REACHABLE ADAPTATION UNDER DELAYED FEEDBACK: BUDGETED PARTIAL UPDATES IN NONSTATIONARY STREAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A deployed predictor may receive labels only after delay while each maintenance cycle can update only a few model blocks. We ask which part of the current error is reachable from the model already in service. We formulate current risk over an anchored reachable family, separating mismatch from delayed or replayed evidence, localization from frozen directions, adaptive route selection, and restricted estimation. Under stated conditional-law and second-moment conditions, projection geometry characterizes localization; a coherence analysis compares an observable risk-change route with a joint-gain population oracle; and replay-aware bounds expose the roles of staleness, effective sample size, and selected dimension. Drift-Decomposed Adaptation (DDA) instantiates this view by differencing anchored block gradients across two delayed windows and refitting only selected blocks. Controlled studies probe these mechanisms and their boundary regimes. On measured Appliances data, DDA has RMSE 0.535 versus 0.537 for Full with 66.7% fewer block updates. On a measured Fjord5G segment, it reduces RMSE by 10.1% relative to GradMag at the same one-block budget. These offline chronological evaluations illustrate regime-dependent resource–error tradeoffs rather than universal selector dominance.

1 INTRODUCTION

A predictor can be inexpensive to run yet expensive to maintain. In sensor and network streams, supervision may arrive only after outcome logging while a maintenance cycle can refresh only a few modules. The decision is therefore *which part of the deployed predictor should receive the available update*. Delayed-label and resource-constrained learning make these restrictions explicit (Csaba et al., 2024; Cai et al., 2020; Lin et al., 2022; Yang et al., 2023).

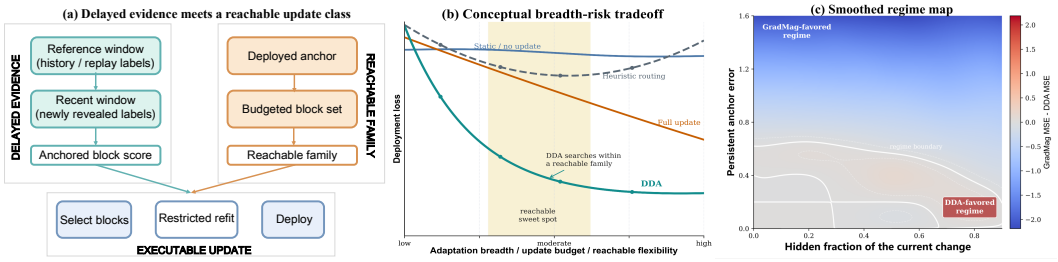


Figure 1: **Update budget changes the relevant comparison.** Measured-stream previews show the full supplied budget grid. On Appliances, similar observed error is attained with fewer updates; on Fjord5G, methods differ at the same budget. Lines join evaluated settings. Reachable objects are defined in Section 3; paired uncertainty appears in Table 2.

Two imperfect resources remain. The deployed predictor retains learned structure but may carry persistent error; delayed labels provide finite evidence from earlier laws, and replay enlarges the

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

sample without making it current. Updating everything ignores the budget, whereas reacting to the largest distributional change ignores whether that direction matters for any feasible next predictor. *The distinction is simple: distributional change asks what changed in the data, whereas maintenance asks what current error can still be reduced by the next feasible update.*

We therefore study *current risk over the family reachable from the deployed anchor*. A selected block set specifies both what may change and what remains fixed. Appendix A gives a nuisance-coordinate shift with positive full-class discrepancy but unchanged reachable losses and zero localization bias: distribution shift can occur in a direction irrelevant to every reachable update. Thus, the relevant comparison class is itself determined by the deployed state and the maintenance action, rather than fixed independently of the decision.

Drift theory, partial-update dynamics, and module-selection methods provide useful tools (Mohri & Muñoz Medina, 2012; Doğançay, 2008; Seo et al., 2025), but they do not by themselves define this anchor-constrained current-risk comparison. The anchor determines the feasible class; delayed/replayed observations determine mismatch; frozen directions determine localization; and data-dependent class choice contributes selection complexity. Our central contribution is to separate these effects at the reachable target.

DDA provides one executable route: compare anchored gradients across delayed windows, choose a feasible block set, and refit without resetting omitted blocks. Projection geometry characterizes costly omissions, while a coherence analysis compares this executable route with a population joint-gain benchmark. Persistent reference error, stale feedback, and block interaction appear explicitly as limits to the comparison.

Our contributions are threefold: **(i)** we formulate adaptation through current risk over anchored reachable families, separating delayed/replay mismatch from frozen-direction localization; **(ii)** we characterize localization through projection geometry, analyze routing relative to the best feasible joint quadratic update, and bound restricted estimation under explicit information and second-moment conditions; and **(iii)** we provide controlled mechanism studies and retrospective measured-stream evaluations of the predicted gains and boundary regimes.

2 Related Work

We organize related work around delayed supervision, distribution shift, and selective updates. Our setting connects them through current risk over an anchor- and budget-dependent reachable family. **1. Delayed-feedback and resource-constrained adaptation.** Delayed-label learning studies prediction when supervision arrives asynchronously (Csaba et al., 2024), while resource-constrained and modular adaptation limits which parameters or modules may be updated (Cai et al., 2020; Lin et al., 2022; Yang et al., 2023; Uzlaner et al., 2025; Seo et al., 2025; Sahoo et al., 2025; Gui et al., 2024). Our distinction is that stale information and restricted movement are treated as separate limitations: the former affects what can be estimated, whereas the latter determines which current errors are reachable at all. **2. Distribution shift over a restricted reachable class.** Domain-adaptation and nonstationary-learning theory compare distributions through hypothesis-class discrepancies and related drift bounds (Ben-David et al., 2010; Mansour et al., 2009; Mohri & Muñoz Medina, 2012; Kuznetsov & Mohri, 2015), with mixing tools handling temporal dependence (Mohri & Rostamizadeh, 2010; Kuznetsov & Mohri, 2017). We instead evaluate mismatch on the class reachable from the deployed anchor by the next feasible update, separating distributional mismatch from localization error due to frozen directions. **3. Partial updates, adaptive selection, and routing.** Partial-update filtering analyzes updates to selected coefficients (Doğançay, 2008; Abadi & Husøy, 2008; Abadi & Moradiani, 2013), while data-dependent-class analysis controls generalization after adaptive class selection (Foster et al., 2019). Our complementary question is how much current-risk reduction a budget-feasible block set can attain and when an observable delayed-data routing score approximates a useful reachable movement. Sequential conformal methods address calibration and risk control under temporal shift rather than the prediction-risk optimization studied here (Gibbs & Candès, 2021; Angelopoulos et al., 2023; Xu & Xie, 2023; Farinhas et al., 2024). Appendix B gives a scope-level comparison of targets, information structures, interventions, and guarantees.

3 CURRENT RISK AND THE REACHABLE INFORMATION STRUCTURE

Notation. For $M \in \mathbb{N}$, write $[M] = \{1, \dots, M\}$. At update time t, P_t and R_t denote the current data law and risk, and $\bar{h}_{t-1}$ the deployed anchor. For $S \subseteq [M]$, $\mathcal{G}_t(S)$ is the anchored reachable class and $B_t(S)$ its localization bias. $\mathcal{H}_{t-}$ denotes the pre-update history, J_t the newly revealed labels, Q_t^H the replay law, ρ_t the fresh-loss mass, and $D_t^H(S)$ the resulting hybrid-to-current mismatch. U_t, V_t denote the reference and recent routing windows, $\widehat{d}_{m,t}$ the DDA block score, and $Q_t(S)$ with $S_{t,k}^\star$ the population joint-gain benchmark. Appendix A gives the complete notation table; theorem-specific symbols are defined at first use.

A comparator is needed before a routing score. The update must start from the deployed model but is judged at the current law, not the law that generated its delayed labels. The target, feasible movement, and information boundary therefore come first.

3.1 DECISION TARGET AND ANCHORED REACHABILITY

At round t, a labeled observation generated at time s is denoted Z_s and is revealed after delay d_s. Let P_t be the current data law and ℓ the prediction loss. The target is

$$R_t(h) = \mathbb{E}_{Z \sim P_t}[\ell(h; Z)] \tag{1}$$

The model has M blocks, $h = (h_1, \dots, h_M) \in \mathcal{H}$, and $\bar{h}_{t-1}$ is the model in service immediately before the update. Block m costs $c_m > 0$ to update. A budget B induces

$$\mathfrak{S}_B = \left\{ S \subseteq [M] : \sum_{m \in S} c_m \le B \right\}. \tag{2}$$

For equal costs we use $\mathcal{S}_{\le K} = \{S \subseteq [M] : |S| \le K\}$. Choosing S does not erase what the other blocks have learned.

Definition 1 (Anchored reachable class and localization bias)**.** *For $S \subseteq [M]$,*

$$\mathcal{G}_t(S) = \{h \in \mathcal{H} : h_m = \bar{h}_{t-1,m} \ \forall m \notin S\}, \tag{3}$$

$$B_t(S) = \inf_{h \in \mathcal{G}_t(S)} R_t(h) - \inf_{h \in \mathcal{H}} R_t(h). \tag{4}$$

$\mathcal{G}_t(S)$ records permitted movement; $B_t(S)$ records its best current-risk loss relative to unrestricted updating. Omitted coefficients retain deployed values rather than resetting to zero (Figure 2).

3.2 DELAYED LABELS, REPLAY, AND THE DEPLOYED ANCHOR

Information available before the update. Let $\mathcal{H}_{t-}$ denote everything known immediately before newly revealed labels are used: the deployed anchor, earlier revealed labels, previous routing decisions, and replay state. The sigma-field notation formalizes which quantities must be predictable. Let J_t index labels newly revealed since the previous update. For a replay measure Q_t^H measurable with respect to this history, the empirical fit and its population comparison are

$$\widehat{R}_t^H(h) = (1 - \rho_t)\mathbb{E}_{Q_t^H}[\ell(h; Z)] + \rho_t \sum_{s \in J_t} \widetilde{\omega}_{t,s} \ell(h; Z_s), \tag{5}$$

$$R_t^H(h) = (1 - \rho_t)\mathbb{E}_{Q_t^H}[\ell(h; Z)] + \rho_t \sum_{s \in J_t} \widetilde{\omega}_{t,s} R_s(h), \tag{6}$$

Here $\widetilde{\omega}_{t,s}$ are predictable normalized fresh weights and $\rho_t \in [0, 1]$ is fresh-loss mass. Replay is fixed after conditioning. *Fresh* means new to this history, not merely in a recent window: sliding windows reuse labels.

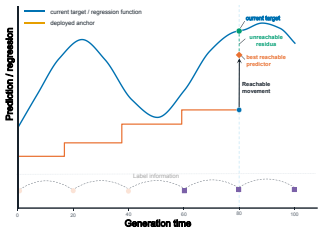


Figure 2: **Reachability and information timing are distinct restrictions.** The deployed anchor limits feasible movement toward the current target, leaving an unreachable residual; delayed labels separately limit which information is available at the update.

We use a loss-composed discrepancy closely related to discrepancy measures in domain adaptation and drifting-distribution analyses (Mansour et al., 2009; Mohri & Muñoz Medina, 2012). For a class $\mathcal{G}$, write

$$\mathrm{disc}_{\ell\circ\mathcal{G}}(P,Q)=\sup_{h\in\mathcal{G}}|\mathbb{E}_P[\ell(h;Z)]-\mathbb{E}_Q[\ell(h;Z)]|.$$

The hybrid fit/current mismatch and the two weight-concentration counts are

$$D_t^H(S)=\sup_{h\in\mathcal{G}_t(S)}|R_t^H(h)-R_t(h)|,\qquad n_{\mathrm{eff},J,t}=\left(\sum_{s\in J_t}\widetilde{\omega}_{t,s}^2\right)^{-1},\qquad n_{\infty,J,t}=\left(\max_{s\in J_t}\widetilde{\omega}_{t,s}\right)^{-1}. \tag{7}$$

$D_t^H(S)$ is the worst-case gap, over updates reachable from the anchor, between the population objective represented by available delayed/replayed evidence and the current target risk. The two counts measure quadratic and maximal fresh-weight concentration; they do not encode temporal dependence.

3.3 Information Conditions and Reliability Regimes

A predictable anchor, predictable weights, and a finite feasible family support the bounded-loss analysis. Conditional independence gives the cleanest display; conditional β-mixing gives a blocked alternative (Mohri & Rostamizadeh, 2010; Kuznetsov & Mohri, 2017). Fast rates additionally require sub-Gaussian score and second-moment control. Squared-loss discrepancy may be infinite on unrestricted classes, so the displayed mismatch must be finite on the reachable class for the bound to be informative.

Conditional-law compatibility. For $s\in J_t$ and $h\in\bigcup_{S\in\mathfrak{S}}\mathcal{G}_t(S)$, we assume

$$\mathbb{E}[\ell(h;Z_s)\mid\mathcal{H}_{t^-}]=R_s(h).$$

This requires newly revealed observations to be conditionally unbiased for the declared time-indexed risks after conditioning on the pre-update history; it is not a conditional-independence assumption. If this compatibility is not appropriate, the population objective should instead be formulated under the corresponding conditional fresh laws. Appendix C summarizes the assumptions used throughout the theory.

4 From Reachable Risk to an Executable Update

We proceed in three layers: first control current target risk for a data-selected reachable class; then compare DDA's executable delayed-window router with a population joint-gain benchmark; finally control estimation within the selected class.

4.1 Selected Target Risk and Delayed-Information Error

For any selected feasible set $\widehat{S}_t$, let Ω be a regularizer and $\lambda\ge 0$ its strength. The reachable fit is

$$\widehat{h}_t\in\arg\min_{h\in\mathcal{G}_t(\widehat{S}_t)}\widehat{R}_t^H(h)+\lambda\Omega(h), \tag{8}$$

The fit freezes omitted blocks. Any selector may reuse its fresh labels. Write $\widehat{\mathfrak{R}}_{J_t}$ for conditional empirical weighted Rademacher complexity (Appendix D.1).

