reduces here to $\mathrm{krank}(S) \geq 2$, the pencil argument is the simultaneous diagonalization of Leurgans et al. (1993), and for binary labels the lemma applies on the logit scale.

D.3 Stability

How precisely do two windows fix the factors? The more their gain profiles differ, the more precisely. Let $\vartheta = \sin\angle(S_{:1}, S_{:2})$ be the sine of the angle between two gain profiles.

Lemma 5 (Stability)**.** *In each of two windows observe n_s pairs with $x \sim \mathcal{N}(0, I_2)$, $y = W\,\mathrm{diag}(s_e)Mx + \varepsilon \in \mathbb{R}^T$ and $\varepsilon \sim \mathcal{N}(0, \sigma_s^2 I_T)$, where $T \geq 2$, and the singular values of W and M, the gains and the profile norms are bounded above and away from zero. The pencil estimator of the columns of W has $\max_j \sin\angle(\hat{w}_j, W_{:j}) \leq C\sigma_s\sqrt{T + \log(1/p_{\mathrm{fail}})}/(\vartheta\sqrt{n_s})$ with probability at least $1 - p_{\mathrm{fail}}$ when the right side is below a constant, and for $T = 2$ every estimator has worst-case expected error at least $c_1 \min\{1, \sigma_s/(\vartheta\sqrt{n_s})\}$.*

The eigengap of equation 12 is exactly

$$|\varrho_1 - \varrho_2| = \frac{|\det[S_{:1}, S_{:2}]|}{|s_{21}s_{22}|} = \frac{\|S_{:1}\|\,\|S_{:2}\|}{|s_{21}s_{22}|}\,\vartheta, \qquad \varrho_j = s_{1j}/s_{2j},$$

the gap that controls perturbation analyses of the pencil (Moitra, 2018; Bhaskara et al., 2014). The proofs of this subsection and of Appendix D.4 use the following standard facts, with $\|\cdot\|$ the operator norm and $\sigma_1 \geq \sigma_2 \geq \cdots$ the singular values.

Lemma A (Standard perturbation facts)**.** *Let Y and E be real matrices of the same size.*

- *(a)* Weyl *(Horn & Johnson, 1991): $|\sigma_j(Y+E) - \sigma_j(Y)| \leq \|E\|$ for every j. Hence, if $\mathrm{rank}\,Y \leq k$, the best rank-k approximation $\tilde{Y}$ of $Y + E$ has $\|\tilde{Y} - Y\| \leq 2\|E\|$; and if Y has full column rank and $\|E\| \leq \sigma_{\min}(Y)/2$, then $\|(Y + E)^+\| \leq 2/\sigma_{\min}(Y)$.*
- *(b)* Inverses*: if Y and $Y + E$ are invertible, $(Y + E)^{-1} - Y^{-1} = -(Y + E)^{-1}EY^{-1}$.*
- *(c)* Wedin, singular subspaces*: if $\mathrm{rank}\,Y = k$ and $\|E\| \leq \sigma_k(Y)/2$, the orthogonal projections onto the column (row) space of Y and onto the span of the leading k left (right) singular vectors of $Y + E$ differ by at most $2\|E\|/\sigma_k(Y)$, and orthonormal bases of the two column spaces can be matched by a $k \times k$ orthogonal matrix to within $C\|E\|/\sigma_k(Y)$.*
- *(d)* Wedin, pseudoinverses*: if $\mathrm{rank}(Y + E) = \mathrm{rank}\,Y$, then $\|(Y + E)^+ - Y^+\| \leq C\|Y^+\|\,\|(Y + E)^+\|\,\|E\|$ with an absolute constant C.*
- *(e)* Bauer–Fike, and eigenvectors *(Moitra, 2018): let $\Psi = V\Lambda V^{-1}$ be $k \times k$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ real, $h_\Lambda = \min_{i \neq j} |\lambda_i - \lambda_j| > 0$, and let E be $k \times k$ with $\mathrm{cond}(V)\|E\| \leq h_\Lambda/4$. Each eigenvalue of $\Psi + E$ lies within $\mathrm{cond}(V)\|E\|$ of some λ_j, each of these k disjoint discs holds exactly one, and a unit eigenvector $\hat{v}_j$ for the eigenvalue near λ_j has $\sin\angle(\hat{v}_j, V_{:j}) \leq 2\,\mathrm{cond}(V)\|E\|/h_\Lambda$. If $\Psi + E$ is real, these eigenvalues are real, since the non-real eigenvalues of a real matrix come in conjugate pairs.*

(f) Gordon, with Gaussian concentration*: if X is $n \times p$ with independent $\mathcal{N}(0,1)$ entries and $t \geq 0$, then $\sigma_1(X) \leq \sqrt{n}+\sqrt{p}+t$ and, for $n \geq p$, $\sigma_p(X) \geq \sqrt{n}-\sqrt{p}-t$, each failing with probability at most $e^{-t^2/2}$.*

Proof. Upper bound. Let $\hat{Y}_e$ be the least-squares coefficient matrices, which estimate the slices $Y_e = W\,\mathrm{diag}(s_e)M$. Lemma A(f), applied to the design and then, given the design, to the noise, gives $\|\hat{Y}_e - Y_e\|_{\mathrm{op}} \leq C_0\sigma_s\sqrt{(T+\log(1/p_{\mathrm{fail}}))/n_s}$ for both windows once n_s exceeds a constant times $1+\log(1/p_{\mathrm{fail}})$. Let $\hat{U}$ be an orthonormal basis of the column space of $\hat{Y}_2$; as both singular values of Y_2 are bounded away from zero, Lemma A(c) puts this space within the same order of the true one, and after matching bases the compressed slices $\hat{\Psi}_e = \hat{U}^\top \hat{Y}_e$ are 2×2 perturbations of $\Psi_e = U^\top Y_e$. Lemma A(b) gives $\|\hat{\Psi}_1\hat{\Psi}_2^{-1} - \Psi_1\Psi_2^{-1}\|_{\mathrm{op}} \leq C_1\sigma_s\sqrt{(T+\log(1/p_{\mathrm{fail}}))/n_s}$. The population matrix is $\Psi_1\Psi_2^{-1} = (U^\top W)\,\mathrm{diag}(\varrho_1,\varrho_2)(U^\top W)^{-1}$, whose eigengap is the display above, comparable to ϑ under the scale bounds. Lemma A(e), with $k=2$, $V = U^\top W$ and $\Lambda = \mathrm{diag}(\varrho_1,\varrho_2)$, bounds each eigenvector's angular error by $C_2\|\hat{\Psi}_1\hat{\Psi}_2^{-1} - \Psi_1\Psi_2^{-1}\|_{\mathrm{op}}/|\varrho_1-\varrho_2|$ when the perturbation is below a fixed fraction of the gap; lifting by $\hat{U}$ adds only the column-space error, of no larger order.

Lower bound. Set $T=2$ and $W_0 = M_0 = I$, and let $W_r = I + r\mathbf{e}_1\mathbf{e}_2^\top$ and $M_r = I - r\frac{S_{:1}^\top S_{:2}}{\|S_{:1}\|^2}\mathbf{e}_1\mathbf{e}_2^\top$. Since $(\mathbf{e}_1\mathbf{e}_2^\top)^2 = 0$, slice e changes by exactly $r\big(s_{e2} - s_{e1}\frac{S_{:1}^\top S_{:2}}{\|S_{:1}\|^2}\big)\mathbf{e}_1\mathbf{e}_2^\top$, so for Gaussian inputs and noise the total Kullback–Leibler divergence is $n_s r^2\|S_{:2}\|^2\vartheta^2/(2\sigma_s^2)$, while the second column of W_r turns by $\arctan|r|$. Taking $r = c_0\min\{1, \sigma_s/(\|S_{:2}\|\vartheta\sqrt{n_s})\}$ keeps the two laws within a fixed distance while their directions stay separated, and Le Cam's two-point bound (Tsybakov, 2009) gives the claim. □

The angle ϑ is scale-invariant (multiplying every gain by 10^{-6} leaves it unchanged but destroys the signal), so a statement for general k must also control gain magnitudes and condition numbers, as Theorem 2 does.

D.4 Theorem 2: formal statement and proof

This subsection states Theorem 2 with its constants as Theorem 2 and proves both bounds; Table 5 lists the constants.

Setting and estimator. Three symbols take a local meaning here. Y_e is window e's $T \times p$ matrix of residual coefficients, the slice as a linear map of x rather than its values at k inputs as in Appendix D.1; the columns of W are rescaled to unit length, with their scales moved into M; and M also absorbs the nonzero gains of past window 2, so that the gain ratios take the place of the gains. The known base is subtracted from every label. Let $\phi(x) = Mx$ with $x \in \mathbb{R}^p$, $M \in \mathbb{R}^{k\times p}$ of full row rank, $k \geq 2$, and $W \in \mathbb{R}^{T\times k}$ of full column rank. With these conventions the residual slices are $Y_2 = WM$, $Y_1 = W\,\mathrm{diag}(\varrho)M$ and $Y_\star = W\,\mathrm{diag}(\varrho^\star)M$, with ratios $\varrho = s_1 \oslash s_2$ and current ratios $\varrho^\star = s^\star \oslash s_2$. Let $h = \min_{i\neq j}|\varrho_i - \varrho_j| > 0$, $\bar{h} = \min(h,1)$, $\beta = \sigma_k(Y_2) > 0$, $b_1 = \|Y_1\|_{\mathrm{op}}$ and $K_\phi = MM^\top$. Each past window supplies n_s units with design Gram matrix at least $n_s\kappa_s I_p$ and independent Gaussian label noise of variance σ_s^2. The estimator fits both residual slices by least squares, truncates $\hat{Y}_2$ to rank k, $\tilde{Y}_2 = \hat{U}\hat{\Lambda}\hat{V}^\top$, diagonalizes $\hat{\Psi} = \hat{U}^\top\hat{Y}_1\hat{V}\hat{\Lambda}^{-1}$, sets $\hat{W}$ to $\hat{U}$ times the matrix of normalized eigenvectors of $\hat{\Psi}$, and sets $\hat{M} = \hat{W}^+\tilde{Y}_2$, so that $\hat{W}\hat{M} = \tilde{Y}_2$ exactly; a complex or unresolved pencil is reported as unavailable. The current gains are fitted by least squares on the estimated features $\hat{z}_t(x) = \mathrm{diag}(\hat{w}_t)\hat{M}x$ of n independent units per fast task, with inputs $\mathcal{N}(0, I_p)$ and Gaussian noise of variance σ^2, independent of the past sample. With $M_t = \mathrm{diag}(w_t)M$, assume full fast coverage $G_D = \sum_{t\in D} M_t M_t^\top \succeq |D|\kappa I_k$. R_{t° is the slow task's excess squared risk.

Theorem 2 (Error from estimated factors)**.** *For any fixed decomposition $\varrho^\star = a_1\mathbf{1} + a_2\varrho + \omega$, put $c_a = 2|a_1| + C(1 + b_1/\beta)|a_2|$ and $d_s = T + p + \log(C/p_{\mathrm{fail}})$. Let C_{cond} depend only on k and on bounds for the scale and condition quantities of the pencil (Lemma 6): upper bounds on $\mathrm{cond}(W)$, $\|M\|_{\mathrm{op}}$, $\lambda_{\max}(K_\phi)$ and b_1, lower bounds on $\sigma_{\min}(W)$, $\lambda_{\min}(K_\phi)$ and β, and upper and lower bounds on the magnitudes of the past gains. Conditional on the past design*

Gram bounds, a sufficient past sample size for the resolved-gap and stable-coverage events is $n_s \geq C_{\mathrm{cond}}\sigma_s^2 d_s/(\kappa_s \bar{h}^2 \min\{1,\kappa\})$. *Together with* $n \geq C[p + \log(|D|/p_{\mathrm{fail}})]$, *this implies, with probability at least* $1 - p_{\mathrm{fail}}$ *over the past label noise and the current input designs,*

$$\mathbb{E}R_{t^\circ} \leq C_{\mathrm{cond}}\Big(1 + \frac{\xi_{t^\circ}^2}{\kappa}\Big)\Big[\frac{c_a^2\sigma_s^2 d_s}{n_s\kappa_s} + \frac{\|\omega\|_\infty^2\sigma_s^2 d_s}{n_s\kappa_s\bar{h}^2}\Big] + \frac{4\sigma^2 k\xi_{t^\circ}^2}{n|D|\kappa}. \tag{14}$$

The expectation is over the current label noise and an independent served input, and ξ_{t° *bounds the estimated slow feature map* $\hat{M}_{t^\circ}$, *with* $\xi_{t^\circ} \leq \|M_{t^\circ}\|_{\mathrm{op}} + \sqrt{\kappa}/4$ *on the event. If the past Gram event is deduced from Gaussian past inputs, also require* $n_s \geq C[p + \log(1/p_{\mathrm{fail}})]$ *and include its failure probability. No restriction to the span of the past gains is imposed on* $\varrho^\star$.

Conversely, on the two-world subfamily constructed below, with a full-support fast task, identical current fast-label laws and different slow means, the minimax risk is at least $c_1(\varrho_i^\star - \varrho_j^\star)^2 \min\{1, \sigma_s^2/(n_s h^2)\}$ *even with unlimited current fast labels, where the slow task contrasts the factors* i *and* j *of the closest ratio pair and* $\varrho_i^\star - \varrho_j^\star$ *is the difference of their current ratios.*

A known base gives an invertible translation of the observations that does not depend on the parameters, so it preserves the upper-bound experiment and the lower bound's likelihood ratios.

Lemma 6 (Two-window matrix pencil)**.** *Suppose* $\max_e \|\hat{Y}_e - Y_e\|_{\mathrm{op}} \leq \epsilon_Y$. *Set* $\bar{\epsilon} = C\epsilon_Y(\beta^{-1} + b_1\beta^{-2})$. *If* $\epsilon_Y \leq \beta/4$ *and* $\mathrm{cond}(W)\bar{\epsilon} \leq h/4$, *the eigenvalues of* $\hat{\Psi}$ *are real and simple, and the factor columns can be matched in order and sign so that*

$$\eta_W := \|\hat{W} - W\|_{\mathrm{op}} \leq C\sqrt{k}\,(\epsilon_Y/\beta + \mathrm{cond}(W)\bar{\epsilon}/h), \qquad \eta_M := \|\hat{M} - M\|_{\mathrm{op}} \leq \tfrac{2(2\epsilon_Y + \eta_W\|M\|_{\mathrm{op}})}{\sigma_{\min}(W)},$$

where the second bound requires $\eta_W \leq \sigma_{\min}(W)/2$. *Stable coverage additionally requires* $\max_{t\in D\cup\{t^\circ\}} \|\hat{M}_t - M_t\|_{\mathrm{op}} \leq \sqrt{\kappa}/4$, *where* $\hat{M}_t = \mathrm{diag}(\hat{w}_t)\hat{M}$. *These are the resolved-gap and stable-coverage events of Theorem 2.*

Proof. Outline. Align the singular subspaces of $\hat{Y}_2$ (Lemma A(c)), compare the compressed pencil with its population version, and separate its eigenvalues (Lemma A(e)); $\hat{M}$ then follows from $\hat{W}\hat{M} = \tilde{Y}_2$.

Step 1: alignment. Rank-k truncation has error at most $2\epsilon_Y$ by Lemma A(a). By Lemma A(d) and (c), the smallest nonzero singular value β of Y_2 bounds its pseudoinverse perturbation by $C\epsilon_Y/\beta^2$ and aligns its left singular basis with a true basis U up to $C\epsilon_Y/\beta$. After this orthogonal alignment the compressed pencil is within $\bar{\epsilon}$ of $\Psi_0 = U^\top Y_1 Y_2^+ U = (U^\top W)\,\mathrm{diag}(\varrho)(U^\top W)^{-1}$, whose diagonalizer $U^\top W$ has condition number $\mathrm{cond}(W)$.

Step 2: eigenvalues and $\hat{W}$. Lemma A(e), applied to Ψ_0, whose diagonalizer has condition number $\mathrm{cond}(W)$, under the gap assumption $\mathrm{cond}(W)\bar{\epsilon} \leq h/4$, gives disjoint neighbourhoods of the simple real eigenvalues, keeps the matched perturbed eigenvalues real because the pencil has real coefficients, and bounds each eigenvector's line error by $C\,\mathrm{cond}(W)\bar{\epsilon}/h$; adding the singular-subspace error and combining the k unit columns gives the first bound.