Theorem 1 (Replay-aware selection under a predictable anchor)**.** *Condition on $\mathcal{H}_{t^-}$. Assume $0\le\ell\le 1$, the conditional-law compatibility condition of Section 3.3, and that the fresh observations $\{Z_s:s\in J_t\}$ are conditionally independent given $\mathcal{H}_{t^-}$, with predictable normalized weights $\widetilde{\omega}_{t,s}$. Let $\mathfrak{S}$ be any finite feasible family and let $\widehat{S}_t\in\mathfrak{S}$ be any selector that may use the same fresh labels as Equation 8. If $\lambda=0$, then with conditional probability at least $1-\delta$,*

$$R_t(\widehat{h}_t)-\inf_{h\in\mathcal{H}}R_t(h)\le 4\rho_t\widehat{\mathfrak{R}}_{J_t}(\ell\circ\mathcal{G}_t(\widehat{S}_t))+6\rho_t\sqrt{\frac{\log(4|\mathfrak{S}|/\delta)}{2n_{\mathrm{eff},J,t}}}+2D_t^H(\widehat{S}_t)+B_t(\widehat{S}_t). \tag{9}$$

The terms separate fitting, feasible-class selection, current-law mismatch, and localization. Simultaneous control over the feasible family plays the same general role as complexity control for data-dependent hypothesis sets (Foster et al., 2019); here the finite family permits a direct uniform bound, allowing label reuse without algorithmic enumeration. For $\mathcal{S}_{\le K}$, the selection complexity is of order $K\log(eM/K)$. The compatibility assumption identifies the conditional fresh means with the population risks in Equation 6; concentration then controls estimation around that objective, while D_t^H keeps the remaining target mismatch explicit.

Corollary 1 (Replay mismatch and feedback age)**.** *Let*

$$D^H_{\mathrm{rep},t}(S) = \mathrm{disc}_{\ell\circ\mathcal{G}_t(S)}(Q_t^H, P_t).$$

Then

$$D_t^H(S) \le (1-\rho_t)D^H_{\mathrm{rep},t}(S) + \rho_t \sum_{s\in J_t} \widetilde{\omega}_{t,s}\,\mathrm{disc}_{\ell\circ\mathcal{G}_t(S)}(P_s, P_t). \tag{10}$$

If the fresh-law discrepancy is at most $L_S(t-s)$, the second term is at most $\rho_t L_S \bar{\tau}_{J,t}$, where $\bar{\tau}_{J,t} = \sum_s \widetilde{\omega}_{t,s}(t-s)$.

Replay and label age create distinct staleness terms. Conditional β-mixing replaces the fresh deviation by block terms with an explicit dependence penalty (Corollary 2, Appendix G). Localization remains: which error can no reachable fit remove?

4.2 Localization Geometry and Risk-Change Routing

Omitted movement matters through its residual outside the active span. Correlated active features can make an omitted change cheap. This property characterizes a set before choosing its router.

Theorem 2 (Localization is a projection residual)**.** *Suppose $\ell(h;(x,y)) = (h(x)-y)^2$, $f_t(x) = \mathbb{E}[Y \mid X = x] \in \mathcal{H}$, and $\mathcal{G}_t(S) = \bar{h}_{t-1} + V_S$ for a closed linear subspace $V_S \subset L_2(P_t)$. Then*

$$B_t(S) = \left\|(I - \Pi_{V_S})(f_t - \bar{h}_{t-1})\right\|^2_{L_2(P_t)}. \tag{11}$$

Here Π_{V_S} projects onto the active $L_2(P_t)$ span. Representable omissions cost nothing. Updating standardized Gaussian X_1 while omitting ΔX_2 leaves $\Delta^2(1-\varrho^2)$ at correlation ϱ. Equation 11 characterizes the cost of a fixed set; routing still requires an observable criterion.

Selective partial-update methods provide the broader precedent for modifying only a subset of coefficients at each update (Doğançay, 2008; Abadi & Husøy, 2008; Abadi & Moradiani, 2013). DDA differs in routing that budget through temporal changes in anchored gradients. DDA uses a reference window U_t and a recent delayed window V_t, both containing only available labels. For grouped squared loss, let d_m be the dimension of block m and define

$$\widehat{g}_m(W;\bar{h}) = \widehat{\mathbb{E}}_{s\in W}[x_{s,m}(y_s - \bar{h}(x_s))].$$

Both windows are evaluated at the *same current anchor*. DDA's risk-change score is

$$\widehat{d}_{m,t} = D^{\mathrm{risk}}_{m,t} = \frac{1}{\sqrt{d_m}}\left\|\widehat{g}_m(V_t;\bar{h}_{t-1}) - \widehat{g}_m(U_t;\bar{h}_{t-1})\right\|_2. \tag{12}$$

Stable gradients cancel: the score measures change, not total current error. With threshold τ and cap K,

$$\widehat{k}_t = \max\Big\{1, \min\Big(K, \#\{m : \widehat{d}_{m,t} > \tau\}\Big)\Big\}, \quad \widehat{S}_t = \mathrm{Top}_{\widehat{k}_t}(\widehat{d}_{1,t},\dots,\widehat{d}_{M,t}). \tag{13}$$

The headline implementation fixes $\tau = -\infty$, hence selects exactly K blocks at each eligible update. Figure 3 makes both the mechanism and its persistent-error limitation explicit.

To compare this observable route with a current-risk oracle, consider equal-dimensional grouped linear squared loss. Let $g_{m,t} = \mathbb{E}_t[X_m(Y - \bar{h}_{t-1}(X))]$, and let $g_{m,U}, g_{m,V}$ denote its population window counterparts. Write $\Sigma_m = \mathbb{E}_t[X_m X_m^\top]$ for the current block second-moment matrix, with eigenvalues in $[\lambda_-, \lambda_+]$, and Σ_{mj} for its cross-block counterpart. Intuitively, block coherence measures normalized overlap between candidate update directions: larger overlap makes isolated block scores less representative of their joint current-risk gain. The cross-block interaction is measured by

$$\mu = \max_{m\ne j}\left\|\Sigma_m^{-1/2}\Sigma_{mj}\Sigma_j^{-1/2}\right\|_{\mathrm{op}}, \qquad (K-1)\mu < 1. \tag{14}$$

For joint gradient $g_{S,t}$ and second-moment matrix Σ_S, define the population quadratic gain and its feasible joint-gain oracle by

$$Q_t(S) = g_{S,t}^\top \Sigma_S^{-1} g_{S,t}, \qquad S_{t,k}^\star \in \arg\max_{|S|\le k} Q_t(S). \tag{15}$$

$Q_t(S)$ measures the joint current-risk gain of a feasible block set, accounting for cross-block correlations rather than summing isolated scores; $S_{t,k}^\star$ is the corresponding population joint-gain oracle. The executable score in Equation 12 allows unequal block dimensions d_m; the oracle-comparison result below specializes to common dimension so that the normalization does not alter the block ordering.

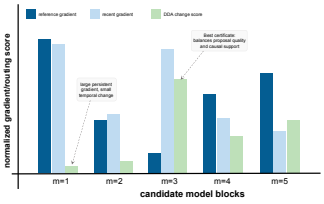


Figure 3: **DDA measures temporal gradient change, not residual size.** A block can retain a large persistent gradient yet receive a small score when its reference and recent gradients are similar; blocks with larger temporal changes are prioritized.

Theorem 3 (Thresholded risk-change routing under block coherence)**.** *Because the blocks have common dimension d, write $\widehat{a}_{m,t} = \sqrt{d}\,\widehat{d}_{m,t}$; this rescales the threshold but not the ordering. Assume $\|g_{m,t}\|_2 \le G$ and, for all blocks,*

$$\|g_{m,U}\|_2 \le \eta_{\mathrm{ref}}, \qquad \|g_{m,V} - g_{m,t}\|_2 \le \eta_{\mathrm{age}}, \qquad |\widehat{a}_{m,t} - \|g_{m,V} - g_{m,U}\|_2| \le \epsilon.$$

These are separate error budgets for persistent reference gradient, recent-to-current mismatch, and score estimation; no common convergence rate is imposed. Let $\xi = \eta_{\mathrm{ref}} + \eta_{\mathrm{age}} + \epsilon$, $\zeta = 2G\xi + \xi^2$, and let $\widehat{k}_t, \widehat{S}_t$ be exactly the threshold-then-top-K rule in Equation 13. For $k = \widehat{k}_t$ define

$$\alpha_{\mu,k} = \frac{1-(k-1)\mu}{\kappa[1+(k-1)\mu]}, \qquad \kappa = \lambda_+/\lambda_-.$$

On the score-error event,

$$Q_t(\widehat{S}_t) \ge \alpha_{\mu,k} Q_t(S_{t,k}^\star) - \frac{2k\zeta}{\lambda_+[1+(k-1)\mu]}. \tag{16}$$

If $\mu = 0$ and $\kappa = 1$, and if the kth and $(k+1)$st population gradient norms (when both exist) are separated by more than 2ξ while the population score is separated from the routing threshold by at least ξ, Equation 13 selects the corresponding oracle set.

Conditioning and interaction weaken the multiplicative comparison, while reference error, label age, and score noise enter through ξ. Theorem 3 compares the executable delayed-window router with a population benchmark: Q_t and $S_{t,k}^\star$ enter the analysis, while deployment uses the score in Equation 12. We next condition on the selected reachable class and bound the error of fitting it from delayed and replayed evidence.

4.3 Restricted Estimation and Deployment

Useful directions still have to be fitted from limited delayed evidence. Here d_S counts active coordinates, $h_{\lambda,S}^H$ is the population ridge minimizer, and $\|\cdot\|_{\psi_2}$ is the conditional sub-Gaussian norm.

Theorem 4 (Replay-aware grouped-ridge fast rate)**.** *Condition on $\mathcal{H}_{t^-}$ and consider grouped linear squared loss. For each feasible S, write $\Sigma_{t,S}^H$ for the conditional hybrid population second-moment matrix and $\Sigma_{s,S}$ for the conditional second-moment matrix of a fresh observation. Assume, for constants $0 < \underline{\lambda}_H \le \overline{\lambda}_H < \infty$ and $\overline{\lambda}_F < \infty$,*

$$\underline{\lambda}_H I \preceq \Sigma_{t,S}^H \preceq \overline{\lambda}_H I, \qquad \Sigma_{s,S} \preceq \overline{\lambda}_F I \quad \textit{for every fresh } s.$$

The fresh covariates satisfy $\|u^\top X_{s,S}\|_{\psi_2} \le K_X (u^\top \Sigma_{s,S} u)^{1/2}$ for every unit u, and the fresh residual at the population ridge point is conditionally sub-Gaussian with scale σ. Assume the conditional-law compatibility condition of Section 3.3 and conditional independence of the fresh observations given $\mathcal{H}_{t^-}$. conditionally independent given $\mathcal{H}_{t^-}$. Let the regularizer in Equation 8 be ridge, $\Omega(h) = \|\theta_S\|_2^2$. Let $d_{\max} = \max_{S\in\mathfrak{S}} d_S$, $D_\delta(S) = d_S + \log(4|\mathfrak{S}|/\delta)$, and $D_{\max,\delta} = d_{\max} + \log(4|\mathfrak{S}|/\delta)$.

$$n_{\mathrm{eff},J,t} \ge C K_X^4 \kappa_{F/H}^2 D_{\max,\delta}, \qquad n_{\infty,J,t} \ge C K_X^2 \kappa_{F/H} D_{\max,\delta}, \qquad \kappa_{F/H} = \overline{\lambda}_F/\underline{\lambda}_H, \tag{17}$$

which is the sufficient restricted-Gram condition of Lemma 2. Define

$$A_{\lambda,t}^H(S) = R_t^H(h_{\lambda,S}^H) - \inf_{h\in\mathcal{G}_t(S)} R_t^H(h).$$

Then, uniformly over the finite feasible family, with conditional probability at least $1-\delta$*,*

$$R_t(\widehat{h}_t) - \inf_{h\in\mathcal{H}} R_t(h) \le C_\Sigma K_X^2\,\rho_t^2\sigma^2\left\{\frac{D_\delta(\widehat{S}_t)}{n_{\mathrm{eff},J,t}} + \frac{D_\delta(\widehat{S}_t)^2}{n^2_{\infty,J,t}}\right\} + 2A^H_{\lambda,t}(\widehat{S}_t) + 2D_t^H(\widehat{S}_t) + B_t(\widehat{S}_t), \tag{18}$$

where one may take $C_\Sigma = C\,\overline{\lambda}_H\overline{\lambda}_F/\underline{\lambda}_H^2$ *for a universal* C*. The score term follows from Lemma 1; the second term inside braces is the Bernstein linear-term price for concentrated weights. For equal fresh weights and* $n_J \gtrsim K_X^4\kappa^2_{F/H}D_{\max,\delta}$*, both effective counts equal* n_J *and the estimation term is* $O(K_X^2\rho_t^2\sigma^2C_\Sigma D_\delta(\widehat{S}_t)/n_J)$*. Corollary 4 gives the corresponding geometric-*β*-mixing specialization.*

The rate depends on selected dimension and family complexity, with $\kappa_{F/H}$ separating fresh-law scale from hybrid curvature. Proposition 1 (Appendix G) gives the matching Gaussian lower-bound scale $\sigma^2\{rd_0 + r\log(eM/r)\}/n$, the grouped analogue of classical sparse-regression minimax rates (Raskutti et al., 2011). Thus active dimension and block-support uncertainty are not merely proof artifacts.

Combining routing and fitting separates the hard budget gap from route, staleness, and estimation terms under the stated conditions.

Corollary 3 (End-to-end DDA current-risk bound)**.** *Under Theorems 3 and 4, let* $k = \widehat{k}_t$*,* $S^\star_{t,K}$ *and* $S^\star_{t,k}$ *be the joint-gain oracles, and suppose the full grouped-linear oracle is reachable by updating all blocks. Then*

$$\begin{aligned} R_t(\widehat{h}_t) - \inf_{h\in\mathcal{H}} R_t(h) \le{}& \mathcal{E}_{t,k} + 2D_t^H(\widehat{S}_t) + 2A^H_{\lambda,t}(\widehat{S}_t) \\ &+ \underbrace{Q_t([M]) - Q_t(S^\star_{t,K})}_{\text{unavoidable } K\text{-budget gap}} + \underbrace{Q_t(S^\star_{t,K}) - Q_t(S^\star_{t,k})}_{\text{threshold under-use}} \\ &+ \underbrace{(1-\alpha_{\mu,k})Q_t(S^\star_{t,k})}_{\text{block interaction / conditioning}} + \underbrace{\frac{2k\zeta}{\lambda_+[1+(k-1)\mu]}}_{\text{routing estimation / staleness}}, \end{aligned} \tag{19}$$

where $\mathcal{E}_{t,k} = C_\Sigma K_X^2\rho_t^2\sigma^2\{D_\delta(\widehat{S}_t)/n_{\mathrm{eff},J,t} + D_\delta(\widehat{S}_t)^2/n^2_{\infty,J,t}\}$ *under the conditional-independence specialization of Theorem 4.*

With the hard budget, $k = K$, so threshold under-use vanishes. The remaining terms separate the unavoidable budget gap from interaction/conditioning, staleness and reference error, and estimation. Algorithm 1 implements the executable update using only revealed labels: the deployed anchor is fixed before routing, U_t and V_t contain only labels available by time t, omitted blocks retain their anchor values, and the refit is committed before the next prediction interval. The linear runner refits a shared intercept outside the feature-block count.