Step 3: $\hat{M}$. From $\hat{W}\hat{M} = \tilde{Y}_2$, $\hat{M} - M = \hat{W}^+[\tilde{Y}_2 - Y_2 - (\hat{W} - W)M]$, and $\|\hat{W}^+\|_{\mathrm{op}} \leq 2/\sigma_{\min}(W)$ (Lemma A(a)) gives the second bound. Here $\beta \geq \sigma_{\min}(W)\lambda_{\min}(K_\phi)^{1/2}$ and $b_1 \leq \|W\|_{\mathrm{op}} \max_j |\varrho_j| \lambda_{\max}(K_\phi)^{1/2}$, which exposes both task and input conditioning. □

Proof of Theorem 2. Outline. A resolved-gap event recovers the factors to within order $\epsilon_Y/\bar{h}$ (Step 1), and a stable-coverage event keeps the estimated fast design well conditioned (Step 2). Comparing with ratios built from the estimated past ratios leaves a bias that costs no gap inside the span of the past gains (Step 3), and the bounds combine (Step 4).

Let $\epsilon_Y = \max_e \|\hat{Y}_e - Y_e\|_{\mathrm{op}}$, and let $\eta_{\mathrm{fac}} = \eta_W\|M\|_{\mathrm{op}} + (\|W\|_{\mathrm{op}} + \eta_W)\eta_M$ be the error of the estimated factors.

Step 1: resolved gap. Given window e's past design X_e, the matrix $(\hat{Y}_e - Y_e)(X_e^\top X_e)^{1/2}$ has independent $\mathcal{N}(0,\sigma_s^2)$ entries, so its operator norm exceeds $\sigma_s(\sqrt{T} + \sqrt{p} + \sqrt{2\log(4/p_{\mathrm{fail}})})$ with probability at most $p_{\mathrm{fail}}/4$ (Lemma A(f) with $t = \sqrt{2\log(4/p_{\mathrm{fail}})}$). With $X_e^\top X_e \succeq n_s\kappa_s I_p$ and

a union bound over the two windows this gives the first bound below, with probability at least $1 - p_{\text{fail}}/2$; the stated past sample size then puts ϵ_Y in the resolved regime of Lemma 6, which gives the second:

$$\epsilon_Y^2 \le C\,\frac{\sigma_s^2 d_s}{n_s \kappa_s}, \qquad \eta_{\text{fac}}^2 \le C_{\text{cond}}\,\frac{\epsilon_Y^2}{\bar{h}^2}.$$

Step 2: stable coverage. The stacked-feature triangle inequality and stable coverage give $\hat{G}_D \succeq G_D/2$. Lemma A(f) gives each fast task's design event $I/2 \preceq X_t^\top X_t/n \preceq 2I$ with probability at least $1 - p_{\text{fail}}/(2|D|)$ for the stated n; on these events

$$\hat{Z}^\top \hat{Z} \succeq \tfrac{n}{2}\,\hat{G}_D \succeq \tfrac{n|D|\kappa}{4}\,I.$$

Step 3: risk split. Use the comparison ratios $\varrho_c = a_1\mathbf{1} + a_2\hat{\varrho} + \omega$; the estimator is still unconstrained least squares. The eigendecomposition gives the second anchored identity $\hat{W}\,\mathrm{diag}(\hat{\varrho})\hat{M} = \tilde{Y}_1 := \Pi_{\hat{U}}\hat{Y}_1\Pi_{\hat{V}}$, and because the true slices share row and column spaces, Lemma A(c) gives $\|Y_1 - \tilde{Y}_1\|_{\text{op}} \le C\epsilon_Y(1 + b_1/\beta)$. With $\hat{W}\hat{M} = \tilde{Y}_2$,

$$Y_\star - \hat{W}\,\mathrm{diag}(\varrho_c)\hat{M} = a_1(Y_2 - \tilde{Y}_2) + a_2(Y_1 - \tilde{Y}_1) + (W - \hat{W})\,\mathrm{diag}(\omega)M + \hat{W}\,\mathrm{diag}(\omega)(M - \hat{M}).$$

Rank-k truncation has error at most $2\epsilon_Y$ (Lemma A(a)), so the first two terms have norm at most $c_a\epsilon_Y$ and the last two at most $\|\omega\|_\infty\eta_{\text{fac}}$, and every row Δ_t of the left side satisfies

$$\|\Delta_t\| \le \eta_{\text{row}} := \epsilon_Y c_a + \|\omega\|_\infty\eta_{\text{fac}}.$$

The part of the current gains in the span of the past gains thus has no eigenvector-gap cost, exactly rather than to first order, although its coefficients a_1, a_2 can still amplify past noise. Stack $X_t\Delta_t$, $t \in D$, into ι_D; the current labels are $y = \hat{Z}\varrho_c + \iota_D + \varepsilon$, and the conditional slow risk is exactly $\|\hat{M}_{t^\circ}^\top(\hat{Z}^\top\hat{Z})^{-1}\hat{Z}^\top\iota_D - \Delta_{t^\circ}\|^2 + \sigma^2\,\mathrm{tr}\{\hat{M}_{t^\circ}\hat{M}_{t^\circ}^\top(\hat{Z}^\top\hat{Z})^{-1}\}$. Since $\hat{Z}(\hat{Z}^\top\hat{Z})^{-1}\hat{Z}^\top$ is an orthogonal projection, Step 2 bounds the trace term by $4\sigma^2 k\xi_{t^\circ}^2/(n|D|\kappa)$ and gives, with $\|\iota_D\|^2 \le 2n\sum_{t\in D}\|\Delta_t\|^2 \le 2n|D|\eta_{\text{row}}^2$ and $(x+y)^2 \le 2x^2 + 2y^2$,

$$\|\hat{M}_{t^\circ}^\top(\hat{Z}^\top\hat{Z})^{-1}\hat{Z}^\top\iota_D\|^2 \le \frac{4\xi_{t^\circ}^2\|\iota_D\|^2}{n|D|\kappa}, \qquad \mathbb{E}R_{t^\circ} \le \Big(2 + \frac{16\xi_{t^\circ}^2}{\kappa}\Big)\eta_{\text{row}}^2 + \frac{4\sigma^2 k\xi_{t^\circ}^2}{n|D|\kappa}.$$

Step 4: final bound. By the union bound the events of Steps 1 and 2 hold together with probability at least $1 - p_{\text{fail}}$. On them, $(x+y)^2 \le 2x^2 + 2y^2$ and Step 1 give

$$\begin{aligned}
\eta_{\text{row}}^2 &\le 2c_a^2\epsilon_Y^2 + 2\|\omega\|_\infty^2\eta_{\text{fac}}^2 \le 2\epsilon_Y^2\Big(c_a^2 + C_{\text{cond}}\frac{\|\omega\|_\infty^2}{\bar{h}^2}\Big) \\
&\le 2C\max\{1, C_{\text{cond}}\}\Big[\frac{c_a^2\sigma_s^2 d_s}{n_s\kappa_s} + \frac{\|\omega\|_\infty^2\sigma_s^2 d_s}{n_s\kappa_s\bar{h}^2}\Big], \\
2 + 16\xi_{t^\circ}^2/\kappa &\le 16\,(1 + \xi_{t^\circ}^2/\kappa).
\end{aligned}$$

Inserting both lines into Step 3's bound and replacing C_{cond} by $32C\max\{1, C_{\text{cond}}\}$ gives equation 14. All expectations over the current label noise remain conditional on the designs. □

Proof of the lower bound in Theorem 2. Outline. Two worlds rotate the closest pair of factors by $\pm r$; the past slices then differ by order hr, the current fast row not at all, and the slow means by order $(\varrho_i^\star - \varrho_j^\star)r$, so Le Cam's two-point bound applies.

Step 1: two worlds. Let $p = T = k$ and choose an orthogonal O whose fast row is $w_D^{0\top} = \mathbf{1}^\top/\sqrt{k}$ and whose slow row is $w_{t^\circ}^{0\top} = (\mathbf{e}_i - \mathbf{e}_j)^\top/\sqrt{2}$, the readouts at $r = 0$, where $|\varrho_i - \varrho_j| = h$. Let Θ_r rotate only factors i and j, by the angle r, and put $W_r = O\Theta_r$, $M_r = \Theta_r^\top$, $Y_{2,r} = O$ and $Y_{1,r} = O\Theta_r\,\mathrm{diag}(\varrho)\Theta_r^\top$; both factor condition numbers are one. Choose bounded current ratios $\varrho^\star$ with $\varrho_i^\star \ne \varrho_j^\star$ and give world r the current ratios $\varrho_{r,l}^\star = [\Theta_r^\top\,\mathrm{diag}(\varrho^\star)w_D^0]_l/[\Theta_r^\top w_D^0]_l$. For small enough $|r|$ the denominators are nonzero, these ratios are bounded and the fast row keeps full support.

Step 2: divergence and separation. For the worlds $\pm r$ the only difference in the past slices has squared norm $2h^2\sin^2(2r)$. The fast row's current coefficient is exactly $w_D^{0\top}\,\mathrm{diag}(\varrho^\star)$ in both

worlds, so even infinitely many current fast labels add no divergence. On the (i, j) plane the current input operator is $\mathrm{diag}(\varrho_i^\star, \varrho_j^\star) - (\varrho_i^\star - \varrho_j^\star)\tan(2r) w_{t^\circ}^0 w_{t^\circ}^{0\top}$. Hence

$$D_{\mathrm{KL}}(\mathbb{P}_r, \mathbb{P}_{-r}) = \frac{n_s h^2 \sin^2(2r)}{\sigma_s^2} \le \frac{4 n_s h^2 r^2}{\sigma_s^2}, \qquad \|\text{slow separation}\|^2 = 4(\varrho_i^\star - \varrho_j^\star)^2 \tan^2(2r).$$

Step 3: conclusion. Taking $r = c_0 \min\{1, \sigma_s/(\sqrt{n_s} h)\}$, Le Cam's two-point inequality (Tsybakov, 2009) gives $\inf_{\hat f} \sup_{\pm r} \mathbb{E} R_{t^\circ} \ge c_1 (\varrho_i^\star - \varrho_j^\star)^2 \min\{1, \sigma_s^2/(n_s h^2)\}$. For $k \ge 3$ the remaining entries of $\varrho^\star$ are chosen so that it lies outside the plane of the past gains, an open condition preserved by a small rotation. □

If all ratios lie in an interval $[\varrho_-, \varrho_+]$, then $h \le (\varrho_+ - \varrho_-)/(k-1)$, so the worst-direction cost grows at least quadratically in k. The other two terms of equation 14 have standard lower bounds on their own subfamilies, a past-reference error and a known-design regression.

D.5 Diagonal gains on a pooled basis

Why must the factor basis be learned together with the gains? A model trained on pooled windows fixes its basis only up to an invertible map (Lemma 4(i)), and diagonal gains on that basis generally cannot represent a new window.

Proposition 15 (Pooled basis)**.** *Assume (A1), (A2) and that every coordinate of $\bar s$ is nonzero, and let a pooled model use the basis $\Theta\phi$, Θ invertible, with the exact pooled residual head $W \mathrm{diag}(\bar s)\Theta^{-1}$. (a) Diagonal adaptation $W \mathrm{diag}(\bar s)\Theta^{-1}\mathrm{diag}(s')\Theta\phi$ represents all current task means only if $\Theta^{-1}\mathrm{diag}(s')\Theta = \mathrm{diag}(s^\star \oslash \bar s)$ for some s', which fails for generic Θ when the normalized current gains $s^\star \oslash \bar s$ are non-scalar. With $\bar s = \mathbf{1}$, the $k = 2$ instance of Remark 6 keeps the slow error $(s_1^\star - s_2^\star)^2/2$ as $n \to \infty$. (b) A zero fast residual certifies nothing about the slow task: in the one-head instance of Remark 6, diagonal gains fit the fast task exactly, with slow excess risk 5.*

Proof. (a) By (A1) and (A2), the representation holds for all tasks if and only if $\mathrm{diag}(\bar s)\Theta^{-1}\mathrm{diag}(s')\Theta = \mathrm{diag}(s^\star)$. Normalizing $\bar s = \mathbf{1}$ by absorbing it into ϕ, this reads $\mathrm{diag}(s')\Theta = \Theta\,\mathrm{diag}(s^\star)$, that is, $(s_i' - s_j^\star)\Theta_{ij} = 0$ for all i, j; if every entry of Θ is nonzero, it forces all $s_j^\star$ to be equal, and it has a solution for a scaled permutation Θ or a scalar gain; Remark 6 works out the $k = 2$ instance. (b) is the second instance of Remark 6. □

Remark 6 (Two instances of Proposition 15). On a rotated basis, diagonal gains can keep a slow error that no number of labels removes, or fit a fast head exactly yet miss the slow task.

First instance: $k = 2$, $\phi \sim \mathcal{N}(0, I)$, $\Theta = 2^{-1/2}\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$, fast readouts I_2, slow readout $(1, 1)$ and $s_1^\star \ne s_2^\star$. With $s_\pm' = (s_2' \pm s_1')/2$, the effective matrix is $\Theta^\top \mathrm{diag}(s')\Theta = \left(\begin{smallmatrix} s_+' & s_-' \\ s_-' & s_+' \end{smallmatrix}\right)$; the fast least-squares fit minimizes its Frobenius distance to $\mathrm{diag}(s^\star)$, giving $s_+' = (s_1^\star + s_2^\star)/2$ and $s_-' = 0$, so the slow error stays at $(s_1^\star - s_2^\star)^2/2$ as $n \to \infty$, which is 1.125 for $s^\star = (2, 0.5)$. *Second instance:* with the single fast readout $(1, 0.5)$, the slow readout $(1, -0.5)$ and $s^\star = (2, 0.5)$, the gains $s' = (3.5, 1.5)$ fit the fast task exactly, yet the slow excess risk is 5. In the same basis a full $k \times k$ matrix Υ represents the truth with $\Upsilon_\star = \Theta\,\mathrm{diag}(s^\star)\Theta^{-1}$. As $\mathbb{E}[(\Theta\phi)(\Theta\phi)^\top] \succ 0$, a perturbation of Υ leaves every fast function unchanged if and only if the fast heads annihilate it, so the full matrix is identified if and only if the fast heads have rank k, and the slow task alone if and only if its head lies in their row span (Proposition 4). Under full rank and n labels per fast task, the leading slow excess risk is $\frac{\sigma^2 k}{n} \tilde w_{t^\circ}^\top (\tilde W_D^\top \tilde W_D)^{-1} \tilde w_{t^\circ}$, from the precision $n K_{\Theta\phi} \otimes (\tilde W_D^\top \tilde W_D)/\sigma^2$ of $\mathrm{vec}(\Upsilon)$, with $K_{\Theta\phi} = \mathbb{E}[(\Theta\phi)(\Theta\phi)^\top]$; its factor k counts the k^2 free entries of this parameterization, which diagonal gains in the identified basis avoid.

E What changes beyond linear, correctly specified models?

This appendix supports two statements of Section 4, that Lemma 1 applies exactly to window coordinates that enter linearly and locally to the others, and that binary tasks are handled on the logit scale, and it bounds the bias of an approximately diagonal shift or a wrong number of factors. Lemma 1 and

Propositions 2, 3 and 14 use only the fixed designs z_t of an affine family $\ell_t(x;s) = f_t^0(x) + z_t(x)^\top s$, so they hold whether or not $z_t = w_t \odot \phi$, as for the expert gains and the residual branch of Appendix F; Lemma 3 and Theorem 2 need the canonical polyadic form of equation 1.

E.1 Binary tasks

Do identification and the refit's risk carry over to binary tasks? On the logit scale they do, with the excess log loss in place of the squared error. If the true logit of task t is $\ell_t^\star(x)$, the conditional excess log loss of a logit ℓ is $d_{\log}(\ell, \ell_t^\star) = \log(1+e^{\ell}) - \log(1+e^{\ell_t^\star}) - \mathrm{sig}(\ell_t^\star)(\ell - \ell_t^\star)$, where the logistic function sig is the derivative of $\ell \mapsto \log(1+e^{\ell})$. For $|\ell|, |\ell_t^\star| \le \bar{\ell}$ let $c_{\bar{\ell}} = \mathrm{sig}(\bar{\ell})\,\mathrm{sig}(-\bar{\ell}) > 0$; Taylor's formula gives

$$\frac{c_{\bar{\ell}}}{2}(\ell - \ell_t^\star)^2 \le d_{\log}(\ell, \ell_t^\star) \le \frac{1}{8}(\ell - \ell_t^\star)^2, \tag{15}$$

where the upper bound holds without bounded logits. Coverage thus remains the kernel condition through the injective logistic link, and the squared logit error bounds the excess log loss.