Algorithm 1 DDA under a hard block budget

Input: Warm-up data; blocks; fixed K, λ; windows; reveal schedule.
1: Fit warm-up model and transforms; freeze transforms.
2: **for** each eligible scheduled update t **do**
3: Fix $\bar{h}_{t-1}$; form U_t, V_t from labels revealed by t.
4: Score by Eq. 12; select top-K.
5: Fit Eq. 8 on the declared fitting/replay objective, preserving omitted blocks.
6: Commit before the next prediction interval.
7: **end for**
8: Otherwise retain the deployed model.

5 Experiments: Reachable Gains and Their Boundaries

The experiments target specific implications of the analysis rather than a single global performance claim. Figure 4a probes the projection implication of Theorem 2; Figure 4b combines the routing and selected-dimension mechanisms of Theorems 3 and 4; and Figure 4c stresses persistent reference error. Delay and replay separately probe feedback age and the estimation–mismatch tradeoff. The controlled studies ask whether these mechanisms remain visible under finite perturbations; the measured streams examine their combined resource–error behavior.

• **Controlled mechanism studies:** Grouped-linear simulations vary proxyability, ambient dimension, persistent reference error, label visibility, replay, and routing geometry. Simulation truth provides population oracles only as analysis references; DDA itself routes from sampled delayed-window scores. The routing-geometry stress separately varies block coherence and within-block conditioning.

• **Measured chronological streams:** We evaluate Appliances (Candanedo, 2017) and Fjord5G (Denizer et al., 2025) under fixed chronological update schedules using warm-up-fitted transforms and ridge strength 0.05. The headline schedules are summarized in Table 1. These experiments characterize the resource–error tradeoffs of the executable protocol. The equal-block quadratic routing mechanism is evaluated separately in the controlled simulations, where the required population quantities are available.

Table 1: **Headline experimental settings.**

Setting	Configuration
Controlled	Grouped-linear; proxyability, dimension, reference error, delay/replay, and routing geometry varied
Appliances	19,735 rows; 6 blocks; 5,000 warm-up; delay 72; window 2,016; $K = 2$ per 144 rows
Fjord5G	350,000 rows; 4 blocks; 50,000 warm-up; delay 600; window 20,000; $K = 1$ per 5,000 rows
Comparators	DDA, GradMag, Recent gain, Random-K, Full, Static
Uncertainty	600-resample paired moving-block bootstrap

• **Comparators and uncertainty:** Partial selectors share data availability, initialization, fitting code, windows, and per-update budget, although their deployed anchors diverge. GradMag uses recent anchored-gradient magnitude, Recent gain normalizes by local covariance, and Random-K is uninformed; Full and Static are resource endpoints. Table 2 reports fixed-trajectory RMSE with paired moving-block bootstrap intervals from 600 resamples. Complete protocols, seeds, sensitivity grids, and result ledgers appear in Appendix H.

5.1 Main Results and Cross-Setting Robustness

In the unit-signal controlled construction, the projection prediction is quantitative: the residual is $1 - \varrho^2$ (Figure 4a). Table 3 summarizes the active-dimension sweep. At these controlled settings, DDA recovers the oracle active set, so the DDA and Oracle rows coincide at the displayed precision, while Full increasingly pays for ambient dimension.

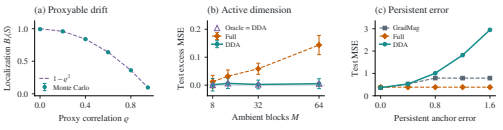


Figure 4: **Controlled tests of theory-facing mechanisms.** (a) Proxyable omissions follow the projection residual. (b) Successful routing reduces the dimension that must be fit. (c) Persistent reference error weakens temporal-change routing. Lines connect evaluated settings.

In the separate routing-geometry construction, increasing coherence from $\mu = 0$ to 0.8 lowers the population joint-gain ratio of the score-selected route from 1.00 to 0.602; increasing κ from 1 to 16 lowers it to 0.591 (finite-window mean 0.615). These are controlled manifestations of the weakening comparison factor, not a claim of universal monotonic behavior.

At the fixed headline settings in Table 2, Appliances DDA has RMSE 0.5354 versus Full's 0.5373 with 66.7% fewer block updates. On Fjord5G, DDA improves 10.1% over GradMag at the same one-block budget and uses 75% fewer block updates than Full. The paired interval excludes zero for the Fjord5G DDA–GradMag difference, whereas the intervals against Recent gain and Full include zero. Intervals that include zero are treated as inconclusive rather than as evidence of equivalence.

Table 2: **Fixed-trajectory RMSE and paired uncertainty.** Intervals are 95% block-bootstrap intervals for DDA minus the comparator; negative favors DDA. Random-K reports ten-seed mean ± SD. Partial/Full block updates per 1,000 rows are 13.845/41.534 (Appliances) and 0.197/0.787 (Fjord5G).

	Appliances		Fjord5G	
Method	RMSE	Paired interval	RMSE	Paired interval
DDA	0.5354	—	1.0905	—
GradMag	0.5345	$[-.0079, .0107]$	1.2131	$[-.2422, -.0192]$
Recent gain	0.5344	$[-.0082, .0118]$	1.2227	$[-.3925, .1600]$
Random-K	$.5418 \pm .0056$	—	$1.3601 \pm .1686$	—
Full	0.5373	$[-.0136, .0109]$	1.1362	$[-.2628, .1853]$
Static	0.7799	$[-.2837, -.2101]$	1.2298	$[-.3631, .0493]$

Across four additional Fjord5G segments, the ranking varies: GradMag, Recent gain, and Static each lead in at least one segment. The three-seed neural check is similarly mixed: Appli-

ances gives 0.5345 for GradMag versus 0.5479 for DDA, while Fjord5G gives 0.7285 and 0.7279, respectively. GALA-style transfers a cosine-alignment criterion (Sahoo et al., 2025) into the same delayed supervised protocol rather than reproducing the original test-time setting. These checks delimit the empirical scope across segments and backbones (Appendix H.4).

5.2 Ablation and Sensitivity Analyses

Resetting omitted blocks changes the reachable class: anchoring-control MSE rises from 0.2522 to 2.2646. Full removes localization, GradMag removes temporal differencing, and the true active set supplies a simulation-only oracle. These ablations target the mathematical ingredients. Delay weakens the relation between recent and current gradients. In the controlled sweep, DDA's recovery falls from 1.00 at zero hidden post-change fraction to 0.22 at 0.9. At the largest measured delays, RMSE reaches 0.5529 on Appliances and 1.1664 on Fjord5G, although Appliances is not monotone at intermediate delays. Additional window and ridge sweeps appear in Appendix H.5; longer windows may also reduce the number of eligible updates.

Table 3: **Active-dimension scaling.** Reported are mean test excess MSEs over 30 seeds. Relative reduction compares DDA with Full; larger values favor DDA.

Number of blocks M	8	16	32	64
DDA	.0015	.0064	.0025	.0053
Full	.0124	.0320	.0587	.1436
Oracle	.0015	.0064	.0025	.0053
Rel. reduction (%)	**88.25**	**80.06**	**95.72**	**96.33**

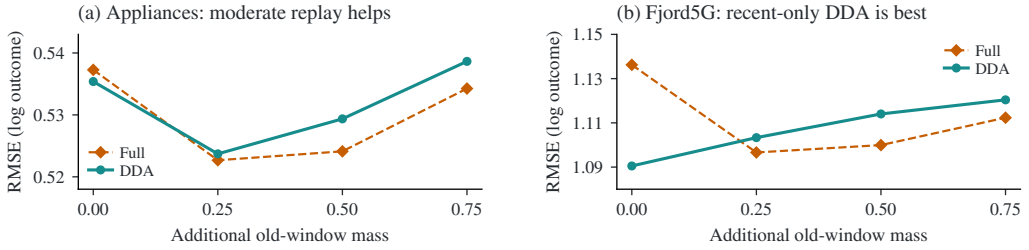


Figure 5: **Adding older evidence does not have a universal sign.** At the same update budget, modest replay first helps and then hurts Appliances, while Fjord5G worsens as older-window mass increases. Full is a different-resource reference.

Replay probes the estimation–mismatch tradeoff. Appliances DDA reaches 0.5237 RMSE at older-window mass 0.25 and 0.5386 at 0.75, while Fjord5G moves from 1.0905 to 1.1204 (Figure 5). Per-update budgets, rather than resulting anchors or trajectories, are fixed; the theoretical components are not separately identified.

5.3 Efficiency and Error Analysis

The headline DDA runs touch 61.7% and 81.7% fewer feature coordinates than Full on Appliances and Fjord5G. In the controlled $M = 128$ timing run, DDA touches 8 of 512 coordinates but takes 17.825 ms versus Full's 17.887 ms, so the empirical claim is reduced update volume rather than proportional wall-clock speedup. At stable reference error 1.2, DDA MSE is 1.8176 versus GradMag's 0.7900, illustrating a boundary where recent change is detected while a larger repair opportunity is missed (Figure 4c). Coherence, close score margins, and unavailable labels provide additional boundaries; saved Fjord5G trajectories are descriptive rather than causal (Appendix H.6).

6 Discussion and Conclusion

The central lesson is that budgeted adaptation should be framed not by *what changed*, but by *what the next feasible update can still fix*. Once the deployed model anchors the admissible movement, distributional shift and reducible current error become different objects: delayed evidence limits what can be inferred, while frozen directions limit what can be repaired. This reachable-current-risk view is the paper's main conceptual contribution. DDA is one executable route within this view rather than the view itself. More broadly, budgeted model maintenance should distinguish *reachability*, *information*, and *routing* rather than compressing them into a single notion of drift.

AI USE STATEMENT

Generative AI tools were used for language editing, structural refinement, formatting assistance, and exploratory discussion during manuscript preparation. The authors are responsible for all theoretical claims, experimental results, analyses, and final conclusions.

ETHICS STATEMENT

This work studies predictive model maintenance under delayed feedback and hard partial-update budgets. The experiments use controlled simulations and previously released Appliances and Fjord5G data and do not involve live deployment or intervention. The proposed analysis is not a safety guarantee; deployment in high-stakes settings would require application-specific validation beyond the scope of this work.

REPRODUCIBILITY STATEMENT

Appendices C–H contain the assumption ledger, full proofs, implementation details, dependent-data complements, experimental protocols, and reconstruction information. The artifact preserves original code and cached outputs, adds an independent panel renderer and a frozen-design reproduction wrapper, and records which checks were executed.

REFERENCES

Mohammad Shams Esfand Abadi and John Håkon Husøy. Mean-square performance of the family of adaptive filters with selective partial updates. *Signal Processing*, 88(8):2008–2018, 2008.

Mohammad Shams Esfand Abadi and Fatemeh Moradiani. A unified approach to tracking performance analysis of the selective partial update adaptive filter algorithms in nonstationary environment. *Digital Signal Processing*, 23(3):817–830, 2013.

Anastasios N. Angelopoulos, Emmanuel J. Candes, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. TinyTL: Reduce memory, not parameters for efficient on-device learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Luis M. Candanedo. Appliances energy prediction. UCI Machine Learning Repository, 2017.

Botos Csaba, Wenxuan Zhang, Matthias Müller, Ser-Nam Lim, Mohamed Elhoseiny, Philip H. S. Torr, and Adel Bibi. Label delay in online continual learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Birkan Denizer, Nils Dohse, and Olaf Landsiedel. Fjord5G: A comprehensive 5G dataset for coastal maritime connectivity. In *2025 IEEE 101st Vehicular Technology Conference (VTC2025-Spring)*, pp. 1–5, 2025.

Kutluyıl Doğançay. *Partial-Update Adaptive Signal Processing: Design, Analysis and Implementation*. Academic Press, 2008.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André F. T. Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations*, 2024.

Dylan J. Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021.

Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. In *International Conference on Learning Representations*, 2024.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. On-device training under 256KB memory. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, pp. 124–138, 2012.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10): 6976–6994, 2011.

Sabyasachi Sahoo, Mostafa ElAraby, Jonas Ngnawe, Yann Batiste Pequignot, Frédéric Precioso, and Christian Gagné. A layer selection approach to test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20237–20245, 2025.

Minhyuk Seo, Hyunseo Koh, and Jonghyun Choi. Budgeted online continual learning by adaptive layer freezing and frequency-based sampling. In *International Conference on Learning Representations*, 2025.

Nicole Uzlaner, Tomer Raviv, Nir Shlezinger, and Koby Todros. Asynchronous online adaptation via modular drift detection for deep receivers. *IEEE Transactions on Wireless Communications*, 24(5): 4454–4468, 2025.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 38707–38727, 2023.

Yuedong Yang, Guihong Li, and Radu Marculescu. Efficient on-device training via gradient filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3811–3820, 2023.

Appendix

CONTENTS

A Notation and Extended Information Setup **14**
A.1 Delayed history, fresh labels, and replay 14
A.2 Feasible families and reachable classes 15
A.3 Why full-class drift can be irrelevant 15
A.4 Population gradients and joint quadratic gain 15

B Related-Work Scope Comparison **15**

C Assumptions, Diagnostics, and Special Cases **15**

D Proof Map and Full Proofs **17**
D.1 Conditional weighted deviation 17
D.2 Proof of Theorem 1 17
D.3 Proof of Corollary 1 18
D.4 Proof of Corollary 2 18
D.5 Proof of Theorem 2 18
D.6 Block coherence lemma 19
D.7 Proof of Theorem 3 19
D.8 Fresh score and covariance concentration 20
D.9 Dependent-data supply of the fast-rate events 22
D.10 Proof of Proposition 1 22
D.11 Proof of Corollary 3 23

E Full Algorithm and Implementation Details **23**
E.1 Executable DDA pipeline 23
E.2 Information use, replay decomposition, and label reuse 24
E.3 Computational path 24
E.4 Optional one-step conservative decision diagnostic 24

F Additional Assumptions and Scope **24**
F.1 Population oracles and executable quantities 24
F.2 Conditional targets, dependence, and finite discrepancy 25
F.3 Covariance and routing scope 25
F.4 Preprocessing and statistical interpretation 25
F.5 Resource interpretation 25

G Additional Statistical Statements **25**

H Current Experiment Protocol and Evidence **25**

H.1 Current data and chronological fitting protocol 26

H.2 Uncertainty and exact saved results 26

H.3 Controlled study designs and failure contrasts 27

H.4 Cross-segment and neural checks 28

H.5 Sensitivity and replay grids 29

H.6 Cost and trajectory diagnostics 31

H.7 Current numeric ledgers 32

A Notation and Extended Information Setup

This section collects the nonstandard notation used throughout the paper and then expands the information model of Section 3.

Table 4: Core notation. Symbols that are standard within a local theorem are omitted.

Symbol	Meaning
t, s, d_s	update time, observation-generation time, and label delay
$P_t,\ R_t(h)$	current data law and current prediction risk
$\bar{h}_{t-1}$	deployed model immediately before update t
$\mathfrak{S}_B,\ \mathcal{S}_{\le K}$	feasible block families under a cost or cardinality budget
$\mathcal{G}_t(S)$	models reachable by changing only blocks in S from the deployed anchor
$B_t(S)$	localization bias of reachable set S relative to unrestricted current risk
$\mathcal{H}_{t^-},\ J_t$	pre-update information and labels newly revealed for the current update
$Q_t^H,\ \rho_t$	replay distribution and fresh-loss mass in the hybrid fitting objective
$D_t^H(S)$	hybrid-to-current risk mismatch restricted to $\mathcal{G}_t(S)$
$n_{\mathrm{eff},J,t},\ n_{\infty,J,t}$	quadratic and maximal fresh-weight effective counts
U_t, V_t	reference and recent delayed routing windows
$\widehat{d}_{m,t}$	DDA temporal risk-gradient change score for block m
$Q_t(S),\ S_{t,k}^\star$	population joint quadratic gain and its size-k oracle set
$\mu,\ \kappa$	normalized cross-block coherence and within-block conditioning factor

Information boundary. The target remains current-time risk after a delayed, budgeted partial update; the additional notation identifies which objects are fixed before newly revealed labels are incorporated.

A.1 Delayed history, fresh labels, and replay

Let $d_s \ge 0$ be the delay of label Y_s and let

$$\mathcal{W}_t = \{s < t : s + d_s \le t\}$$

be the indices whose labels are available by round t. Immediately before the labels newly available for the present update are used, let $\mathcal{H}_{t^-}$ denote the sigma-field generated by all earlier observations, fitted models, routing decisions, hyperparameters fixed from training/validation, and replay state. The deployed anchor $\bar{h}_{t-1}$ is $\mathcal{H}_{t^-}$-measurable.

Let $J_t \subseteq \mathcal{W}_t$ be the indices whose labels became available since the preceding update opportunity. A fit window F_t can be written as a disjoint union $F_t = \mathcal{R}_t \cup J_t^F$, where $\mathcal{R}_t$ contains replayed observations already measurable with respect to $\mathcal{H}_{t^-}$ and $J_t^F \subseteq J_t$ contains newly revealed labels. For a uniform fit over F_t,

$$\rho_t = \frac{|J_t^F|}{|F_t|}, \qquad Q_t^H = \frac{1}{|\mathcal{R}_t|}\sum_{s\in\mathcal{R}_t}\delta_{Z_s}, \qquad \widetilde{\omega}_{t,s} = |J_t^F|^{-1};$$

the weighted case is analogous. This gives exactly Equations 5–6. The fresh-only protocol has $\mathcal{R}_t = \varnothing$ and $\rho_t = 1$.

The reference and recent routing windows U_t, V_t may contain any labels available by round t. Portions already in $\mathcal{H}_{t^-}$ are fixed after conditioning; any newly revealed labels used by the router are handled by uniformity over the feasible family in Theorem 1. No label that is still unavailable at t is used for routing, fitting, tuning, or calibration.

Conditional mixing convention. Whenever a dependent-data statement says that the fresh sequence is conditionally β-mixing given $\mathcal{H}_{t^-}$, the coefficient $\beta(a)$ refers to the conditional law of the ordered fresh observations after fixing $\mathcal{H}_{t^-}$. All coupling probabilities below are therefore conditional on the same pre-update history used in the main theorems; no unconditional stationarity assumption is introduced by the blocking step.

A.2 Feasible families and reachable classes

The cost family is $\mathfrak{S}_B = \{S : \sum_{m \in S} c_m \le B\}$ and the equal-cost family is $\mathcal{S}_{\le K} = \{S : |S| \le K\}$. Theorem 1 needs only a finite family $\mathfrak{S}$; hardware placement constraints, expert availability, or cost knapsacks can therefore be represented by changing the family rather than the proof. For each feasible S, the anchor determines

$$\mathcal{G}_t(S) = \{h \in \mathcal{H} : h_m = \bar{h}_{t-1,m}\ \forall m \notin S\}.$$

After conditioning on $\mathcal{H}_{t^-}$, the entire family $\{\mathcal{G}_t(S) : S \in \mathfrak{S}\}$ is fixed even though the anchor was learned from earlier data.

A.3 Why full-class drift can be irrelevant

Let $X = (X_a, X_n)$ and $Y = X_a$. Under laws P and Q, let X_a have the same centered distribution and let X_n be independent of (X_a, Y) with variance one under P and variance four under Q. For squared-loss predictors $h_{a,b}(x) = aX_a + bX_n$, suppose the next update may change a while the anchor fixes $b = 0$. Every reachable loss $(aX_a - Y)^2$ has the same law under P and Q, so

$$\mathrm{disc}_{\ell \circ \mathcal{G}}(P, Q) = 0.$$

The oracle $h_{1,0}$ is reachable, hence $B(S) = 0$. In the full class, $h_{1,b}$ changes risk by $3b^2$, so full-class discrepancy is positive whenever $b \neq 0$. The example isolates the setting in which a full-class drift detector spends attention on a direction the next update neither needs nor can use.

A.4 Population gradients and joint quadratic gain

For a law P, anchor $\bar{h}$, and block m, define

$$g_m(P; \bar{h}) = \mathbb{E}_P[X_m(Y - \bar{h}(X))].$$

For a weighted window, $g_m(W; \bar{h})$ is the corresponding population mixture gradient and $\widehat{g}_m(W; \bar{h})$ its empirical estimate. For a block set S, g_S stacks the gradients and $\Sigma_S = \mathbb{E}_t[X_S X_S^\top]$. In linear squared loss, optimizing the active coefficients from the anchor reduces current risk by

$$Q_t(S) = g_{S,t}^\top \Sigma_S^{-1} g_{S,t},$$

provided Σ_S is nonsingular. This identity remains valid with correlated blocks; additivity across individual blocks is not required.

Implication. The anchor and replay buffer may depend on all earlier observations, the route may use currently revealed labels, and the stochastic step conditions only on labels that are new relative to the anchor. Evaluation remains under the current law P_t.

B Related-Work Scope Comparison

Table 5 summarizes the modeling scope of the closest methodological lines cited in Section 2. "Explicit" means that the object is directly modeled or analyzed in the cited line, "partial" means that closely related structure appears but is not the primary comparison used here, and "–" means that the object is not central to that line. The table is a scope map rather than a quality ranking, and it does not assert that prior methods cannot be extended.

C Assumptions, Diagnostics, and Special Cases

Conditions and scope. Table 6 records where each condition is used and what changes when it fails. Conditions supporting the method-specific quadratic theorem are separate from the assumptions of the general reachable-class decomposition.

Proxyability. Theorem 2 does not assume that an omitted block is irreplaceable. If the current target change lies in the active span, its localization cost is zero. A positive omission lower bound requires an additional non-proxy condition; none is used in the main upper bounds.

Table 5: Scope-level comparison with representative related work.

Representative line	Delayed feedback	Hard partial-update budget	Deployed-anchor reachability	Current-law target	Replay/age mismatch	Frozen-direction localization	Adaptive route/ class selection
Resource-efficient / on-device training (Cai et al., 2020; Lin et al., 2022; Yang et al., 2023)	–	Explicit	Partial	–	–	–	Partial
Delayed-label learning (Csaba et al., 2024)	Explicit	–	–	Partial	Partial	–	Partial
Domain adaptation and drift (Ben-David et al., 2010; Mansour et al., 2009; Mohri & Muñoz Medina, 2012; Kuznetsov & Mohri, 2015)	Partial	–	–	Explicit	Explicit	–	–
Modular / selective / test-time adaptation (Uzlaner et al., 2025; Seo et al., 2025; Sahoo et al., 2025; Gui et al., 2024)	Partial	Partial	Partial	Partial	Partial	–	Explicit
Partial-update adaptive filtering (Doğançay, 2008; Abadi & Husøy, 2008; Abadi & Moradiani, 2013)	Partial	Explicit	Partial	Partial	–	Partial	Explicit
Reachable-risk view / DDA (ours)	Explicit	Explicit	Explicit	Explicit	Explicit	Explicit	Explicit

Table 6: Assumption ledger and corresponding diagnostics.

Condition	Used for	Operational meaning	If violated	Stress / diagnostic
Finite feasible family $\mathfrak{S}$	Thm. 1	Finite menu of allowed block sets	Replace the union term by complexity of a richer selector family	K/M scaling
Pre-update measurability of anchor/replay	Thm. 1, Thm. 4	$\bar{h}_{t-1}$ and replay are fixed before current fresh-label noise is exposed	Need stability/adaptive-data analysis for a simultaneously refit anchor	replay/fresh protocol check
Predictable fresh weights	Thm. 1	A fresh label cannot choose its own loss weight after its value is observed	Add adaptive-weight correction or split the data	protocol check
Conditional-law compatibility	Thms. 1, 4	Conditional fresh loss means agree with the time-indexed risks in Eq. 6	Reformulate the population objective using conditional target laws	information-model check
Conditional independence of fresh labels	clean Thm. 1 display	Fresh innovations are independent after conditioning	Cor. 2 replaces the sample term under β-mixing	chronological streams
Bounded loss	Thm. 1	Clean bounded-difference route	Use Thm. 4, robust losses, or unbounded-process concentration	measured regression; Huber check
Temporal discrepancy Lipschitzness	feedback-age specialization only	Reachable loss laws do not change arbitrarily fast	Keep the exact discrepancy in Eq. 10	delay sweep
Closed affine reachable space	Thm. 2	Active variations form a closed $L_2(P_t)$ span	Replace projection by distance to the actual reachable set	proxy/correlation sweep
Block coherence $(K-1)\mu < 1$ and conditioning	Thm. 3	Joint active covariance remains nonsingular and not arbitrarily interactive	Approximation factor can collapse while Thm. 1 remains valid	routing-geometry stress (Fig. 7)
Reference error η_{ref}	Thm. 3	Stable directions have small anchored gradient on U_t	Persistent error contaminates temporal change	stable-gradient counterexample
Recent/current error η_{age}	Thm. 3	Recent revealed labels still represent current gradient	Stale routing can miss new directions	delay/gradual drift
Uniform fresh covariance cap $\Sigma_{s,S} \preceq \bar{\lambda}_F I$	Lemmas 1–2, Thm. 4	No single fresh law/window has arbitrarily larger feature variance than the declared scale	Replace n_{eff} by a leverage/heteroskedastic concentration bound	analytic assumption; not directly certified on measured streams
Hybrid covariance bounds $\underline{\lambda}_H I \preceq \Sigma^H_{t,S} \preceq \bar{\lambda}_H I$	Thm. 4	The replay–fresh population objective has restricted curvature and a finite risk scale	Fast-rate curvature/risk conversion is not asserted; Thm. 1 remains available	analytic assumption; active-dimension study probes the resulting estimation mechanism
Radius coverage	one-step diagnostic only	Delayed residual radius reaches target coverage	Prediction-risk theorems remain; certificate is unavailable	fixed/adaptive radius

Dependence and heavy tails. The measured streams are chronological, so the paper never treats independence as an empirical fact. Corollary 2 gives a formal dependent-data route for bounded losses. The squared-loss fast theorem cites the weighted score and restricted-Gram lemmas below; Section D.9 gives an explicit geometric-β-mixing block specialization. Huber regression is an empirical robustness check, not a theorem about arbitrary heavy tails.

Scope. The broad guarantee requires little geometry beyond a finite feasible family and a valid fresh-data concentration route. The oracle-routing guarantee additionally uses coherence, conditioning, and score-margin conditions. Figure 7 isolates the first two in a controlled Gaussian construction; the measured streams are not used to certify these population conditions, and the score-margin requirement remains an analytic condition rather than a measured-data certificate.

D Proof Map and Full Proofs

Proof map. The proof dependencies are

$$\begin{aligned}
\text{conditional fresh deviation} &\Rightarrow \text{Thm. 1} \Rightarrow \text{Cor. 1},\\
\text{independent-block coupling} &\Rightarrow \text{Cor. 2},\\
L_2 \text{ projection} &\Rightarrow \text{Thm. 2},\\
\text{coherence + score perturbation} &\Rightarrow \text{Thm. 3},\\
\text{weighted score concentration} &\Rightarrow \text{the score-error control used in Thm. 3},\\
\text{fresh covariance cap + weighted concentration} &\Rightarrow \text{restricted Gram/score events},\\
\text{restricted events + hybrid curvature} &\Rightarrow \text{Thm. 4},\\
\text{sparse packing} &\Rightarrow \text{Prop. 1}.
\end{aligned}$$

and Corollary 3 substitutes the routing lower bound into the localization term of Theorem 4.

D.1 Conditional weighted deviation

Fix $S \in \mathfrak{S}$ and condition on $\mathcal{H}_{t^-}$. Then $\bar{h}_{t-1}$, Q_t^H, and $\mathcal{G}_t(S)$ are fixed. Let

$$\mathcal{F}_S = \ell \circ \mathcal{G}_t(S), \qquad \Delta_{J,S} = \sup_{f \in \mathcal{F}_S} \left| \sum_{s \in J_t} \widetilde{\omega}_{t,s} \{ f(Z_s) - \mathbb{E}[f(Z_s) \mid \mathcal{H}_{t^-}] \} \right|.$$

Under conditional independence and $f \in [0,1]$, weighted symmetrization followed by bounded differences yields

$$\Delta_{J,S} \le 2\widehat{\mathfrak{R}}_{J_t}(\mathcal{F}_S) + 3\sqrt{\frac{\log(2/\delta_S)}{2n_{\mathrm{eff},J,t}}} \tag{20}$$

with conditional probability at least $1 - \delta_S$. By the conditional-law compatibility assumption, $\mathbb{E}[\ell(h; Z_s) \mid \mathcal{H}_{t^-}] = R_s(h)$ on every feasible reachable class. Since the replay term is identical in Equations 5 and 6,

$$\sup_{h \in \mathcal{G}_t(S)} |\widehat{R}_t^H(h) - R_t^H(h)| = \rho_t \Delta_{J,S}. \tag{21}$$

Taking $\delta_S = \delta/|\mathfrak{S}|$ and a union bound makes the event simultaneous over every feasible set. This is the step that allows the route to reuse the fresh labels.