Proposition 16 (Covered logistic refitting)**.** *Condition on fixed offsets and designs for N arrived binary labels, $y_i \sim \mathrm{Bernoulli}\{\mathrm{sig}(o_i + z_i^\top \theta^\star)\}$, independent given the designs. Let $A = Z^\top Z$ and $d_A = \mathrm{rank}(A)$. Let $\mathcal{C}$ be a compact convex parameter set containing $\theta^\star$, with $|o_i + z_i^\top \theta| \le \bar{\ell}$ for all $\theta \in \mathcal{C}$ and all fitting rows, and let $\hat{\theta}$ minimize the logistic loss over $\mathcal{C}$. For a task with the correct fixed logit $o_t(x) + z_t(x)^\top \theta^\star$, suppose $\ker A \subseteq \ker G_t$, with $G_t = \mathbb{E}[z_t(X) z_t(X)^\top]$, and put $\Lambda_t = \lambda_{\max}(A^{+1/2} G_t A^{+1/2})$. Then its expected excess log loss satisfies*

$$\mathbb{E}\,\mathcal{L}_t(\hat{\theta}) - \mathcal{L}_t(\theta^\star) \le \frac{\Lambda_t d_A}{8 c_{\bar{\ell}}^2}. \tag{16}$$

If $A \succeq N\kappa_N P$ with $P = A^+ A$ and $\|G_t\|_{\mathrm{op}} \le \xi_t^2$, this is at most $d_A \xi_t^2 / (8 c_{\bar{\ell}}^2 N \kappa_N)$. The expectation is over the fitting labels and an independent served input.

Proof. Set $\mathcal{U}_\varepsilon = Z^\top (y - p^\star)$, the score of the label noise, and $\Delta = \hat{\theta} - \theta^\star$. The loss Hessian is at least $c_{\bar{\ell}} A$ on $\mathcal{C}$, and the basic inequality for the constrained minimizer gives $\frac{c_{\bar{\ell}}}{2}\|\Delta\|_A^2 \le \mathcal{U}_\varepsilon^\top \Delta \le \|\mathcal{U}_\varepsilon\|_{A^+} \|\Delta\|_A$, where $\mathcal{U}_\varepsilon \in \mathrm{range}(A)$ even if A is singular. Conditional Bernoulli independence gives $\mathbb{E}\|\mathcal{U}_\varepsilon\|_{A^+}^2 = \mathrm{tr}\{A^+ Z^\top \mathrm{diag}(p_i^\star(1-p_i^\star)) Z\} \le d_A/4$. Coverage implies $\Delta^\top G_t \Delta \le \Lambda_t \|\Delta\|_A^2$; the upper half of equation 15 combines these bounds. □

Corollary 4 (Remainders and an unread task)**.** *If the true fitting logits are $o + Z\theta_c + \iota_D$ and the true served logit is $o_t + z_t^\top \theta_c + \iota_t$ for a fixed $\theta_c \in \mathcal{C}$, with $\bar{\iota}_t = \|\iota_t\|_{L^2}$, then $\mathbb{E}[\mathcal{L}_t(\hat{\theta}) - \mathcal{L}_t(\ell_t^\star)] \le \frac{1}{8}\big[\sqrt{\Lambda_t d_A}/c_{\bar{\ell}} + \sqrt{\Lambda_t}\|\iota_D\|_2/(2c_{\bar{\ell}}) + \bar{\iota}_t\big]^2$, and if two feasible parameters give the arrived labels the same law but served logits at L^2 distance r, both within $[-\bar{\ell}, \bar{\ell}]$, every predictor has worst-case excess log loss at least $c_{\bar{\ell}} r^2/8$.*

The upper bound follows from the argument of Proposition 16, the $\frac{1}{4}$-Lipschitz logistic function and the projection $ZA^+Z^\top$, so past-component error, model error and drift enter through logit remainders; the past fits must be independent of the current labels, and with dependent labels $d_A/4$ becomes $\mathrm{tr}(A^+ \mathrm{Cov}(\mathcal{U}_\varepsilon))$. For the lower bound, clipping a predictor's logit to $[-\bar{\ell}, \bar{\ell}]$ improves both conditional losses, the squared-error two-point bound gives $r^2/4$, and the lower half of equation 15 converts it.

E.2 Nonlinear window coordinates

When is a window coordinate that enters the model nonlinearly identified from the fast labels, and how far does the linearization at the pilot carry? Fix the trained parameters and write the mean or logit as $\ell_t(x; \varsigma \odot s)$, with ς the root-mean-square scale of the past coordinates (Appendix F.1).

The exact condition. Let $\mathcal{A}(s) = (\ell_d(\cdot; \varsigma \odot s))_{d \in D}$ in the direct sum of the fast-input L^2 spaces and $\mathcal{S}_t(s) = \ell_t(\cdot; \varsigma \odot s)$ in the served-input L^2 space. The current mean of task t is identified over an allowed set $\mathcal{K}$ if and only if

$$\mathcal{A}(s) = \mathcal{A}(s') \implies \mathcal{S}_t(s) = \mathcal{S}_t(s') \qquad (s, s' \in \mathcal{K}), \tag{17}$$

by equality of the fast conditional laws, through the injective logistic link for binary labels and gain-independent location noise for regression. For a model linear in the gate values, it suffices to check its fast kernel on the differences between gate values that the gate allows. A Jacobian checks only infinitesimal differences: $2\,\mathrm{sig}(s) - 2\,\mathrm{sig}(2s)$ takes the same value at $s = .5$ and $s \simeq 1.89$, although its derivative at .5 is nonzero and the slow logit $2\,\mathrm{sig}(s)$ differs, so no number of fast labels separates the two values. On a bounded sigmoid gate the unbounded-risk clause of Proposition 2 is replaced by the two-point bound $\|\mathcal{S}_t(s) - \mathcal{S}_t(s')\|^2/4$.

Lemma 7 (Near-affine region)**.** *Let $\mathcal{K}$ be convex and $\mathcal{A}, \mathcal{S}_t$ continuously differentiable on a neighbourhood of $\mathcal{K}$. Fix $s_0 \in \mathcal{K}$, write $J_D = \mathcal{A}'(s_0)$ and $J_t = \mathcal{S}_t'(s_0)$, and bound the drift of the Jacobians by $\sup_{s\in\mathcal{K}} \|\mathcal{A}'(s) - J_D\|_{\mathrm{op}} \le \chi_D$ and $\sup_{s\in\mathcal{K}} \|\mathcal{S}_t'(s) - J_t\|_{\mathrm{op}} \le \chi_t$. If tangent coverage holds, let $\bar\Lambda_t = \sup_\Delta \|J_t\Delta\|^2/\|J_D\Delta\|^2 < \infty$, over Δ with $J_D\Delta \neq 0$. Then, for all $s, s' \in \mathcal{K}$,*

$$\|\mathcal{S}_t(s) - \mathcal{S}_t(s')\| \le \bar\Lambda_t^{1/2}\|\mathcal{A}(s) - \mathcal{A}(s')\| + (\bar\Lambda_t^{1/2}\chi_D + \chi_t)\|s - s'\|. \tag{18}$$

If J_D has full column rank with smallest singular value $\beta_D > \chi_D$ and $\|\mathcal{S}_t'\|_{\mathcal{K}} = \sup_{s\in\mathcal{K}} \|\mathcal{S}_t'(s)\|_{\mathrm{op}}$, then $\mathcal{A}$ is one-to-one on $\mathcal{K}$ and

$$\|s - s'\| \le \frac{\|\mathcal{A}(s) - \mathcal{A}(s')\|}{\beta_D - \chi_D}, \qquad \|\mathcal{S}_t(s) - \mathcal{S}_t(s')\| \le \frac{\|\mathcal{S}_t'\|_{\mathcal{K}}}{\beta_D - \chi_D}\|\mathcal{A}(s) - \mathcal{A}(s')\|;$$

the second constant can be replaced by $\bar\Lambda_t^{1/2} + (\bar\Lambda_t^{1/2}\chi_D + \chi_t)/(\beta_D - \chi_D)$. If J_D is rank deficient, Q projects onto its kernel, $P = I - Q \neq 0$, and $\mathcal{A}'(s)Q = \mathcal{S}_t'(s)Q = 0$ throughout $\mathcal{K}$, the same conclusion holds for task t with β_D the least singular value of J_D on the range of P, and the first bound controls $\|P(s - s')\|$.

Proof. Outline. Along the segment from s' to s each map is its tangent map plus a remainder of at most the Jacobian drift times the step; tangent coverage bounds the tangent parts.

Step 1. With $\Delta = s - s'$, integrating the derivatives along the segment gives

$$\mathcal{A}(s) - \mathcal{A}(s') = J_D\Delta + \tilde\iota_D, \quad \mathcal{S}_t(s) - \mathcal{S}_t(s') = J_t\Delta + \tilde\iota_t, \qquad \|\tilde\iota_D\| \le \chi_D\|\Delta\|, \ \ \|\tilde\iota_t\| \le \chi_t\|\Delta\|.$$

Step 2. Tangent coverage gives $\|J_t\Delta\| \le \bar\Lambda_t^{1/2}\|J_D\Delta\| \le \bar\Lambda_t^{1/2}(\|\mathcal{A}(s) - \mathcal{A}(s')\| + \chi_D\|\Delta\|)$, and the triangle inequality with Step 1 gives equation 18.

Step 3. Under full column rank,

$$\|\mathcal{A}(s) - \mathcal{A}(s')\| \ge \|J_D\Delta\| - \|\tilde\iota_D\| \ge (\beta_D - \chi_D)\|\Delta\|, \qquad \|\mathcal{S}_t(s) - \mathcal{S}_t(s')\| \le \|\mathcal{S}_t'\|_{\mathcal{K}}\|\Delta\|,$$

the second by integrating $\mathcal{S}_t'$; together they give both displayed bounds, and substituting the first into equation 18 gives the alternative constant.

Step 4. Under the common-kernel condition every derivative annihilates $Q\Delta$, and Steps 1–3 run on $P\Delta$. □

If $\mathcal{K}$ lies in a ball of radius $R_{\mathcal{K}}$ around s_0 and the curvature is bounded, $\|\nabla_s^2\ell_t(x; \varsigma \odot s)\|_{\mathrm{op}} \le \upsilon_t(x)$, then $\chi_D \le \upsilon_D R_{\mathcal{K}}$ and $\chi_t \le \upsilon_t R_{\mathcal{K}}$ with $\upsilon_D^2 = \sum_{d\in D} \mathbb{E}\upsilon_d(X)^2$ and $\upsilon_t = \{\mathbb{E}\upsilon_t(X)^2\}^{1/2}$, so $\upsilon_D R_{\mathcal{K}} < \beta_D$ is a sufficient small-curvature condition relative to the strength of the fast design; it must hold on a region containing both the true and the served coordinates, which the pilot's Hessian alone does not establish. Without full rank or a common kernel, equation 18 leaves an approximate-identification floor of at most $(\bar\Lambda_t^{1/2}\chi_D + \chi_t)\,\mathrm{diam}(\mathcal{K})$ in root risk: $\mathcal{A}(s_1, s_2) = s_1$ and $\mathcal{S}(s_1, s_2) = s_1 + rs_2^2$ have covered tangent designs at the origin for every $r > 0$ but are not identified. For a gate with a linear readout $\bar w_{tj}(x)$ of its units, $\ell_t(x; \varsigma \odot s) = o_t(x) + \sum_j \bar w_{tj}(x)\,2\,\mathrm{sig}(\Gamma_{j:}(\varsigma \odot s) + b_{\Gamma,j})$, with $\Gamma_{j:}$ the jth row of the gate matrix Γ of Appendix F.1, take $\upsilon_t(x) = \frac{1}{3\sqrt{3}}\sum_j |\bar w_{tj}(x)|\,\|\Gamma_{j:}\,\mathrm{diag}(\varsigma)\|^2$, since $\sup_z |(2\,\mathrm{sig})''(z)| = 1/(3\sqrt{3})$; a nonlinear tower after the gate adds its derivatives by the chain rule.

The one-step update. At the pilot s_0, let H and $\mathcal{U}_0$ be the information and score of Appendix F.1 ($H = Z^\top\Sigma_\varepsilon^{-1}Z$ and $\mathcal{U}_0 = Z^\top\Sigma_\varepsilon^{-1}(y - \ell_0)$ for continuous labels), $\theta^\star = s^\star - \bar s$, $P = H^+H$, and consider an update $s_\Xi = \bar s + \Xi\hat\theta$ with $\Xi P = \Xi$, as for θ_A and θ_W. Substituting these definitions into $\mathcal{U}_0$ gives, pathwise, $\hat\theta = P\theta^\star + H^+(\mathcal{U}_\varepsilon + \mathcal{U}_\iota)$ and

$$\ell_t(x; \varsigma \odot s_\Xi) - \ell_t(x; \varsigma \odot s^\star) = J_t(x)^\top\{(\Xi - I)\theta^\star + \Xi H^+(\mathcal{U}_\varepsilon + \mathcal{U}_\iota)\} + \iota_t(x; s_\Xi) - \iota_t(x; s^\star).$$

Here $\mathcal{U}_\varepsilon$ is the score of the label noise ($Z^\top\Sigma_\varepsilon^{-1}\varepsilon$, or $Z^\top\varepsilon$ for binary labels); $\mathcal{U}_\iota$ is the score of the remainder $\iota_D = \ell(s^\star) - \ell_0 - Z(s^\star - s_0)$ of the arrived logits, $Z^\top\Sigma_\varepsilon^{-1}\iota_D$, or for binary labels $Z^\top\iota^p$ with $\iota^p = \mathrm{sig}(\ell(s^\star)) - p_0 - \mathrm{diag}\{p_0(1-p_0)\}Z(s^\star - s_0)$ and $|\iota_i^p| \le \frac{1}{4}|\iota_{D,i}| + |\ell_i(s^\star) - \ell_i(s_0)|^2/(12\sqrt{3})$; and $\iota_t(x;s) = \ell_t(x;\varsigma\odot s) - \ell_t(x;\varsigma\odot s_0) - J_t(x)^\top(s - s_0)$. The terms in $\theta^\star$ and $\mathcal{U}_\varepsilon$ are the fixed-affine error, $\Xi H^+\mathcal{U}_\iota$ is fast-label curvature amplified by the inverse information, and the served-label curvature is at most $\frac{\upsilon_t}{2}(\|s_\Xi - s_0\|^2 + \|s^\star - s_0\|^2)$ in L^2. In the correct affine model the remainders vanish and $\pi_t + \nu_t$ is the risk of θ_W (Proposition 14); in the nonlinear procedure π_t and ν_t are local working quantities on the logit scale.

E.3 Approximately diagonal shifts and a wrong number of factors

Errors of the fast tasks' model can be mistaken for shared gain changes and passed to the slow task; more current labels reduce the noise but not this bias, and a deviation private to the slow task needs its own labels or a prior bound.

Identification up to residuals. Write $\mu_t = z_t^\top s^\star + \iota_t$ with $\|\iota_{t^\circ}\|_{L^2} \le \bar\iota_{t^\circ}$ and $\sum_{d\in D}\|\iota_d\|_{L^2}^2 \le \bar\iota_D^2$, and under coverage let $\bar\Lambda_{t^\circ} = \sup\{\Delta^\top G_{t^\circ}\Delta/\Delta^\top G_D\Delta : \Delta^\top G_D\Delta > 0\} < \infty$. Two admissible worlds with identical fast means have slow means within $2\bar\iota_{t^\circ} + 2\bar\Lambda_{t^\circ}^{1/2}\bar\iota_D$: their gain difference has fast norm at most $2\bar\iota_D$, which $\bar\Lambda_{t^\circ}^{1/2}$ carries to the slow task, and the two slow residuals add. Without coverage the impossibility of Proposition 2 remains, and arbitrary residuals private to the slow task allow no finite diameter even under coverage.

Fixed-design bias. If $y = o + Zs^\star + \iota_D + \varepsilon$, the refit of Proposition 2 has exact risk

$$R_t = \left\|z_t^\top Q(\bar s - s^\star) + z_t^\top A^+ Z^\top\iota_D - \iota_t\right\|_{L^2}^2 + \sigma^2\,\mathrm{tr}(G_tA^+), \tag{19}$$

by expanding the estimator and averaging the mean-zero noise. If every bias other than the unread one is at most ϵ_t in norm, $R_t \le \sigma^2\,\mathrm{tr}(G_tA^+) + (\|G_t^{1/2}Q(s^\star - \bar s)\| + \epsilon_t)^2$, the two terms of equation 6 with the added bias inside the square. Under full design coverage $A \succeq N\kappa_N I$, with $\|G_t\|_{\mathrm{op}} \le \xi_t^2$, $\|\iota_t\|_{L^2} \le \bar\iota_t$ and $\|\iota_D\|^2/N \le \bar\iota_{D,N}^2$, projection contraction gives $R_t \le 2\bar\iota_t^2 + 2\xi_t^2\bar\iota_{D,N}^2/\kappa_N + \sigma^2 k\xi_t^2/(N\kappa_N)$, a budget on the actual design; a population budget needs its own concentration step.
Remark 7 (Approximately diagonal past shifts). Systematic errors of the past and current slices enter the bound of Theorem 2 through terms that more past samples cannot remove, and the rate holds only while they stay within the resolved-gap allowance.

In Theorem 2 let the true past slices be $Y_e + \mathcal{B}_e$, where $\mathcal{B}_e$ is $W\Upsilon_e M$, with Υ_e a small off-diagonal shift shared by the tasks, plus a task-private coefficient matrix, and $\max_e\|\mathcal{B}_e\|_{\mathrm{op}} \le \epsilon_{\mathrm{sys}}$; an estimated base adds its projected error to $\mathcal{B}_e$. The pencil and stable-coverage events must then hold for the total perturbation $\epsilon_Y = \epsilon_{\mathrm{stat}} + \epsilon_{\mathrm{sys}}$. If the true current slice is $W\,\mathrm{diag}(\varrho^\star)M + \mathcal{B}_\star$, with slow residual row norm at most $\bar\iota_{t^\circ}$ and average squared fast row norm at most $\bar\iota_D^2$, then on the same events, keeping the comparison ratios ϱ_c of the proof and stacking $X_t(\Delta_t + \mathcal{B}_{\star t})$ into ι_D, with the rows $\mathcal{B}_{\star t}$ of $\mathcal{B}_\star$,

$$\mathbb{E}R_{t^\circ} \le (4 + 32\xi_{t^\circ}^2/\kappa)\,\eta_{\mathrm{row}}^2 + 4\bar\iota_{t^\circ}^2 + 32\xi_{t^\circ}^2\bar\iota_D^2/\kappa + 4\sigma^2 k\xi_{t^\circ}^2/(n|D|\kappa).$$

The events hold when $C\sigma_s\sqrt{d_s/(n_s\kappa_s)} + \epsilon_{\mathrm{sys}} \le c_{\mathrm{cond}}\bar h\min\{1,\sqrt{\kappa}\}$.

A wrong number of factors. If a true past slice has rank k but the fitted width is $k' < k$, its squared approximation error under isotropic inputs is at least $\sum_{j>k'}\sigma_j(Y_2)^2$; a slow task need not use the omitted directions, but coverage in the truncated model cannot certify that it does not. Overestimating the rank breaks uniqueness: duplicate a one-dimensional factor into two columns, give two tasks the heads $(1,0)$ and $(0,1)$ and the past gains (s_e, s_e); every past window is fitted exactly, yet a current label of the first task leaves the second gain arbitrary, whereas the correct one-factor model identifies the common gain.

F How does FLAG compute its update?

This appendix specifies every step of Section 4 with its constants, checks and fallbacks (Algorithm 1), prices a bet on the matured windows, varies the threshold that defines U, and treats a task's

own labels once they arrive, which is what a waiting task collects (Appendix F.2). Table 6 says how each object of the known model of Section 3 carries over to the fitted procedure.

Table 6: From the known model of Section 3 to the fitted FLAG: how each object carries over.

object (known → fitted)	carries over as
gains → coordinates	reparametrization; q below the number of moving factors makes coverage certify nothing about the rest (Appendix E.3)
$z_t \to$ Jacobian at s_0	exact if the window acts diagonally in the basis ((ii): CP; (i), (iii): the architecture's), which only the check of Section 4 tests
Z, G_t	exact for affine constructions, up to the support tolerance; local for (iii) (Appendix E.2)
$\ker Z \to U$	the eigendirections of F with $\lambda_j < 1$; an eigenspace of F, so Theorem 1 holds for the labels' information off U (below)
$Z^\top Z/\sigma^2 \to F \preceq H$	relative to the estimated cluster covariance (Appendix F.1)
$vI \to \bar{s}, \Sigma, v$	empirical Bayes; a true size above $v/3$ keeps the expected gain on every direction read at $\lambda_j \geq 1$ (Appendix C.1)
refit $\to \theta_W, \delta_t^*, \theta_A$	$\theta_W \to$ refit as $v \to \infty$; $\delta_t^* = (I - P_t)\theta_W$; θ_A enters only through its bet $P_t\theta_A$
P_t, predicted gain	exact for affine designs, with G_t estimated on the served inputs; logit-scale working quantities if binary (Appendix E.1); local for the gate
bets	priced on at least 40 matured windows, each rebuilt as it would have been served (below)
selection	chosen on the latest matured window, whose ranking it carries over to the served window (Appendix F.3)
factors, Thm. 2	not guarantees of FLAG: which basis to learn, and the price of the learned part, which the check tests

F.1 WINDOW COORDINATES, PRIOR AND INFORMATION

Which coordinates does FLAG refit, which prior does it place on them, and how much information do the arrived labels carry about them? This subsection fixes each with every constant.

Constructions. An *embedding gate* acts on the frozen hidden representation $\mathcal{H}_t(x)$ of a task tower $\mathcal{T}_t$, $\ell_t(x; \tilde{s}) = \mathcal{T}_t\{x, \mathcal{H}_t(x) \odot 2\,\mathrm{sig}(\Gamma\tilde{s} + b_\Gamma)\}$, with the window embedding $\tilde{s}$ as its coordinate. An added gate has $b_\Gamma = 0$ and centred past embeddings, so $\tilde{s} = 0$ recovers the base because $2\,\mathrm{sig}(0) = 1$; a native gate, such as PEPNet's, has its reference at the mean past embedding. The affine constructions have $\ell_t(x; \tilde{\theta}) = f_t^0(x) + z_t(x)^\top\tilde{\theta}$ with centred innovations and reference $\tilde{\theta} = 0$: *expert gains* on a frozen MMoE (Ma et al., 2018a) rescale expert j by $1 + \tilde{\theta}_j$, so $z_t(x)_j$ is task t's frozen gate weight on expert j times its frozen head applied to the expert's output $\mathcal{H}_j(x)$, a task-specific design rather than a shared factor head; a *residual branch* is model equation 1 on the frozen hidden layer $\mathcal{H}(x)$, $z_t(x) = w_t \odot \{M_{\mathrm{res}}\mathcal{H}(x) + b_{\mathrm{res}}\}$, learned with per-window gains on the past windows.

Coordinates and prior. Let a candidate construction have fitted past coordinates $\tilde{s}_e \in \mathbb{R}^q$. Their root-mean-square scale is $\varsigma_j = (m^{-1}\sum_e \tilde{s}_{ej}^2)^{1/2}$, and the coordinates used throughout are $s_e = \tilde{s}_e \oslash \varsigma$, so that the model in the construction's own coordinates is $\ell_t(x; \varsigma \odot s)$. These rescaled coordinates are the window gains of Section 4: the main text's s, $\bar{s}$, Σ and updates are those of the rescaled coordinates, and its $\ell_t(x; s)$ means $\ell_t(x; \varsigma \odot s)$. With $\bar{s}$ the mean of the s_e and $\theta_e = s_e - \bar{s}$, let $\widehat{\Sigma}_0 = (m-1)^{-1}\sum_e(\theta_e - \bar{\theta})(\theta_e - \bar{\theta})^\top$ be their sample covariance, from which Section 4 builds Σ, and $v_0 = \mathrm{tr}(\widehat{\Sigma}_0)/q$. As in Section 4, the prior covariance averages $\widehat{\Sigma}_0$, of rank at most $m-1$, with $(\mathrm{tr}\,\widehat{\Sigma}_0/q)\,I = v_0 I$ and is scaled by τ; a small ridge keeps it invertible:

$$\Sigma = \tau\Big[\tfrac{1}{2}\widehat{\Sigma}_0 + \tfrac{1}{2}v_0 I + 10^{-3}\max(1, v_0)I\Big],$$

with the prior scale $\tau > 0$ chosen on validation, which scales both priors and the pilot's penalty alike. The size v is the 10% quantile of $\mathrm{tr}(\Sigma_b)/q$ over 200 bootstrap resamples b of the past windows, with ς fixed; the running example, with two past windows, uses $v = \mathrm{tr}(\Sigma)/q$. Writing an affine construction as $\ell_t = o_t + z_t^\top s$ with $o_t + z_t^\top\bar{s} = f_t^0$, the translation $s = \bar{s} + \theta$ turns the priors $\mathcal{N}(\bar{s}, \Sigma)$ and $\mathcal{N}(\bar{s}, vI)$ into zero-centred priors on the innovation without changing the design; for expert gains $\tilde{s}_e = \mathbf{1} + \tilde{\theta}_e$, so $\bar{s} = \mathbf{1} \oslash \varsigma$ and only the innovation is centred. Because $\widehat{\Sigma}_0$ needs at least two past windows (one centred window has no dispersion), the FLAG candidates are excluded when the validation or the serving fit has fewer than two past windows.

Pilot, information and effective precision. The pilot s_0 minimizes the negative log likelihood of the arrived labels plus $\|s - \bar{s}\|^2/(2v)$. At s_0, the Jacobians $\partial\ell_{d_i}(x_i; \varsigma \odot s)/\partial s$ of the arrived labels

are the rows of the design Z of Section 4, and ℓ_0 stacks their predictions. For continuous labels with working noise covariance Σ_ε, $H = Z^\top \Sigma_\varepsilon^{-1} Z$ and the score is $\mathcal{U}_0 = Z^\top \Sigma_\varepsilon^{-1}(y - \ell_0)$; for binary labels, with $p_0 = \mathrm{sig}(\ell_0)$, $H = Z^\top \operatorname{diag}\{p_0(1-p_0)\} Z$ and $\mathcal{U}_0 = Z^\top (y - p_0)$. The updates use the information F of Section 4: the cluster-robust version of the model-based information H, capped at H so that correlation between the labels of one unit cluster can only lower it. The eigenvectors O_r of H whose eigenvalues exceed $10^{-12}\lambda_{\max}(H)$ form its numerical support, with those eigenvalues in the diagonal matrix D_r, and the pseudo-observation is $\hat\theta = H^+\{H(s_0 - \bar s) + \mathcal{U}_0\}$. Summing the per-label scores over all arrived tasks of one unit cluster (user, asker or environment) gives its score $\mathcal{U}_{\mathrm{cl}}$; over the n_{cl} clusters, with mean $\bar{\mathcal{U}}$, these give the sandwich estimate of the score covariance Ω (White, 1982; Liang & Zeger, 1986)

$$\widehat\Omega = \frac{n_{\mathrm{cl}}}{n_{\mathrm{cl}} - 1} \sum_{\text{clusters}} (\mathcal{U}_{\mathrm{cl}} - \bar{\mathcal{U}})(\mathcal{U}_{\mathrm{cl}} - \bar{\mathcal{U}})^\top,$$

$$\Omega_r = D_r^{-1/2} O_r^\top \widehat\Omega O_r D_r^{-1/2}, \qquad F = O_r D_r^{1/2} (\Omega_r \vee I)^{-1} D_r^{1/2} O_r^\top,$$

where $\Omega_r \vee I$ raises the eigenvalues of Ω_r below one to one, and an empty support gives $F = 0$. This is the cap: $F \preceq H$, with equality on the support when no eigenvalue of Ω_r exceeds one, and F^+ dominates the sandwich covariance of $\hat\theta$, so correlation can only lower the information. The updates are $\theta_W = (v^{-1} I + F)^{-1} F\hat\theta$ and $\theta_A = (\Sigma^{-1} + F)^{-1} F\hat\theta$, with $C = (I + vF)^{-1}$. The blind subspace U is spanned by the eigenvectors of F with $vf_j < 1$, the kernel of F included, and B holds an orthonormal basis of it. G_t is the second moment of task t's Jacobian on the unlabelled inputs available before the split, $P_t = B(B^\top G_t B)^+ B^\top G_t$, where the pseudoinverse treats eigenvalues of $B^\top G_t B$ below $10^{-12}\operatorname{tr}(G_t)$ as zero, $\delta_t^* = (I - P_t)\theta_W$, and its predicted gain is $v \operatorname{tr}\{(I - P_t)^\top G_t (I - P_t)(I - C)\}$ (Appendix C.1). On the read directions, with $K = (I - P_t)^\top G_t (I - P_t)$, ζ_t^2 is the largest generalized eigenvalue of (CKC, K); if $\zeta_t > 1$, the size-robust update $(I - P_t)\{\theta_W + (1 - \zeta_t^{-1})(\hat\theta - \theta_W)\}$ replaces δ_t^*, with predicted gain $v \operatorname{tr} K - v\zeta_t^{-2} \operatorname{tr}(CKC) - \operatorname{tr}\{(I - \zeta_t^{-1} C) K (I - \zeta_t^{-1} C) F^{-1}\}$.

Pricing a bet. A bet is priced only when at least 40 matured windows exist, the past windows whose slow labels have all matured by the served window's training time and that have at least two windows before them. On each matured window e, FLAG is rebuilt as it would have been served there, with the construction fitted for the served window: the anchor, Σ and v from the windows before e; the pilot, $\hat\theta$, F and U from e's early labels; G_t and P_t on e's served inputs; and the two candidate bets $P_t\theta_A$ and $P_t\theta_W$ of Section 4. The change that e's matured labels reveal, θ_e, is its past coordinate minus that anchor. For each candidate, $\hat\rho_t = \sum_e \langle \beta_{t,e}, P_t\theta_e \rangle_{G_t} / \sum_e \|\beta_{t,e}\|^2_{G_t}$, with each window's own P_t and G_t. By equation 10, the bets would have changed the task's summed error by $\sum_e \|\beta_{t,e}\|^2_{G_t}(1 - 2\hat\rho_t)$, so a candidate passes if $\hat\rho_t > \frac{1}{2}$, and of two passing candidates the one with the larger summed gain $\sum_e \|\beta_{t,e}\|^2_{G_t}(2\hat\rho_t - 1)$ is served. With fewer than 40 matured windows, or no bet with a positive stake, the projected update is served alone.

States. Per slow task, OFF serves the frozen model, the fallback of Section 4; PROJECTED serves $\bar s + \delta_t^*$, and the task *adapts* if $P_t = 0$ and *waits* on its blind spot otherwise; BET serves $\bar s + \delta_t^* + \beta_t$; and UNAVAILABLE records that a quantity of the update cannot be computed, in which case the frozen model is served. The fast tasks are always served by the fine-tuned copy, and the frozen model of a native gate is its evaluation at the mean past embedding. Algorithm 1 writes each served prediction in the construction's own coordinates, $\ell_t(x; \varsigma \odot (\bar s + \theta))$; in the rescaled coordinates of this subsection, those of Section 4, this is $\ell_t(x; \bar s + \theta)$.

Varying the threshold. The threshold 1 on λ_j is posited; along each eigendirection of F it is where the part of θ_W's risk left to the past equals its label noise (Proposition 14(i)). Any threshold $c > 0$ keeps every statement of Theorem 1 for the experiment that retains the labels' information on the directions with $\lambda_j \geq c$: because U is an eigenspace of F, δ_t^* depends only on the part of $\hat\theta$ off U, whose law does not depend on θ's part in U. Raising c puts more weakly read directions into U, so the projected update corrects fewer directions, its predicted gain falls, and a covered task that is read weakly, such as purchase with the weak reader, waits longer on its blind spot, the switch moving in proportion to c because λ_j grows linearly with the labels per task. Lowering c toward zero shrinks U to the numerical kernel of F, and the projected update then corrects directions read

Algorithm 1 FLAG in one served window, with every check and fallback; $\ell_t(x; \varsigma \odot s)$ is the $\ell_t(x; s)$ of Section 4 in the construction's own coordinates.