D.2 Proof of Theorem 1

Work on the simultaneous event above and abbreviate $\widehat{S} = \widehat{S}_t$, $\mathcal{G} = \mathcal{G}_t(\widehat{S})$, and $D = D_t^H(\widehat{S})$. For $\varepsilon > 0$, choose $h_{t,\widehat{S}}^{\varepsilon} \in \mathcal{G}$ with $R_t(h_{t,\widehat{S}}^{\varepsilon}) \le \inf_{h \in \mathcal{G}} R_t(h) + \varepsilon$. By the definition of D,

$$|R_t(h) - R_t^H(h)| \le D \quad \forall h \in \mathcal{G}. \tag{22}$$

For $\lambda = 0$, the ERM property and Equations 21–22 give

$$\begin{aligned}
R_t(\widehat{h}_t) &\le R_t^H(\widehat{h}_t) + D\\
&\le \widehat{R}_t^H(\widehat{h}_t) + \rho_t \Delta_{J,\widehat{S}} + D\\
&\le \widehat{R}_t^H(h_{t,\widehat{S}}^{\varepsilon}) + \rho_t \Delta_{J,\widehat{S}} + D\\
&\le R_t^H(h_{t,\widehat{S}}^{\varepsilon}) + 2\rho_t \Delta_{J,\widehat{S}} + D\\
&\le R_t(h_{t,\widehat{S}}^{\varepsilon}) + 2\rho_t \Delta_{J,\widehat{S}} + 2D.
\end{aligned}$$

Subtracting $\inf_{h \in \mathcal{H}} R_t(h)$ introduces $B_t(\widehat{S}) + \varepsilon$. Letting $\varepsilon \downarrow 0$ and substituting Equation 20 proves Equation 9. Notice that the argument never requires the anchor to be independent of earlier observations; it is fixed only after conditioning on $\mathcal{H}_{t^-}$.

For equal-cost sets,

$$|\mathcal{S}_{\le K}| = \sum_{j=0}^{K} \binom{M}{j} \le (K+1)(eM/K)^K,$$

so the route-selection price scales as $K \log(eM/K)$ up to lower-order terms.

D.3 Proof of Corollary 1

For each feasible S, the hybrid fit/current difference is a convex combination of a replay part and a fresh-law part. The triangle inequality therefore gives

$$D_t^H(S) \le (1-\rho_t) \sup_{h\in\mathcal{G}_t(S)} |\mathbb{E}_{Q_t^H}[\ell(h;Z)] - R_t(h)| + \rho_t \sum_{s\in J_t} \widetilde{\omega}_{t,s} \sup_{h\in\mathcal{G}_t(S)} |R_s(h) - R_t(h)|,$$

which is Equation 10. If each fresh discrepancy is bounded by $L_S(t-s)$, the second line is at most $\rho_t L_S \bar{\tau}_{J,t}$. No temporal smoothness is required for the replay term; its actual mismatch remains visible.

D.4 Proof of Corollary 2

Write $n_J = |J_t|$, choose a with $N_{\mathrm{blk}} = \lfloor n_J/(2a) \rfloor \ge 1$, and set $q = n_J - 2aN_{\mathrm{blk}} < 2a$. Partition the first $2aN_{\mathrm{blk}}$ observations into consecutive blocks $B_1, \dots, B_{2N_{\mathrm{blk}}}$ of length a and leave the final q observations as a remainder. Let $\widetilde{B}_1, \dots, \widetilde{B}_{2N_{\mathrm{blk}}}$ be independent blocks with the same block marginals. Under the conditional mixing convention of Appendix A, the independent-block coupling changes the probability of any event determined by the alternating blocks by at most $(2N_{\mathrm{blk}}-1)\beta(a)$ (Kuznetsov & Mohri, 2017; Mohri & Rostamizadeh, 2010).

Apply the conditional symmetrization argument of Section D.1 separately to the odd and even independent block sequences, treating

$$\bar{f}(B_j) = a^{-1} \sum_{s\in B_j} f(Z_s)$$

as one $[0,1]$-valued super-observation. A union bound over the two subsequences and $S \in \mathfrak{S}$ gives, with independent-block probability at least $1-\delta$,

$$\Delta_{2aN_{\mathrm{blk}},S} \le 2\max\{\widehat{\mathfrak{R}}^{(a)}_{\mathrm{odd}}(\mathcal{F}_S), \widehat{\mathfrak{R}}^{(a)}_{\mathrm{even}}(\mathcal{F}_S)\} + 3\sqrt{\frac{\log(4|\mathfrak{S}|/\delta)}{2N_{\mathrm{blk}}}}$$

simultaneously for all S. The full equal-weight fresh deviation satisfies

$$\Delta_{J,S} \le \frac{2aN_{\mathrm{blk}}}{n_J}\Delta_{2aN_{\mathrm{blk}},S} + \frac{q}{n_J} \le \Delta_{2aN_{\mathrm{blk}},S} + \frac{2a}{n_J},$$

because each centered bounded-loss summand has absolute value at most one. Coupling back to the actual sequence reduces the success probability by at most $(2N_{\mathrm{blk}}-1)\beta(a)$. The deterministic ERM comparison in Theorem 1 multiplies the fresh deviation by $2\rho_t$, so the remainder $2a/n_J$ at the deviation level becomes $4\rho_t a/n_J$ at the risk level. The drift and localization steps are unchanged, proving Corollary 2. If $\beta(a) \le C_\beta e^{-c_\beta a}$, taking $a \asymp c_\beta^{-1}\log(C_\beta n_J/\delta)$ makes the coupling penalty $O(\delta)$ and yields $N_{\mathrm{blk}} \asymp n_J/\log(n_J/\delta)$.

D.5 Proof of Theorem 2

For squared loss, $f_t(x) = \mathbb{E}[Y \mid X = x]$ satisfies

$$R_t(h) = \mathbb{E}[(Y - f_t(X))^2] + \|h - f_t\|^2_{L_2(P_t)}.$$

If $\mathcal{G}_t(S) = \bar{h}_{t-1} + V_S$, then

$$B_t(S) = \inf_{v\in V_S} \|\bar{h}_{t-1} + v - f_t\|_2^2 = \|(I - \Pi_{V_S})(f_t - \bar{h}_{t-1})\|_2^2$$

by the Hilbert-space projection theorem. For standardized Gaussian blocks (X_1, X_2) with correlation ϱ, target change ΔX_2, and only X_1 active, the best proxy coefficient is $\varrho\Delta$ and

$$B_t(\{1\}) = \Delta^2 \operatorname{Var}(X_2 - \varrho X_1) = \Delta^2(1-\varrho^2).$$

D.6 BLOCK COHERENCE LEMMA

For a set S of size k, let $D_S = \mathrm{blkdiag}(\Sigma_m : m \in S)$ and $C_S = D_S^{-1/2}\Sigma_S D_S^{-1/2}$. The diagonal blocks of C_S are identity and every off-diagonal block has operator norm at most μ. For a block vector $v = (v_m)_{m\in S}$,

$$\begin{aligned}|v^\top(C_S - I)v| &\le \mu \sum_{m\neq j} \|v_m\|\|v_j\| \\ &\le \mu(k-1)\sum_m \|v_m\|^2.\end{aligned}$$

Hence

$$1 - (k-1)\mu \le \lambda_{\min}(C_S) \le \lambda_{\max}(C_S) \le 1 + (k-1)\mu.$$

Combining with $\lambda_- I \preceq D_S \preceq \lambda_+ I$ yields

$$\frac{A_t(S)}{\lambda_+[1+(k-1)\mu]} \le Q_t(S) \le \frac{A_t(S)}{\lambda_-[1-(k-1)\mu]}, \qquad A_t(S) = \sum_{m\in S}\|g_{m,t}\|_2^2. \tag{23}$$

This is a block-Gershgorin bound and is the only place where coherence enters Theorem 3.

D.7 PROOF OF THEOREM 3

Because the theorem assumes equal block dimension d, write $\widehat{a}_{m,t} = \sqrt{d}\,\widehat{d}_{m,t}$; multiplying every score and threshold by the same positive constant does not change Equation 13. Let $a_m = \|g_{m,t}\|_2$. By the reverse triangle inequality,

$$\big|\|g_{m,V} - g_{m,U}\|_2 - a_m\big| \le \|g_{m,V} - g_{m,t}\|_2 + \|g_{m,U}\|_2 \le \eta_{\mathrm{age}} + \eta_{\mathrm{ref}}.$$

Together with the empirical score error, $|\widehat{a}_{m,t} - a_m| \le \xi$. Since $a_m \le G$,

$$|\widehat{a}_{m,t}^2 - a_m^2| \le 2G\xi + \xi^2 = \zeta. \tag{24}$$

Let $k = \widehat{k}_t$. The threshold-then-top-K rule selects exactly the k largest empirical scores: if more than K exceed τ, it keeps the top K; if between one and K exceed τ, those are the global top k; if none exceed τ, the minimum-one fallback keeps the global maximum. Therefore $\widehat{S}_t$ maximizes $\sum_{m\in S}\widehat{a}_{m,t}^2$ over $|S| = k$. Comparing it with the joint-gain oracle $S_{t,k}^\star$ and using Equation 24,

$$A_t(\widehat{S}_t) \ge A_t(S_{t,k}^\star) - 2k\zeta. \tag{25}$$

The lower inequality in Equation 23 and the upper inequality applied to $S_{t,k}^\star$ give

$$\begin{aligned}Q_t(\widehat{S}_t) &\ge \frac{A_t(\widehat{S}_t)}{\lambda_+[1+(k-1)\mu]} \\ &\ge \frac{\lambda_-[1-(k-1)\mu]}{\lambda_+[1+(k-1)\mu]} Q_t(S_{t,k}^\star) - \frac{2k\zeta}{\lambda_+[1+(k-1)\mu]},\end{aligned}$$

which is Equation 16.

If $\mu = 0$ and $\kappa = 1$, joint gain is proportional to $A_t(S)$. If the kth and $(k+1)$st population gradient norms are separated by more than 2ξ, the top-k ordering is preserved. If, in addition, the population score on the intended side of the threshold is at least ξ away from the threshold, the empirical threshold produces the same k. The selected set then equals $S_{t,k}^\star$ exactly. These margin conditions also explain why small changes can fail in the stable-gradient experiment.

Whitened score. A covariance-whitened score $\|\widehat{\Sigma}_m^{-1/2}(\widehat{g}_m(V) - \widehat{g}_m(U))\|$ removes the within-block condition-number mismatch when $\widehat{\Sigma}_m$ is accurate, but it does not remove cross-block interaction. The released implementation uses the simpler standardized Euclidean score and reports conditioning as a boundary.

D.8 FRESH SCORE AND COVARIANCE CONCENTRATION

The fast-rate proof uses two weighted concentration lemmas. Throughout this subsection we condition on $\mathcal{H}_{t^-}$, so the feasible family, anchor, replay measure, and predictable fresh weights are fixed. Write $n_{\mathrm{eff}} = (\sum_s \widetilde{\omega}_{t,s}^2)^{-1}$ and $n_\infty = (\max_s \widetilde{\omega}_{t,s})^{-1}$.

Lemma 1 (Weighted score concentration). *Fix S and let $\psi_{s,S} \in \mathbb{R}^{d_S}$ be conditionally independent, conditionally centered vectors satisfying*

$$\sup_{\|u\|_2=1} \|u^\top \psi_{s,S}\|_{\psi_1} \le \nu_S \qquad \text{a.s. given } \mathcal{H}_{t^-}.$$

For normalized predictable weights $\widetilde{\omega}_{t,s}$ and every $x > 0$, there is a universal C such that

$$\Pr\left(\left\| \sum_s \widetilde{\omega}_{t,s} \psi_{s,S} \right\|_2 > C\nu_S \left[\sqrt{\frac{d_S + x}{n_{\mathrm{eff}}}} + \frac{d_S + x}{n_\infty} \right] \middle| \mathcal{H}_{t^-} \right) \le 2e^{-x}. \tag{26}$$

Consequently, with $x = \log(4|\mathfrak{S}|/\delta)$ and a union bound, the event holds simultaneously over $S \in \mathfrak{S}$ with conditional probability at least $1 - \delta$.

Proof. For a fixed unit vector u, scalar Bernstein for independent centered sub-exponential variables gives

$$\Pr\left(\left| \sum_s \widetilde{\omega}_{t,s} u^\top \psi_{s,S} \right| > C\nu_S \left[\sqrt{x \sum_s \widetilde{\omega}_{t,s}^2} + x \max_s \widetilde{\omega}_{t,s} \right] \middle| \mathcal{H}_{t^-} \right) \le 2e^{-x}.$$

Apply this bound on a $1/2$-net of the unit sphere with cardinality at most 5^{d_S} and use $\|v\|_2 \le 2 \max_{u \in \mathcal{N}} u^\top v$. Absorbing $d_S \log 5$ into the universal constant proves Equation 26. □

Weighted score-concentration consequence used in Theorem 3. For a routing window $W \in \{U_t, V_t\}$ and block m, define

$$\phi_{s,m} = X_{s,m}(Y_s - \bar{h}_{t-1}(X_s)) - \mathbb{E}[X_{s,m}(Y_s - \bar{h}_{t-1}(X_s)) \mid \mathcal{H}_{t^-}].$$

Under the weighted score-concentration assumptions stated for the routing analysis, $\sup_{\|u\|=1} \|u^\top \phi_{s,m}\|_{\psi_1} \le C K_X \sigma_{\mathrm{sc}} \sqrt{\overline{\lambda}_{\mathrm{sc}}}$. Lemma 1, applied to each of the two equal-weight windows and union-bounded over $2M$ block–window pairs, gives

$$\max_m \|\widehat{g}_m(W) - g_m(W)\|_2 \le C K_X \sigma_{\mathrm{sc}} \sqrt{\overline{\lambda}_{\mathrm{sc}}} \left[\sqrt{\frac{d + \log(4M/\delta)}{n_{\mathrm{sc}}}} + \frac{d + \log(4M/\delta)}{n_{\mathrm{sc}}} \right]$$

for both windows simultaneously, after absorbing the factor of two into C. The reverse triangle inequality then yields the same-order bound for $|\widehat{a}_{m,t} - \|g_{m,V} - g_{m,U}\|_2|$, proving the score-error control used in Theorem 3.

Lemma 2 (Weighted restricted-Gram concentration). *Fix S and let $X_{s,S}$ be conditionally independent, centered or uncentered sub-Gaussian covariates with conditional second-moment matrix $\Sigma_{s,S} = \mathbb{E}[X_{s,S} X_{s,S}^\top \mid \mathcal{H}_{t^-}]$. Assume the* uniform fresh covariance cap

$$\Sigma_{s,S} \preceq \overline{\lambda}_F I \quad \text{for every fresh } s, \qquad \|u^\top X_{s,S}\|_{\psi_2} \le K_X (u^\top \Sigma_{s,S} u)^{1/2} \quad \text{for every unit } u.$$

There is a universal C such that, for every $x > 0$, with conditional probability at least $1 - 2e^{-x}$,

$$\left\| \sum_s \widetilde{\omega}_{t,s} \{X_{s,S} X_{s,S}^\top - \Sigma_{s,S}\} \right\|_{\mathrm{op}} \le C K_X^2 \overline{\lambda}_F \left[\sqrt{\frac{d_S + x}{n_{\mathrm{eff}}}} + \frac{d_S + x}{n_\infty} \right]. \tag{27}$$

Consequently, if the hybrid population covariance satisfies $\Sigma_{t,S}^H \succeq \underline{\lambda}_H I$, then, simultaneously over $S \in \mathfrak{S}$, the fresh covariance deviation is at most $\underline{\lambda}_H / 2$ whenever

$$n_{\mathrm{eff}} \ge C K_X^4 \kappa_{F/H}^2 D_{\max,\delta}, \qquad n_\infty \ge C K_X^2 \kappa_{F/H} D_{\max,\delta}, \quad \kappa_{F/H} = \overline{\lambda}_F / \underline{\lambda}_H, \tag{28}$$

where $D_{\max,\delta} = d_{\max} + \log(4|\mathfrak{S}|/\delta)$. For equal weights, $n_{\mathrm{eff}} = n_\infty = n_J$.

Proof. For each fixed direction u, $(u^\top X_{s,S})^2 - \mathbb{E}[(u^\top X_{s,S})^2 \mid \mathcal{H}_{t^-}]$ is conditionally sub-exponential with scale at most $CK_X^2 u^\top \Sigma_{s,S} u \le CK_X^2 \overline{\lambda}_F$. Hence the Bernstein variance proxy is bounded by $CK_X^4 \overline{\lambda}_F^2 \sum_s \widetilde{\omega}_{t,s}^2$ and the maximal weighted scale by $CK_X^2 \overline{\lambda}_F \max_s \widetilde{\omega}_{t,s}$. Weighted scalar Bernstein followed by a $1/4$-net argument for symmetric matrices yields Equation 27; solving its square-root and linear terms against $\underline{\lambda}_H/2$ gives Equation 28. □

Proof of Theorem 4. Fix S and parameterize the reachable affine class as

$$h_{\theta,S}(x) = \bar{h}_{t-1,-S}(x_{-S}) + x_S^\top \theta, \qquad \widetilde{Y}_{s,S} = Y_s - \bar{h}_{t-1,-S}(X_{s,-S}),$$

so frozen-block contributions are absorbed into a residualized response. Let

$$L_S^H(\theta) = R_t^H(h_{\theta,S}) + \lambda\|\theta\|_2^2, \qquad \widehat{L}_S^H(\theta) = \widehat{R}_t^H(h_{\theta,S}) + \lambda\|\theta\|_2^2,$$

and let $\theta_{\lambda,S}^H$ and $\widehat{\theta}_S$ minimize these two strongly convex quadratic objectives. Define

$$\psi_{s,S} = X_{s,S}(\widetilde{Y}_{s,S} - X_{s,S}^\top \theta_{\lambda,S}^H) - \mathbb{E}[X_{s,S}(\widetilde{Y}_{s,S} - X_{s,S}^\top \theta_{\lambda,S}^H) \mid \mathcal{H}_{t^-}].$$

The uniform fresh covariance cap and the residual condition imply $\sup_{\|u\|=1} \|u^\top \psi_{s,S}\|_{\psi_1} \le CK_X\sigma\sqrt{\overline{\lambda}_F}$, so Lemma 1 gives, simultaneously over S,

$$\left\| \sum_s \widetilde{\omega}_{t,s} \psi_{s,S} \right\|_2^2 \le CK_X^2 \sigma^2 \overline{\lambda}_F \left\{ \frac{D_\delta(S)}{n_{\mathrm{eff},J,t}} + \frac{D_\delta(S)^2}{n_{\infty,J,t}^2} \right\}. \tag{29}$$

Lemma 2 and Equation 17 give

$$\left\| \sum_s \widetilde{\omega}_{t,s} \{ X_{s,S} X_{s,S}^\top - \mathbb{E}[X_{s,S} X_{s,S}^\top \mid \mathcal{H}_{t^-}] \} \right\|_{\mathrm{op}} \le \underline{\lambda}_H/2 \tag{30}$$

simultaneously over the feasible family. The empirical hybrid covariance differs from $\Sigma_{t,S}^H$ only through ρ_t times this fresh deviation, so $\rho_t \le 1$ and $\Sigma_{t,S}^H \succeq \underline{\lambda}_H I$ imply curvature at least $\underline{\lambda}_H/2$.

At $\theta_{\lambda,S}^H$, the replay gradients cancel exactly between $\widehat{L}_S^H$ and L_S^H, leaving

$$\nabla \widehat{L}_S^H(\theta_{\lambda,S}^H) - \nabla L_S^H(\theta_{\lambda,S}^H) = -2\rho_t \sum_s \widetilde{\omega}_{t,s} \psi_{s,S}.$$

Strong monotonicity therefore yields

$$\|\widehat{\theta}_S - \theta_{\lambda,S}^H\|_2 \le C \frac{\rho_t}{\underline{\lambda}_H} \left\| \sum_s \widetilde{\omega}_{t,s} \psi_{s,S} \right\|_2.$$

Let $\theta_{0,S}^H$ minimize the unregularized hybrid population risk over the same reachable affine space. For quadratic risk,

$$R_t^H(h_{\widehat{\theta}_S}) - R_t^H(h_{\theta_{0,S}^H}) = \|\widehat{\theta}_S - \theta_{0,S}^H\|_{\Sigma_{t,S}^H}^2 \le 2\|\widehat{\theta}_S - \theta_{\lambda,S}^H\|_{\Sigma_{t,S}^H}^2 + 2\|\theta_{\lambda,S}^H - \theta_{0,S}^H\|_{\Sigma_{t,S}^H}^2.$$

The second term is exactly $2A_{\lambda,t}^H(S)$. For the first, $\Sigma_{t,S}^H \preceq \overline{\lambda}_H I$ together with Equation 29 gives

$$R_t^H(\widehat{h}_S) - \inf_{h \in \mathcal{G}_t(S)} R_t^H(h) \le CK_X^2 \frac{\overline{\lambda}_H \overline{\lambda}_F}{\underline{\lambda}_H^2} \rho_t^2 \sigma^2 \left\{ \frac{D_\delta(S)}{n_{\mathrm{eff},J,t}} + \frac{D_\delta(S)^2}{n_{\infty,J,t}^2} \right\} + 2A_{\lambda,t}^H(S). \tag{31}$$

Finally, the same two-discrepancy comparison used in Theorem 1 gives

$$R_t(\widehat{h}_S) - \inf_{h \in \mathcal{H}} R_t(h) \le [R_t^H(\widehat{h}_S) - \inf_{h \in \mathcal{G}_t(S)} R_t^H(h)] + 2D_t^H(S) + B_t(S).$$

The concentration events are simultaneous over $\mathfrak{S}$, so S can be replaced by the data-dependent $\widehat{S}_t$. This proves Theorem 4.

D.9 Dependent-data supply of the fast-rate events

The preceding lemmas are stated for conditionally independent fresh observations. For dependent fresh data, the same proof is applied to coupled block averages, but the effective counts and coupling penalty are stated explicitly rather than hidden in an abstract N_t.

Corollary 4 (Geometric β-mixing supply for the fast rate)**.** *Assume equal fresh weights, conditional geometric mixing $\beta(a) \le C_\beta e^{-c_\beta a}$, and that the block-averaged score vectors and block-averaged Gram deviations obey the same conditional ψ_1 bounds as Lemmas 1–2, with the fresh covariance scale controlled by $\overline{\lambda}_F$ up to universal block-scale constants. Choose*

$$a = \left\lceil c_\beta^{-1} \log \frac{8 C_\beta n_J}{\delta} \right\rceil, \qquad N_{\mathrm{blk}} = \left\lfloor \frac{n_J}{2a} \right\rfloor.$$

If $N_{\mathrm{blk}} \ge C K_X^4 \kappa_{F/H}^2 D_{\max,\delta}$, then the score and restricted-Gram events required in the proof of Theorem 4 hold with failure probability at most $\delta + O(N_{\mathrm{blk}}\beta(a))$, and the estimation term has order

$$C K_X^2 \Gamma_{\mathrm{cov}} \rho_t^2 \sigma^2 \frac{D_\delta(S)}{N_{\mathrm{blk}}} = O\left(K_X^2 \Gamma_{\mathrm{cov}} \rho_t^2 \sigma^2 \frac{D_\delta(S) \log(n_J/\delta)}{n_J} \right). \tag{32}$$

Thus $N_{\mathrm{eff}}^{\mathrm{mix}} \asymp n_J / \log(n_J/\delta)$ for this explicit geometric specialization.

Justification. Couple alternating length-a blocks to independent copies as in Section D.4. Treat each block average as one super-observation and apply Lemmas 1 and 2 to the N_{blk} odd blocks and separately to the N_{blk} even blocks; a union bound combines the two parities and the feasible family. The chosen a makes $(2N_{\mathrm{blk}} - 1)\beta(a) = O(\delta)$. The displayed block-Orlicz condition is stated explicitly because arbitrary within-block dependence does not by itself preserve an i.i.d.-level sub-Gaussian constant. No claim beyond this declared block-level regularity is needed for the main fast-rate theorem.

D.10 Proof of Proposition 1

Let $p = Md_0$, $X \sim \mathcal{N}(0, I_p)$ and $Y = X^\top \theta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of X. For a fresh test point, prediction excess equals $\|\widehat{\theta} - \theta\|_2^2$. We construct a finite packing of block-r-sparse parameters.

A Varshamov–Gilbert support packing provides supports $S_1, \dots, S_N \subset [M]$, each of size r, with pairwise symmetric difference at least $r/2$ and

$$\log N \ge c_1 r \log(eM/r).$$

Within every chosen support, a binary code on the rd_0 active coordinates provides at least $\exp(c_2 r d_0)$ sign patterns with pairwise Hamming distance at least $rd_0/8$. Combining the support and sign codes gives a block-sparse packing $\{\theta^1, \dots, \theta^{N'}\}$ with

$$\log N' \ge c_3 \left(r d_0 + r \log \frac{eM}{r} \right), \qquad \|\theta^u - \theta^v\|_2^2 \ge c_4 b^2 r d_0, \tag{33}$$

when every nonzero coordinate has magnitude b; constants can be adjusted when the support and sign packings have different cardinalities.

For n i.i.d. observations, the covariate law is the same under all parameters and the conditional response is Gaussian. Therefore

$$\begin{aligned} \mathrm{KL}(P_{\theta^u}^n \| P_{\theta^v}^n) &= \frac{n}{2\sigma^2} \mathbb{E}[(X^\top(\theta^u - \theta^v))^2] \\ &= \frac{n}{2\sigma^2} \|\theta^u - \theta^v\|_2^2. \end{aligned}$$

Choose b so that the maximum pairwise KL is at most $(1/8)\log N'$; equivalently, the squared packing separation is of order

$$\frac{\sigma^2}{n} \left(r d_0 + r \log \frac{eM}{r} \right).$$

Fano's inequality then gives a constant lower bound on the probability of confusing two packing elements and hence the same order lower bound on expected squared prediction error. This proves Equation 34. The rate is the grouped analogue of the classical sparse-regression minimax lower bound (Raskutti et al., 2011).

D.11 Proof of Corollary 3

For grouped linear squared loss, the best current risk obtained by updating set S from the anchor is $R_t(\bar{h}_{t-1}) - Q_t(S)$. If the full linear oracle is reachable with all blocks,

$$B_t(S) = Q_t([M]) - Q_t(S).$$

Let $k = \widehat{k}_t$. Theorem 3 gives

$$Q_t(\widehat{S}_t) \geq \alpha_{\mu,k} Q_t(S^\star_{t,k}) - \frac{2k\zeta}{\lambda_+[1 + (k-1)\mu]}.$$

Hence

$$\begin{aligned} B_t(\widehat{S}_t) \leq Q_t([M]) - Q_t(S^\star_{t,K}) + [Q_t(S^\star_{t,K}) - Q_t(S^\star_{t,k})] \\ + (1 - \alpha_{\mu,k}) Q_t(S^\star_{t,k}) + \frac{2k\zeta}{\lambda_+[1 + (k-1)\mu]}. \end{aligned}$$

Substitute this bound for $B_t(\widehat{S}_t)$ in Theorem 4. The first term is unavoidable under a hard K-block budget, the second is the cost of threshold-induced under-use, the third is the coherence/conditioning approximation, and the fourth is routing estimation/reference/delay error. This proves Equation 19.

Proof summary. The history dependence of the anchor is handled by conditioning, the threshold/fallback rule is analyzed directly, cross-block interaction enters through μ, and the squared-loss result follows from the grouped-ridge concentration and strong-convexity argument above.

E Full Algorithm and Implementation Details

Scope. This section expands the executable pipeline summarized in Section 3.2. The pseudocode below uses the same loss-aware risk-change score, threshold/budget rule, and anchored restricted fit as the main text.