Require: frozen model; matured past labels; arrived pairs $\mathcal{O}$ with unit clusters; unlabelled served inputs

Step 1: configuration and fallback ▷ §4; App. F.3

If fewer than two past windows precede the validation window or the served window, serve the frozen model on the slow tasks and **return**

On the latest matured window, rebuild FLAG for each configuration (construction, q, τ) and keep the one with the lowest mean slow-task loss; mark OFF each slow task whose loss under it exceeds the frozen model's by more than one standard error

Fine-tune a copy of the frozen model on the arrived fast labels; serve the fast tasks from it and the OFF tasks from the frozen model ▷ Prop. 1

Step 2: anchor and size ▷ §4; App. F.1

Fit the selected construction on the matured past labels (or use the native gate); freeze all but its window coordinate

Fix ς and $\bar{s}$; $\Sigma \leftarrow \tau[\frac{1}{2}\widehat{\Sigma}_0 + \frac{1}{2}v_0 I + 10^{-3}\max(1, v_0)I]$ ($\widehat{\Sigma}_0$: sample covariance, $v_0 = \operatorname{tr}\widehat{\Sigma}_0/q$); $v \leftarrow$ lower 10% bootstrap quantile of $\operatorname{tr}\Sigma/q$ over past windows

If these are invalid, no label has arrived or fewer than two clusters are labelled, serve the frozen model on slow tasks, record UNAVAILABLE, **return**

Step 3: pilot, design and information ▷ §4; sandwich

Fit the pilot s_0 on $\mathcal{O}$; if its first-order condition fails or its Hessian is not positive, record UNAVAILABLE, **return**

Form Z, $H, \mathcal{U}_0$, the support (O_r, D_r) of H above $10^{-12}\lambda_{\max}(H)$, and $\hat{\theta} = H^{+}\{H(s_0 - \bar{s}) + \mathcal{U}_0\}$

Sum scores per cluster; form $\widehat{\Omega}$, Ω_r and $F \preceq H$; if any is invalid, record UNAVAILABLE, **return**

Step 4: updates and blind subspace ▷ Thm. 1

$\theta_W \leftarrow (v^{-1}I + F)^{-1}F\hat{\theta}$; $\theta_A \leftarrow (\Sigma^{-1} + F)^{-1}F\hat{\theta}$; $C \leftarrow (I + vF)^{-1}$

$B \leftarrow$ orthonormal eigenvectors of F with $vf_j < 1$, its kernel included

Step 5: projected update and bet, per slow task not OFF ▷ §4

for each such slow task t **do**

  Form G_t on the unlabelled served inputs; if it or a score is invalid, or $\operatorname{tr}(G_t) \le 0$, serve the frozen model, record UNAVAILABLE, and continue

  $P_t \leftarrow B(B^\top G_t B)^{+}B^\top G_t$; $\delta_t^* \leftarrow (I - P_t)\theta_W$; if $\zeta_t > 1$, replace δ_t^* by the size-robust update; record the predicted gain

  if at least 40 matured windows exist and a candidate $\beta_t \in \{P_t\theta_A, P_t\theta_W\}$ has $\hat{\rho}_t > \frac{1}{2}$ **then**

    serve $\ell_t(x; \varsigma \odot (\bar{s} + \delta_t^* + \beta_t))$ with the passing candidate of larger summed gain; record BET

  else

    serve $\ell_t(x; \varsigma \odot (\bar{s} + \delta_t^*))$; record PROJECTED, adapting if $P_t = 0$ and waiting otherwise ▷ App. F.2

Step 6: scoring ▷ §5.2

When the slow labels mature, score every served update and split its regret by equation 9 (Appendix H.4)

less precisely than a typical change, where the label noise can exceed what not adapting leaves, as the weak reader's .53 against .124 shows. In the replay of Section 5.1, thresholds of $\frac{1}{4}$, $\frac{1}{2}$, 2 and 4 change FLAG's mean slow-task error by at most .003 on every dataset, and it stays below that of the best baseline.

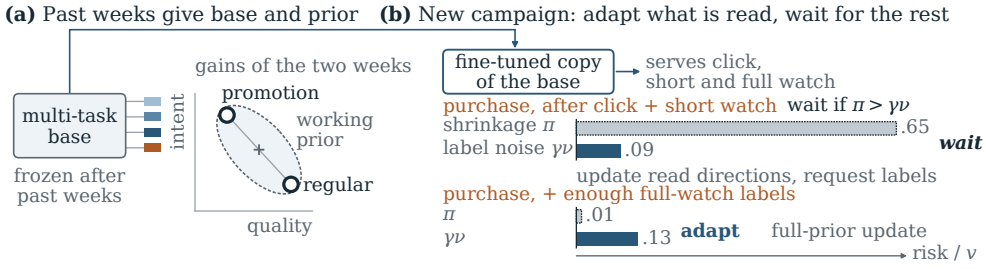


Figure 6: (a) The prior from the two weeks; the projected update uses only its size v. (b) Purchase, 30 early labels per fast task: intent counts as read once $\lambda \ge 1$; bars show the error the projected update removes, beyond label noise.

F.2 What a task's own labels add

What do a waiting task's own labels add once they arrive, and what changes when some slow labels have arrived by decision time? They enter Z and F like any other label, and a task's own labels read its blind spot, so U shrinks for it. Let $\mathcal{O}_{\mathfrak{a}}$ contain every matured fast-label pair and every slow-label pair that has arrived, indexed by a nested arrival level $\mathfrak{a}$. One window coordinate serves all pairs of the likelihood, each task has its own working noise variance (or a joint logistic objective), and the pilot, Jacobians, information and cluster scores use exactly $\mathcal{O}_{\mathfrak{a}}$; G_t still concerns the served inputs.

Proposition 17 (Arrived slow labels)**.** *Assume the fixed affine model of Lemma 1, a fixed reference and positive working weights, and condition on a nested sequence of arrival masks and their inputs. Put* $H_{\mathfrak{a}} = \sum_{(i,d)\in\mathcal{O}_{\mathfrak{a}}} \varpi_{id} z_d(x_i) z_d(x_i)^\top$ *with the working weights* ϖ_{id}*. Then* $H_{\mathfrak{a}'} \succeq H_{\mathfrak{a}}$ *and* $\ker H_{\mathfrak{a}'} \subseteq \ker H_{\mathfrak{a}}$ *for* $\mathfrak{a}' \geq \mathfrak{a}$*, and task* t *is identified on the observed design exactly when* $\ker H_{\mathfrak{a}} \subseteq \ker G_t$*; if labels of* t *arrive with positive probability on a design whose kernel lies in* $\ker G_t$*, its own labels eventually cover it, and it adapts. For independent Gaussian noise of variances* $1/\varpi_{id}$*, a fixed size* v *and* $U_{\mathfrak{a}} = \ker H_{\mathfrak{a}}$*, the predicted gain of the projected update does not decrease with* $\mathfrak{a}$*. If all* $H_{\mathfrak{a}}$ *commute, their eigenvalues do not decrease, and a direction that leaves the thresholded* U *of Section 4 does not return.*

Proof. Each arrived pair adds a positive semidefinite outer product, which gives the order and the kernel inclusion, and identification is Lemma 1 with $H_{\mathfrak{a}}$ in place of G_D; once $\ker H_{\mathfrak{a}} \subseteq \ker G_t$, $P_t = 0$. For the predicted gain, the law $\mathcal{N}(0, vI)$ of the whole change belongs to Γ_v at every level. The projected update at level $\mathfrak{a}$ is computed from labels that level $\mathfrak{a}'$ also has and has no weight on $U_{\mathfrak{a}} \supseteq U_{\mathfrak{a}'}$, so it meets the condition of Lemma 2 at level $\mathfrak{a}'$; by Theorem 1(i) there, its Bayes regret under $\mathcal{N}(0, vI)$ is at least that of the projected update at level $\mathfrak{a}'$, and by Theorem 1(ii) the two Bayes regrets are minus the two predicted gains. With commuting $H_{\mathfrak{a}}$ the eigenvectors are shared and each eigenvalue grows with $\mathfrak{a}$. □

Own labels read what the task depends on. A label of task t at a served input x with noise variance σ^2 adds $z_t(x)z_t(x)^\top/\sigma^2$ to $H_{\mathfrak{a}}$, whose expectation over the served inputs is G_t/σ^2: each own label reads every direction in exactly the proportion that t depends on it, whereas an early label reads it in proportion to its fast task's readout. In the running example, in the coordinates of Appendix C.3, one purchase label adds about 34 prior units (v times the information) on intent and one weak-reader label about .02, so a waiting purchase is read along intent far faster by its own labels than by more early labels. *Label-dependent arrival*, as when positives are recorded early and negatives only at a horizon, multiplies an arrived label's odds by $\Pr(\text{arrived} \mid y = 1, x, e)/\Pr(\text{arrived} \mid y = 0, x, e)$ and biases the arrived-only logistic score.

F.3 Selection and fallback

How does validation choose the configuration and the fallback? On the latest window whose slow labels have matured, FLAG is rebuilt for each configuration, construction, width q and prior scale τ, within the tuning budget of Section 5.1, and the configuration with the lowest mean slow-task loss is chosen, ties going to a smaller τ and then a smaller width. Then, for each slow task, its loss under that configuration and under the frozen model are computed, with the standard error of FLAG's loss over the window's unit clusters, and the task falls back to the frozen model (OFF) if FLAG's loss exceeds the frozen model's by more than one standard error, that is, if it is not within one standard error of the lower. The one-standard-error rule was chosen on development windows earlier than the served ones. The fast tasks' fine-tuned copy may be fully fine-tuned: the slow predictions stay fixed because the copy writes nothing to the parameters, states or statistics of the frozen FLAG path.

Seeds. In each seeded run an added branch, construction (ii), is trained once, on the frozen base and the matured past labels, and that single branch has its own past coordinates, anchor, size, pilot, information and updates (lines 4–18 of Algorithm 1), which Theorem 1 covers. All methods, FLAG included, report the mean over three seeds.

G How were the known-truth experiments run?

This appendix specifies the two known-truth experiments of Section 5.2, in which refits follow the shift (Appendix G.1) and waiting follows what is read (Appendix G.2), both shown in Figure 7, and reports numerical checks of the other results (Appendix G.3, Tables 9 and 10).

G.1 Refits follow the shift: the running example

Which methods follow the campaign of the running example, and how does their error change with the campaign's distance from the past weeks and with the sample size?

Table 7: The running example of Section 3.1: what each task reads, and the gains of the two past weeks.

	readout w_t				gains s_e	
factor	click	short watch	full watch	purchase	regular week	promotion week
appeal	1	.2	.3	$-.4$	1	1.2
quality	.3	1	.6	1	1.8	.6
purchase intent	0	0	1	.8	.5	1.8

Setup. The base is zero, $x \sim \mathcal{N}(0, I_3)$ and $\phi(x) = x$; the readouts and gains are those of Table 7: the regular week has gains $(1, 1.8, .5)$ and the promotion week $(1.2, .6, 1.8)$, and the campaign has $s^\star(r) = (1.1, 1.2, 1.15) + r\,u$ with $u = (.831, -.339, -.441)$, the unit normal of the weeks' plane, for $r \in \{0, .25, .5, 1, 1.5\}$. Labels add Gaussian noise with $\sigma = .1$. Each past week has $n_s = 2000$ users with all four labels; the campaign has $n = 2000$ users with the fast labels only. The fast labels are full watch (scenario A), click and short watch (B), or all three (C), and purchase is never labelled in the campaign. Every method outputs a linear purchase predictor $\hat{b}_0 + \hat{b}^\top x$, scored by its exact excess squared error $\hat{b}_0^2 + \|\hat{b} - w_{\text{purchase}} \odot s^\star\|^2$; results are geometric means over 20 runs.

Methods. *Refit* fits each past week by least squares, recovers the factors by the pencil equation 12, and corrects the mean past gains by the minimum-norm least-squares fit to the fast labels; singular directions of the design below 5% of the largest are treated as unread and stay at the past mean. *Mixing* fits two weights on the past weeks' fitted slices to the fast labels. *Regression* fits purchase on the fast labels, with an intercept, in the past weeks and applies it to the campaign's fitted fast slices. The *pooled full adapter* fits a free 3×3 matrix on the pooled heads by minimum-norm least squares. *Blind* is pooled least squares. The floors are those of Proposition 4: $\text{dist}_{G_{t^\circ}}(s^\star, L)^2 = .21\,r^2$ for mixing, and $\text{dist}_{K_\star}(w_{\text{purchase}}, \text{span}\{w_{\text{full}}\})^2$ for any combination of full watch.

Results. In scenario A (Table 8; Figure 7a) the refit stays at noise level at every distance, mixing stays just above its floor $.21\,r^2$, and regression and the adapter stay at or above the full-watch floor, because purchase is no multiple of full watch. At $r = 0$ the campaign lies in the weeks' plane and mixing, with fewer parameters, is slightly more precise than the refit.

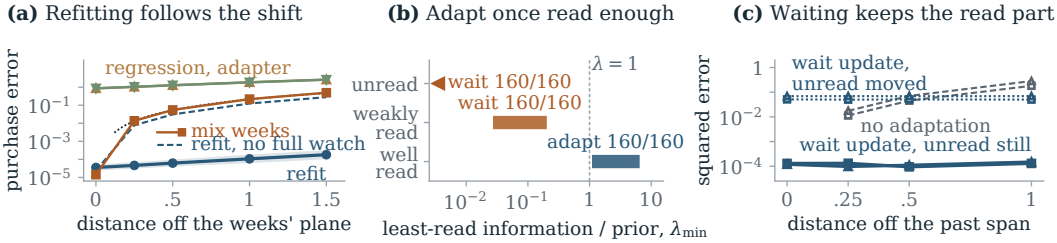


Figure 7: (a) Only the full-watch refit follows the campaign. (b) G2F: a factor read below $\lambda = 1$ is left to wait, above it is corrected. (c) Waiting errs only when the unread factor moves.

Table 8: Running example, scenario A (full watch the only fast task): purchase excess squared error. Top: 2,000 users per window, against the floors of Proposition 4. Bottom: at $r = 1$, as the users per window grow in the past weeks and the campaign together.

r	refit	mixing	regression	pooled full adapter	blind	mixing floor	full-watch floor
0	3.6×10^{-5}	1.4×10^{-5}	.861	.861	4.0×10^{-4}	0	.861
.25	4.6×10^{-5}	.014	1.05	1.04	.022	.013	1.03
.5	6.2×10^{-5}	.055	1.28	1.27	.087	.053	1.22
1	1.1×10^{-4}	.219	1.87	1.85	.349	.211	1.52
1.5	1.8×10^{-4}	.492	2.63	2.60	.786	.475	1.47

users per window	500	2,000	8,000	32,000	128,000
refit	5.9×10^{-4}	1.1×10^{-4}	6.3×10^{-5}	7.4×10^{-6}	3.6×10^{-6}
mixing	.221	.219	.219	.219	.219

Only the refit's error falls as samples grow. As the users per window grow from 500 to 128,000 at $r = 1$, the refit's error falls from 5.9×10^{-4} to 3.6×10^{-6}, a log–log slope of -0.93 (Table 8, bottom), while the slopes of the other methods lie between -0.001 and $+0.001$: they stay at their bias limits, .219 for mixing, 1.87 for regression, 1.85 for the adapter and .349 for the blind model.

Unread and fully read directions. In scenario B purchase intent is unread and the refit errs by $.124\, r^2 = (w_{\text{purchase,intent}}\, u_{\text{intent}}\, r)^2$, exactly the unread change of equation 6 with the purchase-intent gain held at its past mean; mixing, fitted to click and short watch, errs by 3.28 at $r = 1$. In scenario C the refit, the adapter and the regression are all near noise level, the regression at 1.8×10^{-4} to 4.9×10^{-4} because of errors-in-variables attenuation. With $\sigma = 10^{-6}$ the refit errs by at most 3×10^{-14} in scenario A and matches $.124\, r^2$ in B. Without the rank cutoff, scenario B errs by .26 to 1.24 even at this noise level: estimation leaves the unread direction weakly read, and inverting it amplifies noise, the first term of equation 6. This is why FLAG weighs information rather than rank.

More early labels help only a covered task. A sweep over the number of early labels, not shown in a figure, fixes $r = 1$ and $\sigma = .1$, gives each past week $n_s = 100{,}000$ labels per task, so that estimating the factors costs little, and varies the early labels per fast task from $n = 30$ to 30,000; errors are means over 100 independently seeded runs. With click and short watch, purchase's error stays at about .124 at every n, the unread change of Proposition 2. With full watch added to them, it falls from about 4.5×10^{-4} at $n = 30$, the label-noise term of equation 6, to about 2.2×10^{-6} at $n = 30{,}000$, roughly as $1/n$ until it reaches the noise floor left by the past weeks.

G.2 WAITING FOLLOWS WHAT IS READ: THE PROJECTED UPDATE ON SEMI-SYNTHETIC G2F DATA

Does the implemented projected update wait exactly where the early labels read too little, when that is known?