E.1 Executable DDA pipeline

At each scheduled partial-update round t:

1. **Reveal delayed labels.** Form $\mathcal{W}_t = \{s < t : s + d_s \leq t\}$.
2. **Construct windows.** Select $U_t, V_t, F_t, C_t \subseteq \mathcal{W}_t$ using only past indices and fixed/validation-selected lengths.
3. **Score risk change.** For every block,
$$D^{\text{risk}}_{m,t} = d_m^{-1/2} \|\widehat{g}_m(V_t; \bar{h}_{t-1}) - \widehat{g}_m(U_t; \bar{h}_{t-1})\|_2.$$
4. **Route the budget.** For equal-cost experiments, keep scores above threshold τ and truncate to the largest K. The measured-stream implementation keeps the single largest score when none cross, so every scheduled update opportunity performs at least one block update and the threshold controls whether additional blocks are included. This protocol choice is logged and analyzed explicitly by Theorem 3; a zero-update fallback is also admissible in the general feasible-family analysis because $\varnothing \in \mathcal{S}_{\leq K}$. For unequal costs, a score-per-cost/knapsack router can be substituted without changing Theorem 1.
5. **Fit the reachable class.** Freeze all omitted blocks at the anchor and minimize Equation 8 over $\mathcal{G}_t(S_t)$. In grouped linear regression this is ridge/least squares on selected columns after subtracting the frozen-anchor contribution.
6. **Calibrate optionally.** Compute delayed residuals on C_t. The adaptive row uses an upper residual quantile; the fixed-radius ablation uses the same predictor and selected blocks but disables adaptation of this quantile.
7. **Deploy and log.** Record selected blocks, score margins, update volume, prediction, radius, and wall-clock components, then commit the fitted predictor as the next anchor.

E.2 Information use, replay decomposition, and label reuse

At each update the runner records the last label index available at the previous update. Labels revealed after that index form the fresh set J_t. The configured fit window F_t may also contain earlier revealed records; those records are replay and are measurable with respect to $\mathcal{H}_{t^-}$. Uniform fitting over F_t is therefore exactly Equation 5 with $\rho_t = |F_t \cap J_t|/|F_t|$ and Q_t^H the normalized empirical measure on $F_t \setminus J_t$. The core logger writes `fresh_labels`, `replay_labels`, `fresh_weight`, and `replay_weight` at every update. The fresh-only variant sets $F_t = J_t$, hence $\rho_t = 1$. This realizes the measurability and replay decomposition used in Theorem 1: replay is fixed by the pre-update state and only newly revealed labels contribute fresh stochastic variation. It does not by itself verify the theorem's conditional-law or dependence assumptions: replay is fixed by the pre-update history and only newly revealed labels contribute fresh stochastic variation.

Routing may reuse fresh labels that also appear in the fit. The simultaneous event over $S \in \mathfrak{S}$ handles that reuse conditional on the anchor. Calibration is evaluated separately because its coverage claim concerns the radius rather than the regression-risk theorem.

E.3 Computational path

The algorithm does not enumerate $\binom{M}{K}$ subsets. Dense risk-change scores are vectorized over the selected window; equal-cost routing uses top-K selection, and the restricted fit touches only active columns plus the intercept. The $\log|\mathfrak{S}|$ term in Theorem 1 is a statistical uniformization price, not the runtime of a combinatorial search.

E.4 Optional one-step conservative decision diagnostic

For completeness, a one-step downstream action diagnostic can be formed from the calibrated prediction radius. When a downstream action $a \in \mathcal{A}_t$ has known base cost $c_t(a)$ and unknown current penalty $\mu_t(a)$ estimated by prediction $p_t(a)$ and radius $u_t(a)$, let $\gamma_{\mathrm{sw}}, \gamma_{\mathrm{pen}} \geq 0$ and define

$$J_t(a) = c_t(a) + \gamma_{\mathrm{sw}} \mathbf{1}\{a \neq a_{t-1}\} + \gamma_{\mathrm{pen}} \mu_t(a),$$

$$\widetilde{J}_t(a) = c_t(a) + \gamma_{\mathrm{sw}} \mathbf{1}\{a \neq a_{t-1}\} + \gamma_{\mathrm{pen}} [p_t(a) + u_t(a)].$$

If $|\mu_t(a) - p_t(a)| \leq u_t(a)$ for all actions and a_t minimizes $\widetilde{J}_t$, then for a current-cost oracle $a_t^\star$,

$$J_t(a_t) - J_t(a_t^\star) \leq 2\gamma_{\mathrm{pen}} \max_{a \in \mathcal{A}_t} u_t(a).$$

Indeed $J_t(a) \leq \widetilde{J}_t(a) \leq J_t(a) + 2\gamma_{\mathrm{pen}} u_t(a)$ and the result follows by optimality of a_t. No transition model, cumulative reward, or future safety dynamics enter. Invalid radius coverage therefore invalidates only this one-step interpretation, not Theorems 1–3.

Implication. The executable pipeline uses the same score, threshold/fallback rule, restricted fit, and optional calibration layer defined in the main text.

F Additional Assumptions and Scope

This section collects interpretation boundaries that are useful when applying the main results. The theorem assumptions themselves are stated in the main text and in Table 6; the points below explain how the analytic objects relate to the executable protocol.

F.1 Population oracles and executable quantities

The joint-gain quantity $Q_t(S)$ and the oracle set $S_{t,k}^\star$ use current-law population gradients and covariances. They are comparison objects for Theorem 3 and controlled simulations, not quantities assumed available to DDA. The executable router uses only the two delayed windows in Equation 12. Likewise, the exact localization bias $B_t(S)$ is an analytic current-risk quantity rather than an online certificate.

F.2 Conditional targets, dependence, and finite discrepancy

Theorems 1 and 4 use the conditional-law compatibility condition stated in Section 3.3. Conditional independence controls fresh stochastic variation but does not replace that compatibility assumption. Corollary 2 gives a bounded-loss route under conditional β-mixing. For unrestricted squared-loss classes, discrepancy can be infinite under covariance shift; the risk bounds should therefore be read on reachable classes for which the displayed mismatch is finite.

F.3 Covariance and routing scope

Theorem 4 separates the uniform fresh covariance cap $\overline{\lambda}_F$ from hybrid curvature $\underline{\lambda}_H$ and uses their ratio $\kappa_{F/H}$. Theorem 3 assumes equal block dimension, bounded coherence, conditioning, and explicit score margins. Current measured blocks have unequal dimensions, and their population coherence and error parameters are not estimated from deployment logs. Accordingly, the measured studies illustrate resource–error behavior under the executable protocol, while the controlled studies are used for direct mechanism tests of the quadratic theory.

F.4 Preprocessing and statistical interpretation

Numerical imputation, scaling, and within-block whitening statistics in the current measured-stream runners are fit on the warm-up prefix. The saved preprocessing pipeline fixes missingness-indicator columns by checking which columns contain missing values over the stored stream; this schema rule is shared by all compared methods and is documented so that the chronological experiment can be reproduced exactly. The paired bootstrap resamples stored chronological prediction-error trajectories and does not refit the learner within each resample; its intervals therefore quantify paired trajectory differences rather than prospective online confidence or equivalence.

F.5 Resource interpretation

The hard budget counts updated feature blocks or coordinates. Scoring can still read all candidate features, and the shared intercept is refit outside the feature-block count in the linear implementation. Update-volume reductions therefore should not be interpreted as proportional wall-clock or hardware speedups without an additional systems model.

G Additional Statistical Statements

The following dependent-data specialization and minimax lower bound complement the main-text results. Their proof ingredients appear in Appendix D.

Corollary 2 (Nonstationary β-mixing fresh labels)**.** *Suppose the fresh sequence is conditionally β-mixing given $\mathcal{H}_{t^-}$ with coefficient $\beta(a)$ and equal fresh weights. Write $n_J = |J_t|$, choose a block length a, set $N_{\mathrm{blk}} = \lfloor n_J/(2a) \rfloor \geq 1$, and let $q = n_J - 2aN_{\mathrm{blk}} < 2a$. Apply the independent-block coupling to the first $2aN_{\mathrm{blk}}$ observations and bound the q-sample remainder directly. Then the conclusion of Theorem 1 holds with the fresh deviation replaced by the maximum of the even/odd block Rademacher terms, an $O\Big(\rho_t \sqrt{\log(|\mathfrak{S}|/\delta)/N_{\mathrm{blk}}}\Big)$ concentration term, and an additional risk-level remainder at most $4\rho_t a/n_J$, with failure probability increased by at most $(2N_{\mathrm{blk}} - 1)\beta(a)$.*

Proposition 1 (Active-block minimax lower bound)**.** *In Gaussian linear regression with M orthogonal blocks of dimension d_0, noise variance σ^2, and an unknown parameter supported on at most r blocks, there are universal constants $c, c' > 0$ such that*

$$\inf_{\widehat{h}} \sup_{\|\operatorname{supp}_{\mathrm{block}}(\theta)\|_0 \leq r} \mathbb{E}\big[R(\widehat{h}) - R(h_\theta)\big] \geq c\,\frac{\sigma^2}{n}\left(rd_0 + r\log\frac{eM}{r}\right). \tag{34}$$

H Current Experiment Protocol and Evidence

This appendix supports Figures 4–5, Figure 7, and Tables 2 and 3. It records the current chronological runners and saved outputs used for the reported results. Every publication panel has a plotted-data export and a source/hash record in the experiment package.

H.1 CURRENT DATA AND CHRONOLOGICAL FITTING PROTOCOL

Appliances uses all 19,735 rows and 31 feature coefficients, grouped as rooms 1–3, rooms 4–6, rooms 7–9, weather, usage/time, and nuisance. Their dimensions are 6, 6, 6, 6, 5, and 2. Fjord5G uses 350,000 rows at source offset 5.3M and 32 coefficients grouped into motion (9), LTE link (12), NR/mode (9), and time (2). Exact feature names and missingness indicators are stored in the two `real_linear/*_meta.json` files.

Targets use the supplied logarithmic transformations. Medians, means, standard deviations, and within-block whitening are estimated on warm-up. Standardized values are clipped at 8 before and after whitening; whitening adds a 0.05 diagonal ridge. The indicator-column schema follows the saved stream-level missingness rule described in Appendix F; numerical imputation, scaling, and whitening parameters remain warm-up fitted.

The initial model uses the offline warm-up period. At scheduled index t, the first deployment interval has already been predicted. Available labels end at $t - d$; the Python window slices exclude that endpoint. Reference and recent windows are adjacent. After scores are computed at the same current anchor, the selected coefficients and common intercept are refit on the recent window, or its declared mixture with an older available window. Unselected coefficients are copied from the anchor. Predictions for the next interval use the committed model. Evaluation starts after warm-up; the final incomplete interval is included. This is delayed chronological evaluation, not a static train/validation/test split.

Table 7: Current linear headline settings, directly from saved metadata and runner defaults. The block budget is hard, not validation-selected.

Quantity	Appliances	Fjord5G
Warm-up rows	5,000	50,000
Evaluation rows	14,735	300,000
Delay (rows)	72	600
Update interval (rows)	144	5,000
Reference/recent window lengths	2,016 / 2,016	20,000 / 20,000
Block budget / total blocks	2 / 6	1 / 4
Ridge strength	0.05	0.05
Bootstrap block length	144	5,000
Bootstrap repetitions / seed	600 / 17	600 / 17
Random-K seeds	0–9	0–9

The default ridge and update settings are fixed in the runner; the archive does not establish a preregistered tuning search. GradMag uses recent gradient norm divided by square-root block dimension. Recent gain uses the local second moment plus ridge, and divides its quadratic gain by block dimension. DDA uses the norm of the gradient difference divided by square-root dimension. Ties are broken by block name. All candidate blocks are scored, so reducing the number fitted does not remove the cost of reading all features.

H.2 UNCERTAINTY AND EXACT SAVED RESULTS

For each bootstrap replicate, the released function samples contiguous error blocks with replacement, concatenates them, truncates to the evaluation length, and recomputes the RMSE difference. Reported endpoints are its 2.5% and 97.5% quantiles; the saved median is a bootstrap median, not the observed RMSE difference. All methods' prediction vectors are paired on the same rows. The original bootstrap was rerun from the saved prediction vectors for all eight comparisons, reproducing its endpoints and medians within floating-point tolerance. This is not an independent raw-data replication.

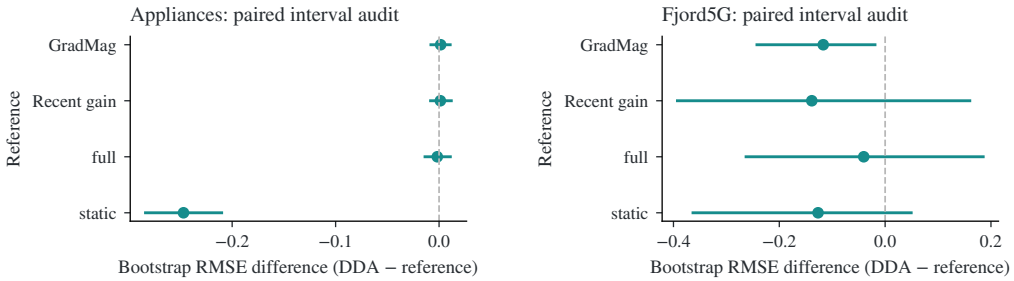


Figure 6: **Uncertainty qualifies the headline comparison.** Stored paired moving-block intervals for fixed prediction trajectories. The plotting layer uses the saved interval definition, not an equivalence threshold or a prospective confidence claim.

H.3 Controlled study designs and failure contrasts

Projection uses 20 seeds, 12,000 fitting observations, correlations $0, 0.2, 0.4, 0.6, 0.8, 0.95$, and unit omitted signal. Active dimension uses 30 seeds, two coordinates per block, two active blocks, $M = 8, 16, 32, 64$, 1,000 fitting and 5,000 test observations, unit signal/noise, and ridge 0.03. The oracle knows the active set; it is not implementable from the measured deployment logs. The anchoring experiment uses 150 seeds and contrasts anchored refitting against zero-reset omitted coefficients. Its default design has eight blocks, three coordinates per block, fitting size 600, test size 4,000, noise 0.5, change 0.7, and ridge 0.03.

The delayed-visibility experiment uses 150 seeds, eight blocks of four coordinates, fitting size 500, noise 0.8, change 0.6, and ridge 0.03. The hidden post-change fractions are $0, 0.25, 0.5, 0.75, 0.9$. The stable-error study uses 120 seeds, eight blocks of four coordinates, 500 observations in each reference/recent window, test size 3,000, noise 0.6, new change 0.65, $K = 1$, and ridge 0.03; persistent-error amplitudes are $0, 0.4, 0.8, 1.2, 1.6$. All raw and summary CSVs remain in `results/synthetic`. These are controlled model regimes, not measured estimates of population reliability.

Routing-geometry stress. To isolate the multiplicative geometry in Theorem 3, we use four Gaussian blocks of dimension two with $K = 2$, a zero anchor, 5,000 observations in each reference/recent window, outcome-noise standard deviation 0.5, and 100 independently sampled window pairs per setting. The reference population has zero anchored gradient and the recent law equals the current law, so the construction removes population reference and age error and varies only covariance geometry plus finite-window score noise. DDA selects from empirical gradient differences; Q_t and $S^{\star}_{t,2}$ are computed afterward from known simulation moments.