Setup. This experiment runs the projected update of Section 4 with $\tau = 1$ and the threshold $\lambda_j < 1$ for U, without the nonlinear gate, the bets or the selection of Appendix F, on real G2F inputs and a fixed linear factor-gain model in which the blind directions are known: the model fitted to G2F for the 2020 window, the shared-bottom network of Section 5.1 with per-window gains on its $k = 8$ shared factors, with its real inputs, its real pattern of missing labels and its residual noise, permuted within each block afresh in each run; a run is one seeded replicate. The true readouts are the fitted ones, except that the fast tasks and grain moisture do not read $j^\star$, the factor that the fast tasks read least; yield and test weight still read it, so coverage holds for grain moisture alone. The past gains are the fitted 2018 and 2019 gains, and the current gains add to their mean $r \in \{0, .25, .5, 1\}$ times a unit direction of the read factors orthogonal to both past gains, and $r' \in \{0, .5\}$ times the unread factor, eight shifts in all. Each instance thus has 160 shift–run pairs (eight shifts, 20 runs) and 480 outcomes (three harvest traits). In two further instances, *weakly read* and *read*, the fast tasks read $j^\star$ through their fitted loadings, scaled down until its nominal information, before cluster adjustment, is .25 or 8 times the prior precision. Coverage then holds for every task, but in the weakly read

instance yield and test weight are meant to wait on $j^\star$, since it carries less information than the prior; there their outcomes are scored against the threshold $\lambda_j \geq 1$ that defines U (Section 4), not against coverage. The prior comes from the two past gains, with environments as clusters.

Results. In the blind instance all 480 outcomes agree with the exact coverage check: grain moisture does not read the blind factor, so it is covered and corrected in full ($P_t = 0$) in all 160 shift–run pairs, and yield and test weight read it and wait on it in all 160. Grain moisture is corrected in full in all 160 pairs of the other two instances too, and so are yield and test weight in all 160 pairs of the read instance, where the cluster-adjusted information in the least-read direction reaches the prior precision, so all 480 outcomes there agree with coverage. In the weakly read instance, where that information stays below the prior precision, yield and test weight wait on it in all 160 pairs, as the threshold requires (Figure 7b). A waiting task is served the projected update: its error stays at noise level while the unread factor holds still ($r' = 0$) and rises by its unread change when it moves, whereas not adapting, exact at $r = 0$, errs more the further the window moves (Figure 7c). No run failed an implementation check ($F \preceq H$, $Q\theta_W = 0$ and the pilot's first-order condition).

G.3 Numerical checks of the remaining results

Many simulated shifts. The simulation scores every update of Section 3 against not adapting over 2,000 simulated windows per setting, all updates sharing one draw of the early labels per window; errors are population errors on the served inputs. In the random-task setting each window is a fresh instance: six factors with $\mathbb{E}\phi\phi^\top = .6I + .4\mathbf{1}\mathbf{1}^\top$, three fast and three slow tasks with sparse random readouts, six past windows whose gains follow a random low-rank co-movement, and a current window whose change mixes that co-movement with an independent direction, 30 labels per fast task and $\sigma = .1$. Fine-tuning trains the last layer to convergence with heads $\bar{s} \odot w_t$; the surrogate regresses slow on fast changes over the past windows; the mixture fits unconstrained weights to the early labels. The share of slow-task windows worse than not adapting is 0.9% for the projected update, whose losses are all at label-noise level, 4% for θ_W and for θ_A, 15% for the mixture, 25% for fine-tuning and 31% for the surrogate on changes (1.5, 11, 11, 41, 28 and 34% on uncovered tasks). Other updates can have a lower mean regret in some settings, at the price of harming more often. In the running example with click and short watch, where their readouts span purchase's read part, fine-tuning coincides with the refit and harms 2.5% of windows; with correlated factors it harms 14 to 24%.

Stress tests. The same random-task setting, 6,000 slow-task windows, varies FLAG's constants and breaks its assumptions. *Constants.* There every direction is read either strongly or not at all (the smallest positive λ has median 35), so thresholds from $\frac{1}{4}$ to 4 and a size v halved or doubled leave the projected update's share of losing windows at 0.9%. With small leak weights, up to $.05$, added to every readout (median smallest λ of $.86$), raising the threshold from $\frac{1}{4}$ to 4 lowers that share from 3.1 to 1.6% at a smaller median gain, and doubling v raises it from 2.1 to 3.2%, as Proposition 10(i) predicts. The bootstrap size of Section 4 is a median $.44$ of $\operatorname{tr}\Sigma/q$, and about 2.5 times below the realized size of the change. *Covariate shift.* When 10, 30 or 60% of the served inputs come from a group with its own second moment that never arrives early, the projected update with G_t on the served inputs loses in 0.9% of windows in each case, and with G_t on the arrived inputs in 1.1, 1.9 and 4.1%; fine-tuning loses in 25%. *(A0) violations.* A drift of the slow task's readout on the factors it reads, carrying 10, 30 or 60% of its change's variance, makes the projected update lose to not adapting in 10, 23 and 33% of windows, against 29, 35 and 40% for fine-tuning; a factor only the slow task reads makes it lose in 3.2, 9.2 and 24%, against 26, 28 and 34%. In every case its mean error stays below not adapting's, by 73 to 31% with the drift and 74 to 33% with the new factor, against 23 to 10% and 24 to 11% for fine-tuning. Both violations change the slow task's labels only, so no early label can detect them. Do the remaining results hold in small instances where everything is known? Tables 9 and 10 check the remaining results in small controlled instances: the first checks identification, floors and the pooled basis, the second the rates of the stability and estimated-factor results and the coverage check of Proposition 6. Unless stated otherwise the checks use a two-factor variant of the running example: click, short watch, full watch and purchase read (appeal, quality) as $(1, 0)$, $(0.9, 0)$, $(0.8, 0.6)$ and $(-0.4, 1)$, the promotion week has gains $(2.0, 0.5)$ and the regular week $(0.6, 1.8)$, $\sigma = 0.1$, and the input law is the same in every window. Errors are geometric means over seeded runs, and pencil angles, with $T = 2$, medians over 200 trials

Table 9: Numerical checks of identification, floors and the pooled basis. Columns: the result checked; the quantity measured; the cases over which it is measured; what the result predicts; the measured value, or its smallest and largest value over the cases.

result	check	cases	predicted	measured
Lemmas 4(i), 5	pooled training's angle	all cells	not identified	25–32°
	pooled training's angle, factors $\tanh(2x_j)$	all cells	not identified	48–54°
	the gain model's angle, factors $\tanh(2x_j)$	all cells	identified	0.55–1.4°
Lemma 1	purchase error of gains fitted on full watch	campaigns	noise level	4–5×10^{-5}
	the same, blind model	campaigns	above noise	.005 to 1.17
	the same, pooled diagonal gains	campaigns	above noise	.11 to 21.9
	the same, pooled 2×2 adapter	campaigns	above noise	.008 to 1.36
	the same, pooled 2×2 adapter with click added	campaigns	noise level	.0002–.0003
Proposition 4	refit, $\lvert D\rvert \in \{1, 2, 3, 5\}$ fast tasks	four $\lvert D\rvert$	noise level	4.0×10^{-5}–1.5×10^{-4}
	pooled 3×3 adapter, $\lvert D\rvert \le 2$: heads cannot span $\mathbb{R}^3$	each $\lvert D\rvert$	above noise	1.40; 1.46
	pooled 3×3 adapter, $\lvert D\rvert = 3, 5$: heads span $\mathbb{R}^3$	each $\lvert D\rvert$	noise level	.003; .0001
Proposition 15(b)	fast residual on the rotated basis	runs	noise variance	.0098–.0104
	slow error on the rotated basis	runs	5	4.92–5.00
	slow error on the aligned basis	—	noise level	below 10^{-4}

Criteria. Noise level: the result predicts no bias, only estimation error from label noise; above noise: a bias that more labels do not remove; noise variance: $\sigma^2 = .01$; identified or not: the angle error falls toward zero or stays away from it. *Units.* Errors are excess squared errors of purchase (of the slow task when $k = 3$); angles are in degrees.

Table 10: Numerical checks of rates and of the coverage check. Columns: the result checked; the quantity measured; the cases; what the result predicts; the measured value, or its range over the cases. With two factors the gap of Theorem 2 is $h = \|S_{:1}\| \, \|S_{:2}\| \, \vartheta / |s_{21} s_{22}|$.

result	check	cases	predicted	measured
Lemma 5	pencil's median angle error, $n_s = 1000$	$\vartheta = .05$ to $.4$	$\propto 1/(\vartheta\sqrt{n_s})$	2.49° at .05, 0.30° at .4
	the same, $n_s = 4000$	$\vartheta = .05$ to $.4$	$\propto 1/(\vartheta\sqrt{n_s})$	1.18° at .05, 0.15° at .4
Lemmas 4(i), 5	trained networks' angle error, slope in $\vartheta \le 0.2$	three n_s	-1	-1.22 to -0.98
Theorem 2	error slope in ϑ, $s_1^\star \neq s_2^\star$	contrasts, n_s	-2	-1.66 to -2.08
	error slope in n_s, $s_1^\star \neq s_2^\star$	contrasts, ϑ	-1	-1.00 to -1.05
	error slope in ϑ, $s_1^\star = s_2^\star$	n_s	0	-0.10 to -0.02
	error slope in n_s, $s_1^\star = s_2^\star$	ϑ	-1	-1.02 to -1.07
	error slope in n, $\vartheta = 0.8$, $n_s = 20{,}000$	$n = 50$ to 1600	-1	-0.98
	the same with click added	$n = 50$ to 1600	-1	-0.95
Proposition 6	coverage check, fast tasks reading only appeal	runs flagged	fires	20 of 20
	coverage check otherwise	runs flagged	silent	0 of 40
	coverage check, change confined to purchase	$k = 3$	silent	unflagged
	the same change: purchase error	$k = 3$	above noise	1.01

Criteria. Slopes are log–log slopes of the error; fires or silent: the coverage check flags the change or not; above noise: a bias that more labels do not remove. *Units.* Errors are excess squared errors of purchase (of the slow task when $k = 3$); angles are in degrees; counts are runs.

per cell; the checks of Lemma 1 use four campaign windows, one near the past average and three farther away, and those of Proposition 4 use $k = 3$, $T = 8$ and three past windows. In the check of Theorem 2, a closed-form pipeline fits least squares in each past window, recovers the readouts by the pencil, fits the current gains on full watch alone ($n = 20{,}000$) and predicts purchase, with campaign gains centred at 1.2 with contrast $s_1^\star - s_2^\star \in \{0, .8, 1.6\}$, over $\vartheta \in \{.05, .1, .2, .4\}$ and $n_s \in \{500, 2000, 8000\}$. With two factors, this angle and the ratio gap of Theorem 2 are related by $h = |\varrho_1 - \varrho_2| = \|S_{:1}\| \, \|S_{:2}\| \, \vartheta / |s_{21} s_{22}|$ (Appendix D.3), and every current gain lies in the span of the past gains, but its coefficient on the ratio direction grows as $(s_1^\star - s_2^\star)/\vartheta$, so the first term of equation 14 predicts an error proportional to $(s_1^\star - s_2^\star)^2/(\vartheta^2 n_s)$ plus a reference term free of ϑ.

H How were the deployment experiments run?

How were the real-data experiments of Sections 5.1–5.3 run? This appendix details the replayed deployment of Section 5.1, with its per-window and per-task results (Appendix H.3) and its ablations and the split of each method's regret behind Section 5.2 (Appendix H.4); it sets up the triage queue of Section 5.3 (Appendix H.5); and it computes the drift statistics quoted in Section 1 (Appendix H.6).

Table 11: The four datasets: unit (bootstrap cluster), window length (number of windows), served windows, and fast and slow tasks.

dataset	unit (cluster)	window	served	fast tasks	slow tasks
KuaiRand-Pure	exposure (user)	2 days (15)	windows 4–15	click, long view	like, follow, comment, forward, hate
RetailRocket	(visitor, item) pair from its first view (visitor)	week (18)	weeks 7–18	view again or add to cart within 24 hours	purchase within 14 days
CrossValidated	question (asker)	quarter (58)	2021Q2–2024Q1	answer within 1 hour, within 1 day	accepted answer within 7 days, upvote within 30 days
G2F maize	(environment, hybrid) plot mean (environment)	year (8)	2016–2021	pollen shed, silking, plant height, ear height	yield, grain moisture, test weight

Table 12: KuaiRand-Pure served windows: split time, exposures in the early and late parts, and late-part positives of click, long view, like, follow, comment, forward and hate.

window	split	early	late	late-part positives
Apr 15–16	Apr 16, 01:00	61,612	58,184	30,006 / 21,937 / 1,423 / 71 / 162 / 59 / 48
Apr 17–18	Apr 17, 20:00	35,473	33,110	15,187 / 10,929 / 561 / 42 / 63 / 33 / 15
Apr 19–20	Apr 20, 01:00	21,302	19,992	8,695 / 6,233 / 342 / 26 / 59 / 21 / 8
Apr 21–22	Apr 22, 02:00	21,333	20,971	9,444 / 6,665 / 437 / 25 / 45 / 10 / 14
Apr 23–24	Apr 23, 22:00	23,693	21,192	9,289 / 6,428 / 368 / 28 / 50 / 21 / 15
Apr 25–26	Apr 26, 00:00	14,911	14,530	6,400 / 4,525 / 218 / 17 / 38 / 7 / 10
Apr 27–28	Apr 28, 00:00	14,328	13,972	5,861 / 4,050 / 232 / 17 / 30 / 16 / 6
Apr 29–30	Apr 30, 08:00	18,927	18,116	8,358 / 5,860 / 373 / 25 / 41 / 21 / 7
May 1–2	May 2, 00:00	20,790	19,554	8,874 / 6,275 / 369 / 23 / 57 / 14 / 29
May 3–4	May 4, 00:00	18,943	17,600	7,828 / 5,525 / 286 / 29 / 51 / 13 / 17
May 5–6	May 6, 01:00	13,934	13,290	5,660 / 3,939 / 208 / 13 / 32 / 12 / 13
May 7–8	May 8, 02:00	14,758	14,676	6,653 / 4,720 / 259 / 24 / 38 / 16 / 10

H.1 Data, splits and label timing

Which data does the replay use, how is each served window split, and when does each label arrive? Windows are numbered from 1 in time order within each dataset, so KuaiRand-Pure serves windows 4–15 of 15 and RetailRocket weeks 7–18 of 18.

Datasets and inputs. Table 11 lists the four datasets. Model inputs use only information available before each unit's anchor time: KuaiRand's released features; for RetailRocket, hashed visitor and item buckets, the anchor's hour and weekday, the item's views and add-to-carts over the past 7 days and purchases over the past 30, and the visitor's earlier views; for CrossValidated, a hashed asker bucket, the hour and weekday, title and body lengths, whether the question contains code, its number of tags, the asker's earlier questions and the 100 most common tags of the first eight quarters; for G2F, the hybrid, its two parents and the trial location, the planting day, latitude and longitude. CrossValidated's outcome fields, such as the score or the accepted answer, are not used as inputs. Test-weight labels outside $[550, 900]$ kg/m^3 are set missing, and an environment with more than half of them outside loses all of them (1,974 of 70,870 labels; four environments). G2F's traits are standardized by the past windows' mean and standard deviation. CrossValidated's texts are the dump's final versions and may include later edits.

Splits. In the three logged datasets each served window is split in time, at the first time that leaves at least half of its units before it, without splitting a time point: by the hour for KuaiRand (Table 12), and at anchor times for RetailRocket and CrossValidated. The rarest KuaiRand slow tasks have 6 to 71 late-part positives, so their intervals are wide. G2F is not split in time: the in-season traits of the served year are known at serving and its harvest traits are not; both are measured on the same plots, and only the harvest traits are scored.

Label maturity. In RetailRocket (horizons of 1, 1 and 14 days) and CrossValidated (1 hour, 1 day, 8 and 31 days, since votes carry only a date and count at the end of their day), a past label is used only if its anchor time plus its horizon precedes the training time, and an early fast label only if it precedes the split; an immature negative is never read as zero. At each served window's training time about 190,000 to 250,000 of RetailRocket's past 14-day purchase labels are still immature

and masked, and about 60 to 70% of its early fast labels have arrived by the split. Every slow label of a served window is withheld, even CrossValidated upvotes that would arrive within the quarter. CrossValidated records when its labels arrive; elsewhere the delays follow the order of events: KuaiRand's likes and follows are held until the window ends, a RetailRocket purchase is known 14 days after the view, and G2F flowering is recorded about 66 days after planting and harvest at about 162. KuaiRand and G2F record no per-unit arrival, so their past labels count as matured and their early fast labels, including G2F's undated heights, as arrived.

H.2 Protocol, baselines and tuning

How is every method trained, updated and scored, and how are its hyperparameters chosen?

Protocol and metrics. Before each served window, every method is retrained on all earlier windows, with the masks above. Every method may update online on the arrived labels and then predicts the late part, or in G2F the slow traits. Normalized entropy is the log loss divided by the entropy of the training base rate (He et al., 2014); G2F's metric is the squared error divided by the past variance, with Pearson's correlation as a secondary metric. Means and standard errors are over independently seeded training runs; the worst window is the worst served window of the seed-averaged value. Intervals come from 1,000 bootstrap draws per window over clusters (users for KuaiRand and RetailRocket, askers for CrossValidated, environments for G2F), shared by all methods and seeds, and differences use the paired bootstrap.

Baselines. The eight baselines of Table 1 use three tuned architectures: a shared-bottom network (an encoder, a k-dimensional shared layer and linear heads), MMoE (Ma et al., 2018a), and this network with window-embedding gates after PEPNet (Chang et al., 2023). *Past windows only* trains the shared-bottom network on past windows, and *GroupDRO* trains it with GroupDRO over past windows (Sagawa et al., 2020). *Shared-bottom, fine-tuned* fine-tunes it on the early fast labels, and *MMoE, fine-tuned* does the same for MMoE. The *window-embedding gate* fits the gated network's new embedding alone on those labels. The *pooled full adapter* fits a $k \times k$ matrix on the pooled shared layer on those labels. The *past-window combination* weights the gated network's logits under the past embeddings, fitted on those labels. The *fast-to-slow regression* regresses each slow task on the fast logits, fitted on past windows and fed at serving by the fine-tuned network, with scikit-learn's default penalty in every served window: logistic with L2 inverse strength 1 on the binary datasets, ridge with penalty 1 on standardized inputs on G2F; an almost unpenalized fit is unstable where the fast logits are nearly collinear (RetailRocket's all-task normalized entropy 2.01, against $.898$ with the default). The pooled full adapter, the past-window combination and the fast-to-slow regression adapt complete predictions, the comparator classes of Appendix A.3. Delayed-feedback corrections (Chapelle, 2014; Yang et al., 2021) reweight or model a slow task's labels in transit; we withhold all slow labels of a served window so that every method faces the same decision, and delayed-feedback corrections are complementary to FLAG.

Tuning. Hyperparameters follow DomainBed (Gulrajani & Lopez-Paz, 2021): 20 random configurations per dataset, architecture and served window, the best on the most recent window whose late-part labels have all matured by the served window's training time, retrained on all past windows. This validation window is the last past window for KuaiRand and G2F, three windows back for RetailRocket and two back for CrossValidated, since the window just before would expose slow labels that have not matured; earlier windows train, with the same masks. The search uses neither the served window's slow labels nor its late part. Serving starts once two past windows exist; FLAG's prior needs two past windows before its validation window, so in G2F's first served year it serves a fallback. FLAG's base is MMoE on KuaiRand, RetailRocket and G2F and the gated network on CrossValidated.

H.3 Per-window and per-task results

KuaiRand-Pure. KuaiRand-Pure's slow tasks are five rare engagements whose labels we hold back until each two-day window ends, so the delay is emulated, and their base rates move between windows: every method scores near the entropy of predicting the training base rate, and in the window whose base rates move most every method is above it. What separates the methods there

Table 13: Slow-task error over the served windows (normalized entropy; G2F: squared error over past variance): mean±s.e. over seeds, the worst served window of the seed-averaged value, and the 95% cluster-bootstrap interval of the mean's difference from past windows only. FLAG's row gives point values; in G2F's first served year it serves a fallback.

method	mean	worst window	interval
KuaiRand-Pure: 12 two-day windows, Apr 15–May 8			
past windows only	1.0236±.0033	1.1989	—
GroupDRO	1.0474±.0036	1.2536	[+.0156, +.0323]
shared-bottom, fine-tuned	1.0222±.0052	1.1960	[−.0036, +.0009]
MMoE, fine-tuned	1.0147±.0006	1.2117	[−.0207, +.0017]
window-embedding gate	1.1342±.0086	1.3507	[+.0934, +.1323]
pooled full adapter	1.0201±.0017	1.2077	[−.0077, +.0002]
past-window combination	1.1296±.0064	1.3425	[+.0881, +.1283]
fast-to-slow regression	1.0959±.0123	1.3115	[+.0528, +.0968]
FLAG	.9784	1.1468	—
RetailRocket: weeks 7–18			
past windows only	.8290±.0007	.9619	—
GroupDRO	.8320±.0020	.9549	[−.0006, +.0071]
shared-bottom, fine-tuned	.8303±.0023	.9641	[−.0026, +.0060]
MMoE, fine-tuned	.8290±.0035	.9684	[−.0060, +.0067]
window-embedding gate	.8304±.0017	.9661	[−.0024, +.0059]
pooled full adapter	.8311±.0016	.9684	[+.0004, +.0039]
past-window combination	.8306±.0016	.9605	[−.0022, +.0064]
fast-to-slow regression	.8418±.0012	.9645	[+.0076, +.0185]
FLAG	.8013	.9107	—
CrossValidated: quarters 2021Q2–2024Q1			
past windows only	.9656±.0042	.9930	—
GroupDRO	.9739±.0035	1.0062	[+.0055, +.0113]
shared-bottom, fine-tuned	.9711±.0059	1.0066	[+.0040, +.0070]
MMoE, fine-tuned	.9738±.0063	1.0383	[+.0061, +.0105]
window-embedding gate	.9300±.0092	.9792	[−.0460, −.0250]
pooled full adapter	.9757±.0118	1.0135	[+.0069, +.0131]
past-window combination	1.2239±.0285	1.9349	[+.2354, +.2820]
fast-to-slow regression	.9709±.0012	1.0149	[+.0002, +.0105]
FLAG	.8852	.9114	—
G2F maize: years 2016–2021			
past windows only	.8792±.0134	1.6031	—
GroupDRO	.9012±.0117	1.6683	[−.0272, +.0744]
shared-bottom, fine-tuned	.9413±.0088	1.8339	[+.0072, +.1251]
MMoE, fine-tuned	.9380±.0515	1.6739	[−.0227, +.1630]
window-embedding gate	.9675±.0385	1.8670	[+.0169, +.1553]
pooled full adapter	1.0877±.0219	1.9623	[+.1103, +.3050]
past-window combination	.9781±.0219	1.6872	[+.0264, +.1668]
fast-to-slow regression	.9407±.0048	1.7187	[−.0017, +.1257]
FLAG	.8421	1.5098	—

is mostly calibration to the moving base rates rather than what the early labels identify; FLAG's mean, .978, is the only one below that entropy. Table 13 gives every method's mean and worst window. Averages over windows and tasks can hide a failure in one window or one slow task, so this subsection breaks the results of Table 1 down. Table 13 gives every method's slow-task error over the served windows and in the worst of them, FLAG's as point values, and is the source of the *worst* column of Table 1; Tables 14–16 give every method's per-task results. FLAG serves the fast tasks from its base's adapted copy, so its fast-task values are those of MMoE, fine-tuned, and on CrossValidated, whose base is the gated network, those of the window-embedding gate. The primary metric is normalized entropy because log loss is a proper scoring rule, and dividing it by the entropy of the training base rate puts tasks with very different base rates on one scale (He et al., 2014). As a secondary metric that measures ranking alone, Table 17 gives each binary slow task's AUC, and Table 16 G2F's Pearson correlations, so that the comparisons can be checked not to rest on calibration alone.

H.4 ABLATIONS, AND WHERE EACH METHOD'S REGRET SITS

What does each part of FLAG contribute, and where do the baselines lose? The two rows under FLAG in Table 1 change one part of the configuration that FLAG selected in a served window, keeping its construction, prior scale, pilot, information, anchor and U, without reselecting hyperparameters, and act only where FLAG is served. *Step 1* serves $\bar{s} + \theta_W$, the refit in FLAG's coordinates with the past's

Table 14: KuaiRand-Pure, per-task log loss on the late parts (lower is better), mean over served windows and worst window; best baseline underlined, FLAG bold where better. Click and long view are fast.

	click		long view		like		follow		comment		forward		hate	
method	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst
past windows only	.5786	.5885	.5271	.5356	.0650	.0696	.0093	.0097	.0161	.0184	.0071	.0089	.0055	.0085
GroupDRO	.5808	.5895	.5306	.5401	.0666	.0717	.0094	.0098	.0166	.0191	.0073	.0093	.0057	.0090
shared-bottom, fine-tuned	.5745	.5810	.5243	.5314	.0652	.0697	.0093	.0096	.0162	.0185	.0071	.0089	.0055	.0084
MMoE, fine-tuned	.5768	.5853	.5264	.5360	.0658	.0704	.0090	.0093	.0162	.0186	.0071	.0090	.0055	.0085
window-embedding gate	.5874	.5971	.5369	.5455	.0777	.0901	.0102	.0107	.0169	.0194	.0076	.0092	.0063	.0102
pooled full adapter	.5752	.5819	.5248	.5325	.0650	.0697	.0093	.0097	.0161	.0184	.0071	.0089	.0055	.0085
past-window combination	.5895	.5971	.5392	.5454	.0783	.0901	.0102	.0107	.0169	.0194	.0074	.0092	.0063	.0102
fast-to-slow regression	.5835	.5915	.5327	.5415	.0779	.0905	.0098	.0107	.0168	.0196	.0073	.0094	.0060	.0090
FLAG	.5768	.5853	.5264	.5360	**.0631**	**.0690**	**.0088**	**.0092**	**.0156**	**.0181**	**.0068**	**.0087**	**.0053**	**.0082**

Table 15: RetailRocket (first three tasks) and CrossValidated (last four), per-task log loss on the late parts (lower is better), mean over served windows and worst window; best baseline underlined, FLAG bold where better. The first two tasks of each dataset are fast.

	RetailRocket						CrossValidated							
	view again 1 d		cart 1 d		buy 14 d		answer 1 h		answer 1 d		accepted 7 d		upvote 30 d	
method	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst
past windows only	.3610	.3701	.0893	.0923	.0418	.0475	.3797	.3938	.6721	.6870	.4485	.4587	.7367	.7803
GroupDRO	.3610	.3701	.0894	.0925	.0419	.0477	.3800	.3976	.6785	.6923	.4497	.4701	.7461	.7894
shared-bottom, fine-tuned	.3612	.3701	.0895	.0923	.0418	.0474	.3658	.3932	.6412	.6555	.4450	.4568	.7479	.8100
MMoE, fine-tuned	.3607	.3697	.0888	.0913	.0418	.0481	.3663	.3925	.6420	.6568	.4448	.4628	.7518	.7954
window-embedding gate	.3609	.3698	.0893	.0923	.0418	.0478	.3662	.3925	.6378	.6488	.4330	.4509	.7084	.7385
pooled full adapter	.3611	.3699	.0895	.0922	.0419	.0476	.3675	.4021	.6380	.6581	.4517	.4673	.7461	.7959
past-window combination	.3610	.3698	.0894	.0922	.0419	.0478	.3656	.3927	.6378	.6487	.4377	.4652	1.0853	1.9705
fast-to-slow regression	.3613	.3701	.0898	.0923	.0424	.0476	.3679	.3932	.6454	.6555	.4447	.4602	.7481	.8094
FLAG	.3607	.3697	.0888	.0913	**.0404**	**.0460**	.3662	.3925	.6378	.6488	**.4145**	**.4312**	**.6715**	**.7005**

size and the early labels' information but without the projection; it leaves the unread coordinates at the anchor, and by equation 9 its gap to FLAG is the bet that the anchor places. Together with the window-embedding gate, a baseline, it forms a ladder: gate, then step 1, then FLAG. *Always bet* serves $\bar{s} + \delta_t^* + P_t\theta_A$, the projected update plus the bet on the past windows' co-movement, on every served slow task. Table 18 gives the differences with intervals, and repeats the ladder with the construction fixed to the gate, construction (iii), in every window. Always betting on the past's co-movement keeps a fifth to under a third of the gain, and its worst window is higher than FLAG's on every dataset.

Testing (A0) on matured windows. For each pair of consecutive matured windows, the frozen model with the earlier window's coordinates is refitted twice to the later window's labels of all tasks: once in its window coordinates only, and once with every weight free. The share of the squared logit error of the slow tasks that the coordinates remove, relative to the free refit, measures how much of the change moving gains explain: .81 on KuaiRand-Pure, .93 on RetailRocket, .86 on CrossValidated and .88 on G2F, the last on the relative squared error. FLAG is not told this share; it bounds what any model of moving gains can capture. Three controls sharpen it. Refitting only each task's intercept removes .28 to .41 of the same error, and a generic adapter with as many free parameters as the window coordinates, a rank-one update of the last shared layer, removes .52 to .67, so the gains' structure, not their number, carries the fit. And, out of sample, coordinates fitted to the later window's labels of all tasks but one slow task, which is how FLAG uses (A0), remove .71 to .84 of that held-out task's error, against .38 to .55 for the same adapter; a gap between this share and the in-sample one measures how much of the task's change the other tasks do not read.

Seeds and scale. FLAG is run on the baselines' three seeds; the standard error of its mean slow-task error over seeds is at most .003 on every dataset, and the squared error of the logit, the scale of Theorem 1, ranks FLAG, step 1 and the best baseline in the order of Table 1 on every dataset.

Table 16: G2F, harvest traits: squared error over past variance and Pearson correlation, mean over served years and worst year; best baseline underlined, FLAG bold where better. In-season traits are known at serving and not scored.

	squared error (lower is better)						Pearson r (higher is better)					
	yield		moisture		test weight		yield		moisture		test weight	
method	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst
past windows only	1.0209	1.3820	.6387	1.4080	.9781	1.9156	.3588	.3385	.7290	.7183	.2189	.1257
GroupDRO	1.0320	1.5005	.7400	1.7038	.9316	1.7459	.3176	.2555	.7029	.6362	.1309	−.0610
shared-bottom, fine-tuned	.9518	1.2815	.7771	1.7565	1.0950	2.3243	.3974	.3819	.6994	.6665	.1623	−.0446
MMoE, fine-tuned	1.0228	1.4552	.6472	1.4798	1.1441	2.0027	.4116	.3888	.7315	.6585	.1174	.0842
window-embedding gate	1.0018	1.5836	.8510	2.0409	1.0498	1.8667	.3695	.2521	.6932	.5155	.1777	.0699
pooled full adapter	.9916	1.4880	1.2162	2.3220	1.0553	1.9167	.3172	.2227	.5461	.4279	.0663	−.0704
past-window combination	.8949	1.3300	.9845	1.7365	1.0550	1.9204	.3996	.2604	.5524	.2011	.1029	−.0618
fast-to-slow regression	.9976	1.6055	.8896	1.5849	.9350	1.8499	.3683	.3255	.5482	.2860	.1972	.1079
FLAG	.9627	**1.2634**	**.6125**	**1.3689**	.9511	1.8533	**.4178**	**.3937**	**.7364**	**.7226**	**.2247**	**.1306**

Table 17: Binary slow tasks, AUC on the late parts (higher is better), mean over served windows and worst window; best baseline underlined, FLAG bold where better. G2F's Pearson correlations are in Table 16.

	KuaiRand-Pure										RetailRocket		CrossValidated			
	like		follow		comment		forward		hate		buy 14 d		accepted 7 d		upvote 30 d	
method	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst	mean	worst
past windows only	.881	.874	.755	.656	.756	.720	.748	.716	.733	.653	.791	.782	.597	.594	.573	.564
GroupDRO	.870	.844	.755	.645	.724	.698	.726	.701	.715	.648	.789	.782	.578	.562	.560	.546
shared-bottom, fine-tuned	.881	.874	.756	.658	.755	.720	.749	.719	.731	.647	.790	.782	.591	.585	.572	.559
MMoE, fine-tuned	.879	.870	.797	.720	.764	.725	.752	.730	.731	.672	.797	.787	.598	.591	.568	.558
window-embedding gate	.795	.766	.684	.657	.695	.649	.695	.616	.620	.533	.794	.782	.592	.575	.574	.555
pooled full adapter	.882	.874	.760	.656	.762	.741	.747	.709	.747	.703	.791	.782	.572	.510	.567	.531
past-window combination	.789	.757	.680	.657	.696	.653	.700	.637	.630	.533	.794	.782	.584	.552	.553	.488
fast-to-slow regression	.818	.758	.711	.632	.681	.661	.664	.594	.639	.570	.786	.780	.534	.526	.534	.523
FLAG	**.885**	**.877**	**.806**	**.728**	**.771**	**.744**	**.758**	**.735**	**.751**	**.706**	**.801**	**.791**	**.605**	**.601**	**.584**	**.573**

The split of a method's regret. Once a served window's slow labels have matured, its change θ is estimated by fitting the window coordinates of FLAG's selected construction to all of the window's matured labels, fast and slow, and subtracting the anchor. A baseline changes a slow task's predictions, not the coordinates, so its change relative to not adapting, Δf_t, is mapped to an equivalent update δ by least squares of $\Delta f_t(x)$ on $z_t(x)$ over the served inputs, on the logit scale for a binary task. The update is split into $\delta_V = (I - P_t)\delta$ and $\delta_B = P_t\delta$, with the U and P_t of FLAG's projected update in that window, and the method's gap to the projected update, $\mathrm{Reg}_t(\delta;\theta) - \mathrm{Reg}_t(\delta_t^*;\theta)$, into the transport error and the bet of equation 9, both in the metric G_t and for the realized draw of the labels. The *remainder* is what the model does not represent: the method's realized loss difference from the projected update on the task's matured labels minus the model's gap, which includes the residual of the least-squares map. Summing each part over the 114 slow-task windows and dividing by the summed gap gives the shares of Section 5.2, with the two fine-tuned baselines pooled as fine-tuning and the past-window combination and the fast-to-slow regression pooled as reuse. FLAG's realized gain in a window is minus the realized regret of its projected update, which its predicted gain, computed before any slow label, forecasts.

Served windows. Table 20 counts, for each dataset, the slow-task windows, the (served window, slow task) pairs, that the check of Section 4 served with the frozen model or with FLAG, and the windows in which FLAG lost to not adapting; a pair's outcome is the one of the majority of seeds. In all 114, $\zeta_t \leq 1$ and Δ_t is under 1% of the predicted gain; the fallback fired once, in G2F's first served year, and without it FLAG is still worse than not adapting in none of the 89 windows.

Table 18: Ablations of Table 1 minus FLAG: slow-task normalized entropy (G2F: squared error over past variance), mean over served windows with its 95% paired cluster-bootstrap interval; positive values favour FLAG. The last two rows fix the construction to the gate.

ablation	KuaiRand-Pure	RetailRocket	CrossValidated	G2F maize
step 1 (θ_W)	$+.017$ [.011, .023]	$+.013$ [.008, .018]	$+.020$ [.014, .026]	$+.016$ [.008, .024]
always bet	$+.029$ [.021, .037]	$+.021$ [.014, .028]	$+.032$ [.024, .040]	$+.028$ [.018, .038]
gate construction: step 1	$+.037$ [.029, .045]	$+.022$ [.017, .027]	$+.034$ [.029, .039]	$+.033$ [.025, .041]
gate construction: projected	$+.028$ [.022, .034]	$+.016$ [.013, .019]	$+.025$ [.022, .028]	$+.025$ [.019, .031]

Table 19: Frozen fast route on the three datasets that score fast tasks: all-task normalized entropy, mean over served windows; lower is better.

variant	KuaiRand-Pure	RetailRocket	CrossValidated
FLAG	.9479	.8845	.8645
frozen fast route	.9560	.8890	.8960

Table 20: Slow-task windows by what the check of Section 4 served, and those in which FLAG lost to not adapting.

dataset	slow-task windows	frozen model	FLAG	lost to not adapting
KuaiRand-Pure	60	0	60	5
RetailRocket	12	0	12	2
CrossValidated	24	0	24	2
G2F maize	18	3	15	2
total	114	3	111	11

H.5 THE TRIAGE-QUEUE CASE STUDY

How is the weekly triage queue of Section 5.3 replayed? It uses the public CrossValidated dump (Stack Exchange, Inc., 2024) and the four tasks of Table 11.

Weeks, labels and inputs. The replay serves the 47 complete weeks, Monday to Monday, that start from 3 April 2023 to 19 February 2024. The early questions of a week are those posted in its first 48 hours; at the start of the week's Wednesday, the cutoff, their matured fast labels are available, 3,999 one-hour and 1,959 one-day labels from 4,063 early questions over the 47 weeks, and none of their slow labels. Every question posted after the cutoff in the same week, 8,704 in all, is scored when it is posted. The labels mature 1 hour, 1 day, 8 days and 31 days after posting, since votes carry only a date; a label enters a fit only after it has matured, and no slow label of a served week enters any method. The inputs are those a question has when it is posted: its initial title, body and tags, recovered from the post history rather than the final and possibly edited version, the hour and weekday of posting, and the asker's number of earlier questions.

Base network and methods. Every Monday a base network is trained on the preceding 104 complete weeks with the labels matured by then: a shared bottom over the inputs, one head per task and a gate after PEPNet (Chang et al., 2023) whose four-dimensional embedding is fitted per week. Every method starts from this network in each week and seed and sees the same early labels. Table 2 shows every baseline. *Frozen* predicts with the mean of the past weeks' embeddings; *fine-tuned* adapts the network to the early labels; the *fast-to-slow regression* fits, on past weeks, a logistic regression of each slow task on the fast tasks' logits and applies it to the fast logits of the window-embedding gate adapted to the early labels; and the *window-embedding gate* fits the week's embedding to the early labels alone, with a Gaussian penalty toward the past embeddings. FLAG refits the same embedding, construction (iii) of Section 4, with the prior of Section 4, serves the projected update, and prices its two candidate bets as in Appendix F.1 on the matured weeks of the base network's 104 training weeks, of which more than 40 have matured slow labels at every cutoff. *Never bet* serves the projected update alone and *always bet* adds the bet $P_t\theta_A$ in every week, as the replay ablation of that name does (Appendix H.4). The hyperparameters of every method, such as the fine-tuning rate and number of steps, the embedding penalty and FLAG's τ, are chosen on the 52 weeks of 2022. All methods, FLAG included, report the mean over three seeds.

The other baselines of Table 1. For the weekly column of Table 1, four more baselines are run in every week and seed, with hyperparameters chosen on the 52 weeks of 2022 before any served week: GroupDRO, training the base network with exponentiated-gradient weights over the 104 training weeks as groups (step .01 of $\{.01, .1\}$); an MMoE with four 64-unit experts and one gate per task on the same input encoder, fine-tuned on the early labels (rate 10^{-4}, 100 steps, of rates $\{10^{-4}, 10^{-3}\}$ and $\{25, 100\}$ steps); a pooled full adapter $h \mapsto Ah + b$ on the shared hidden layer, fitted to the early labels with penalty 10 toward the identity (of $\{.1, 1, 10\}$); and a combination of the eight latest training weeks' predictors with weights fitted to the early labels by ridge-penalized logistic regression toward uniform weights (penalty 10 of $\{.1, 1, 10\}$). The count of weeks worse than not adapting uses the rule of Table 1, a slow-task mean above frozen by more than the sum of two standard errors over the three seeds.

Delayed-feedback baselines. Votes are recorded by day, so an event dated day d is known from the start of day $d + 1$. At a time T, a question's slow outcome is *seen* if its event is known by T and falls within the task's horizon; a seen outcome is always a final positive. With e the time from the start of the posting day to T and F_t the empirical distribution, over the matured positives of the 104 training weeks, of the delay from the start of the posting day to the day the event becomes known, an unseen outcome has likelihood $1 - \mathrm{sig}(\ell_t)F_t(e)$ and a seen one $\mathrm{sig}(\ell_t)$ up to a constant (Chapelle, 2014); a matured label has $F_t = 1$. By the cutoff F_t predicts that 51% of the early questions' acceptances and 80% of their upvotes are seen, against 45% and 77% observed. Five delayed-feedback arms were run, each in 52 weeks of 2022 and 47 served weeks, three seeds: an intercept shift per slow task; the week's embedding refitted on the early fast labels and seen slow outcomes; fine-tuning on both; a base network trained with the delay-corrected likelihood on the unmatured rows of its training weeks, served as it is; and that network with the refitted embedding. The last has the lowest slow-task loss on 2022 (.977 against .979 for frozen) and is the one in Table 2.

Over the served weeks it lowers the slow-task entropy by .0090 [.0047, .0129] relative to frozen, .0043 on acceptance and .0137 on upvotes, and changes the queue by $+0.09$ [-0.00, 0.19] per 100; it is worse than frozen in 11 of the 47 weeks, on acceptance in 19 and on upvotes in 11, and its worst week, 6 November 2023, has entropy 1.033. Fine-tuning with the seen outcomes would do better on the served weeks, .958, but is worse than frozen on 2022 (.988), with a worst served week of 1.083, so the rule applied to every method does not choose it; the intercept shift is worse than frozen on both (.999 and .975). Trained with the delay-corrected likelihood, the base network alone gains .0022, and refitting its embedding on the fast labels alone gains .0021: the seen outcomes, not the training, carry most of the gain. FLAG with the seen outcomes runs on the same delay-trained network, adds each seen or unseen early outcome to its likelihood as above, and is otherwise unchanged.

Queue and scores. At the end of each scored day, the day's new questions are ranked by their predicted probability of having no accepted answer within seven days, and the top 20%, rounded up, are queued: 1,841 slots over 235 days. A queued question is caught if it has no accepted answer seven days after posting. As 84.1% of scored questions have none, the queues are compared at the same daily capacity rather than by their catch rate alone. The slow-task entropy is each slow task's log loss divided by that of its base rate over the training weeks, averaged over weeks and then over the two slow tasks. Intervals are 95% percentile intervals from 1,000 bootstrap draws that resample weeks within their quarter, with all methods paired within each draw.

Information arrival. The check of Section 5.3 reruns every method on the same scored questions and weekly base networks with two nested sets of early labels at the cutoff, the matured one-hour labels alone and the one-hour and one-day labels together, and counts, for each slow task, the weeks in which its projected update keeps weight on the blind spot under each, $P_t \neq 0$; a task's outcome in a week is the majority over the three seeds.

Fine-tuning's relation and bet. For the check of Proposition 1, the tangent kernel of the base network at its Monday weights, over all shared parameters, is evaluated between the week's early questions and its scored questions; applied to the fine-tuned network's changes on the early questions' two answer tasks, it gives the first-order prediction of the fine-tuned network's change in each scored question's upvote logit, and the reported share is that of the change's variance over the scored questions of all 47 weeks that the prediction explains. The fine-tuned network's bet on upvotes in a week is the blind-spot part of its equivalent update, mapped and split as in Appendix H.4.

H.6 THE DRIFT STATISTICS OF SECTION 2

How were the correlations, the fine-tuning results, the rotation and the count of directions quoted in Sections 1 and 2 computed?

CrossValidated correlations. In each quarter with at least 300 questions (55 of the 58), the 30-day-upvote label and the answered-within-a-day label are z-scored within the quarter, and the first is regressed by least squares on the second alone: a linear-probability fit whose slope is their correlation (Figure 3a, which also shows the answered-within-an-hour label). Intervals are ± 1.96 cluster-robust (CR1) standard errors with askers as clusters. The slope stays between $-.12$ and .09 through 2020 and reaches .39 in 2023.

Fine-tuning, task by task. Figure 1 gives, for CrossValidated, each task's change in log loss when the shared-bottom network is fine-tuned on the early labels, relative to past windows only, from the per-task values of Table 15. Both fast tasks and acceptance improve and upvotes get worse; all tasks together improve, while the slow tasks' normalized entropy rises from .966 to .971 (Table 1).

G2F correlations and rotation. On the (environment, hybrid) plot means, after the test-weight cleaning of Appendix H.1, each trait is centred within its environment and z-scored within the year, and yield is regressed on each in-season trait alone, with ± 1.96 CR1 standard errors clustered by environment. For the rotation, the angle between the leading eigenvectors of adjacent years' correlation matrices of the seven traits, minus a within-year noise floor, has a median of 15.0° over the seven pairs of adjacent years.

Directions of change. A label-only screen counts, for three of the timestamped datasets, the directions in which the task covariance of the standardized labels changes between consecutive windows: it stacks the changes, weights them by their precision, and counts the directions that carry 90% of the energy above a split-half noise floor. The count is three in G2F after the test-weight cleaning (signal-to-noise 31) and five in CrossValidated (signal-to-noise 3.5), where most of the change is in base rates. In RetailRocket, the weak-shift dataset, the change stays near the noise floor (signal-to-noise 1.9, below the screen's threshold of 3).

I WHERE ELSE DO REAL DATA SHOW THE STRUCTURE OF SECTION 3?

Do other public multi-task datasets show shifts of the kind Section 1 describes, in which the relation between tasks changes from one environment to another? This appendix measures the blind bias of Proposition 5 on three of them and fits gain models to two.

Blind bias: Beijing, Tox21, KuaiRand. On Beijing air quality (12 stations, six pollutants), Tox21 (four molecular-weight bins, twelve assays) and KuaiRand (four exposure environments, seven feedback types), the part of a blind model's calibration bias left after subtracting task and environment means, which no per-task or per-environment correction removes (Proposition 5), holds .60 [.56, .63], .69 [.59, .79] and .31 [.25, .42] of it (95% bootstrap intervals); its size $\|\mathcal{I}\|_F$ is 2.12 in standardized log units and 4.19 and 4.71 in logit units, and a model given the environment as an input keeps .57, .84 and .68 of it. The fitted gain profiles of all seen environments are separated by .19 [.09, .27], .02 [.01, .05] and .04 [.01, .09].

Gain models: wheat and KuaiRand. In a two-factor gain model of wheat (132 varieties), the three pairs of irrigation regimes have gain-profile separations of .51 to .97 (Lemma 3); the three-factor model has a yield-dominated direction that the early traits barely read, so the design amplifies noise 32 to 3,830-fold (Proposition 2). In KuaiRand, with the random-exposure log held out, an environment embedding fitted on its click and long-view labels alone improves the level calibration of four of five slow tasks over a blind model; for like, the median level error falls from 1.51 to 0.15, against 0.07 for joint training with all labels of the new environment.

J HOW DOES THIS WORK RELATE TO ITS CLOSEST PRECEDENTS?

What do the closest precedents assume, and what is specific here? The proofs use canonical polyadic identification (Kruskal, 1977; Leurgans et al., 1993), linear estimability and least-squares risk, and connect them for delayed tasks: a designated slow task's risk with factors estimated from finitely many past windows (Theorem 2), the exact floors of fixed combinations (Proposition 4), which fine-tuning inherits (Proposition 1), and the update whose gain over not adapting is identified (Theorem 1). Table 21 compares when they reach an unobserved cell. Synthetic interventions (Agarwal et al., 2020) need the target unit in the span of donor units (their Assumption 4) and the prediction direction in the observed row space (Assumption 8); with windows as donors this is mixing past windows, as synthetic control mixes donors (Abadie et al., 2010), equation 7 is its error when the span condition fails, and coverage takes the place of their row-space condition for a task with no labels of its own and factors learned from past windows. TensorDG (Li & Zhang, 2023) needs the observed body to keep the full mode ranks, whereas two past windows have window rank two and the running example's complete array rank three; the readouts of Feng et al. (2022) are products of domain-index factors, the form of equation 1. Yang & Hospedales (2015; 2017) write readouts times diagonal gains times shared features and generate a new domain's gains from a known descriptor; we infer them from other tasks' labels. GDFM recovers conversions from post-click behaviour through a stable outcome-to-behaviour channel (Yang & Zhan, 2022, Lemma 3.1), whereas we fix factors and readouts and let the gains move. Multi-task recommenders share towers, experts or gates across tasks and scenarios (Ma et al., 2018a; Tang et al., 2020; Ma et al., 2018b; Sheng et al., 2021; Chang et al., 2023); plant breeding predicts yield from early secondary traits (Rutkoski et al., 2016) and fits three-way models of varieties, environments and traits (Basford et al., 1991). Scaling layers (Perez et al., 2018; Requeima et al., 2019; Liu et al., 2022; Lian et al., 2022; Wang et al., 2021) fit scales from a task's or domain's own labels or inputs, and Proposition 15 applies whenever the basis was not learned with the scales; multi-task representation theory identifies the shared subspace, not its basis

(Tripuraneni et al., 2020; Du et al., 2021); nonlinear ICA identifies latents from environment-wise modulation of their distributions (Hyvärinen & Morioka, 2016; Khemakhem et al., 2020), whereas here the windows modulate how tasks read fixed factors; and covariate- and label-shift corrections see a change of $p(y \mid x)$ only through $p(x)$ (Shimodaira, 2000; Lipton et al., 2018), which our shift need not change (Proposition 5). The lower bound of Theorem 2 for a designated slow task has no counterpart in these works; TensorDG's only lower bound concerns the whole tensor.

Table 21: Closest precedents. "Cells": the observed task–environment (or unit–intervention) cells.

work	structure	cells	guarantee
Feng et al. (2022)	learned representation, product readouts	domains drawn uniformly at random	average risk over domains
Li & Zhang (2023)	linear Tucker low rank	body plus arms; body of full mode ranks	minimax rates
Agarwal et al. (2020)	unit–intervention–time model	target in the donors' span; prediction direction in the observed row space	consistency for the missing entry
ours	learned representation, canonical polyadic, diagonal gains	past windows and fewer than k fast tasks of the current window	slow-task risk and lower bound with estimated factors; combination floors; coverage