In the coherence sweep, each block has identity covariance, blocks 0 and 1 have cross-covariance μI_2, $\mu \in \{0, .1, \ldots, .8\}$, and $\kappa = 1$. Their population gradient norms are 1.00 and 0.98, while an independent third block has norm 0.90, making the two highest-score blocks increasingly redundant as μ grows. In the conditioning sweep, blocks are independent and have covariance $\mathrm{diag}(1, \kappa)$ for $\kappa \in \{1, 1.5, 2, 3, 4, 6, 8, 12, 16\}$; one high raw gradient lies along the high-variance coordinate, so the Q_t oracle increasingly discounts it. This construction is designed to expose the theorem's geometry, not to claim that empirical gain ratios must decrease monotonically in arbitrary data.

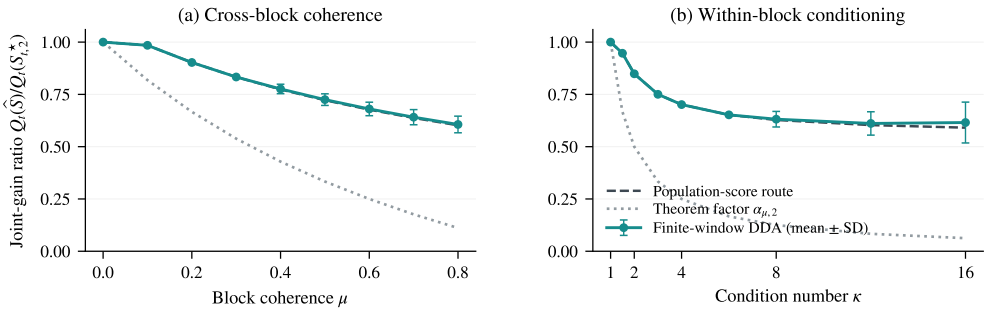


Figure 7: **Routing geometry weakens the joint-gain comparison in the controlled construction.** DDA uses only sampled reference/recent gradient differences, while the denominator $Q_t(S^\star_{t,2})$ uses known simulation population moments. Solid points are finite-window means $\pm$ one SD over 100 seeds; the dashed curve is the population-score route and the dotted curve is the theorem factor $\alpha_{\mu,2}$, shown as a comparison factor rather than a fitted prediction. (a) As coherence rises from 0 to 0.8 at $\kappa = 1$, the population route's joint-gain ratio falls from 1.00 to 0.602. (b) As κ rises from 1 to 16 at $\mu = 0$, it falls to 0.591 (finite-window mean 0.615).

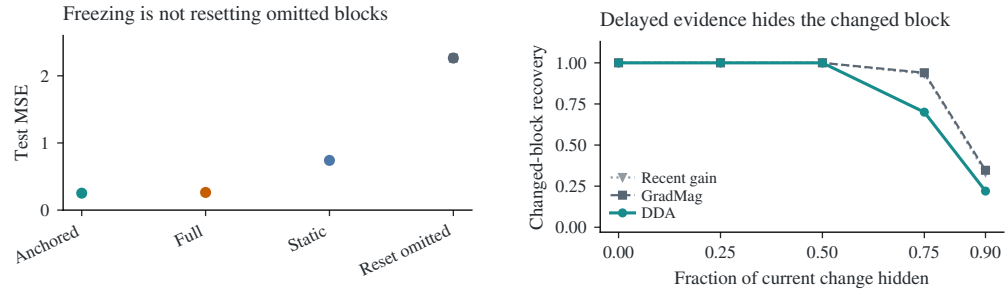


Figure 8: **An anchor and available change information play different roles.** The anchoring control tests preservation of omitted coefficients; the delay control tests active-set recovery as more changed observations remain hidden. Their seed aggregations are those of the original saved summaries.

H.4 Cross-segment and neural checks

Four disjoint 200,000-row Fjord5G segments begin at offsets 0, 2M, 4M, and 8M. Each uses a 50,000-row warm-up, a 600-row delay, 5,000-row updates, 20,000-row windows, $K = 1$, and ridge 0.05. These fixed row-index segments are distinct from the 350,000-row mixed-mode headline case. A favorable headline does not imply that the same selector is best across these segments.

The neural check uses a grouped additive MLP with hidden width 8, 300 pretraining steps, 10 update steps, and three seeds (0, 1, 2). Appliances uses 5,000 warm-up rows, interval 288, delay 72, windows 2,016, $K = 2$, learning rate 0.003, batch size 512, and 1,536 score samples. Fjord5G uses only 120,000 rows at offset 5.3M, 30,000 warm-up rows, interval 10,000, delay 600, windows 20,000, $K = 2$, learning rate 0.002, batch size 1,024, and 1,024 score samples. Thus neural and linear rows are not identical-budget or identical-horizon backbone replacements. GALA-style is an adapted selection criterion within this check, not the source method's complete test-time protocol.

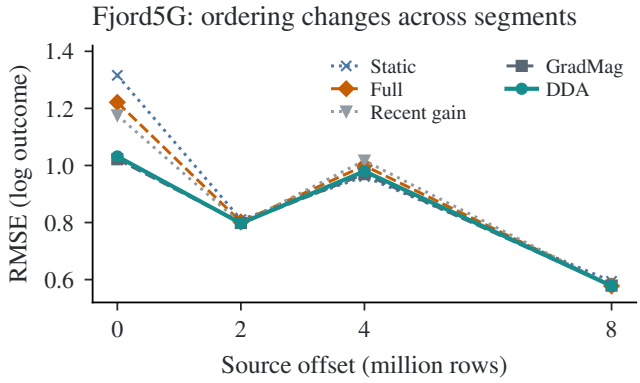


Figure 9: **Route rankings vary across segments.** Segment index is a descriptive regime axis rather than an independent randomization.

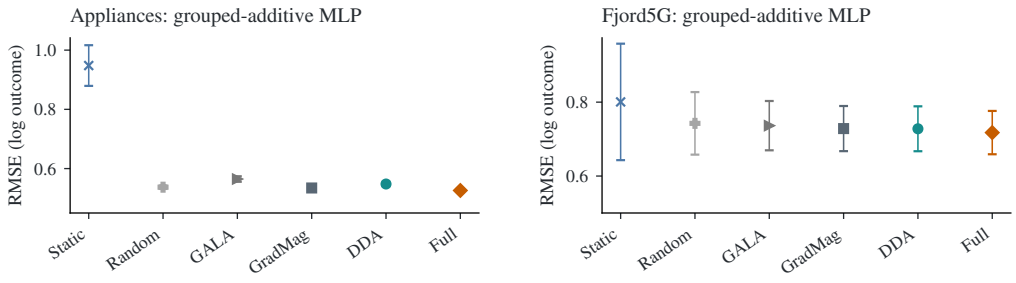


Figure 10: **Route rankings also vary under the neural backbone.** Means and standard deviations are over three seeds; the SDs are descriptive rather than formal superiority tests.

H.5 Sensitivity and replay grids

Appliances varies delay over 0, 72, 144, 288; recent-window length over 504, 1,008, 2,016, 4,032 while keeping reference length 2,016; and ridge over 0.01, 0.05, 0.2. Fjord5G varies delay over 0, 600, 1,200, 3,000; jointly varies both windows over 10,000, 20,000, 40,000, 60,000; and varies the same ridge grid. Each row uses the DDA route and the headline budget. The two window sweeps have different meanings, and large windows can delay eligibility. Saved update counts accompany every result.

The replay study mixes a recent delayed fit window with the immediately preceding available window, at older-window masses $0, 0.25, 0.5, 0.75$. The older window has length 2,016 on Appliances or 20,000 on Fjord5G. A second sweep fixes mass 0.5 and varies older-window length to 0.5, 1, or 2 times that length. Within each method the hard block budget is fixed. DDA's resulting scores and selections can nevertheless change as its anchor changes. The source column `fresh_weight` must be interpreted as recent-window mass, not an empirical measurement of the theorem's fresh-label fraction.

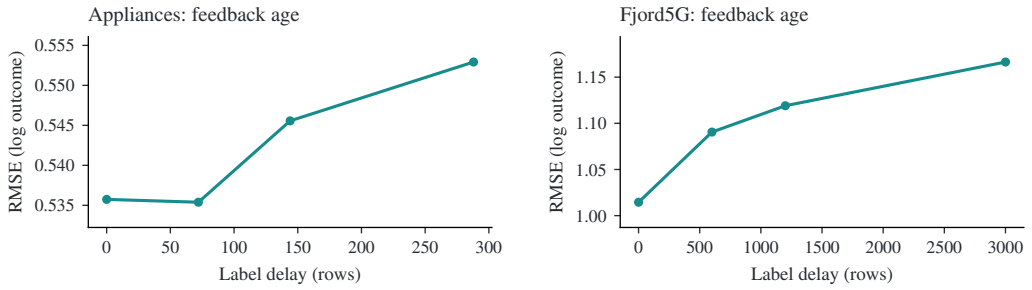


Figure 11: **Large delays degrade the recorded runs, while intermediate responses are non-monotone.** Points are the evaluated deterministic delay settings.

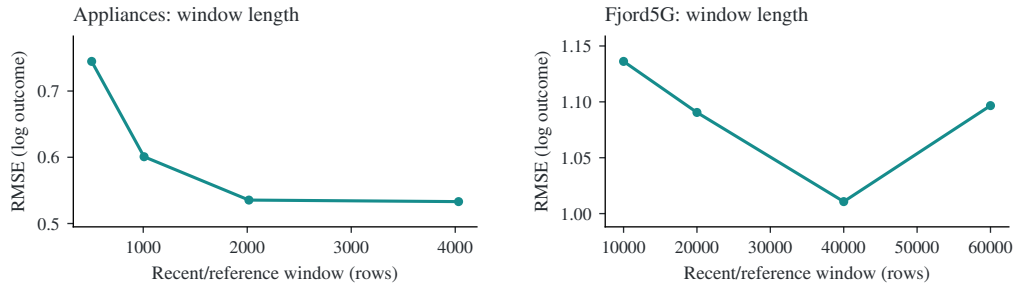


Figure 12: **Window length changes the evidence and can change eligibility.** Appliances varies only the recent window; Fjord5G varies both adjacent windows. Realized update counts are retained in the accompanying numeric tables.

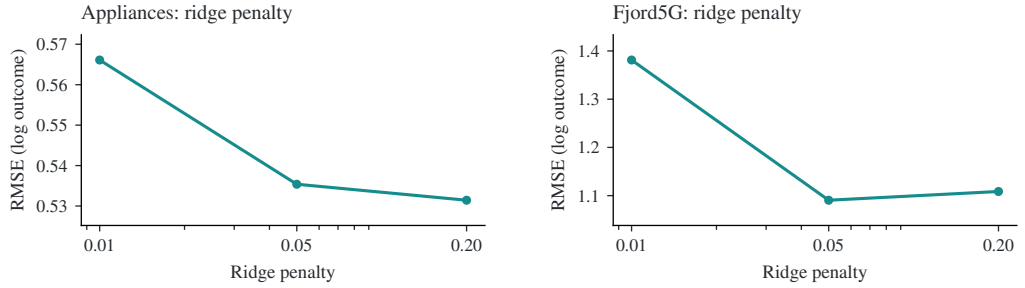


Figure 13: **Ridge sensitivity is a declared three-point check, not a validation-selected optimum.** All supplied regularization settings are plotted.

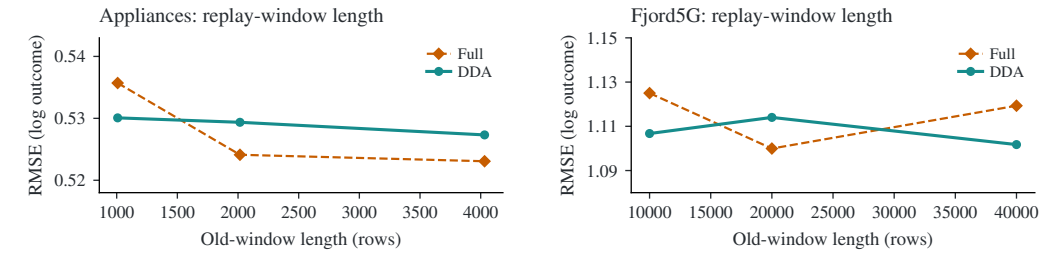


Figure 14: **Older-window length is a separate replay axis from its mixture mass.** These points fix older-window mass at 0.5 and use the three original length settings. They do not identify current-law discrepancy.

H.6 Cost and trajectory diagnostics

The current controlled timing file uses four coordinates per block, $K = 2$, sample size 4,000, and 15 timing repetitions at each $M = 8, 16, 32, 64, 128$. Touched feature coordinates are 8 for DDA and $4M$ for Full. This deterministic count does not include the shared intercept. Timing covers the original benchmark operation on its original execution environment; it should not be read as a portable GPU-throughput claim. The measured-stream timing disagreement is recorded in Appendix F, not replaced by this separate benchmark.

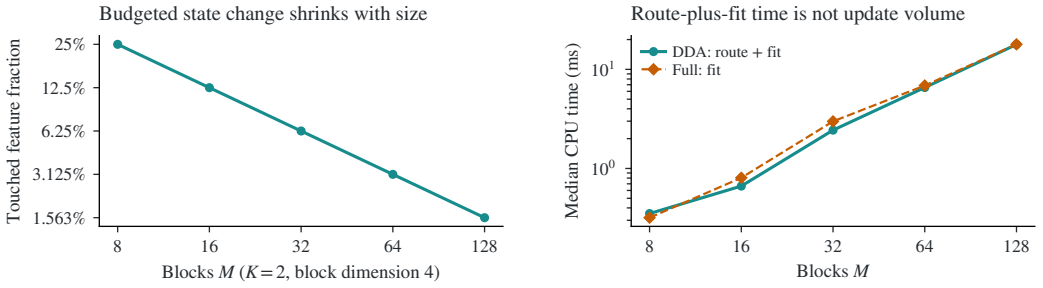


Figure 15: **Update volume and elapsed time scale differently.** Touched-coordinate fractions come from the selected dimension; times come from the saved controlled benchmark. No proportional CPU saving is claimed.

The Fjord5G trajectory display retains the original 5,000-row rolling analysis of squared prediction error and NR availability. These are descriptive rolling statistics, not uncertainty bands. Selection marks come directly from saved DDA update logs. The highlighted late regime, originally examined by the supplied plotting code, is not renamed a representative or randomly selected episode. Co-occurrence of a mode change, a block choice, and an error change is not a causal effect of that choice.

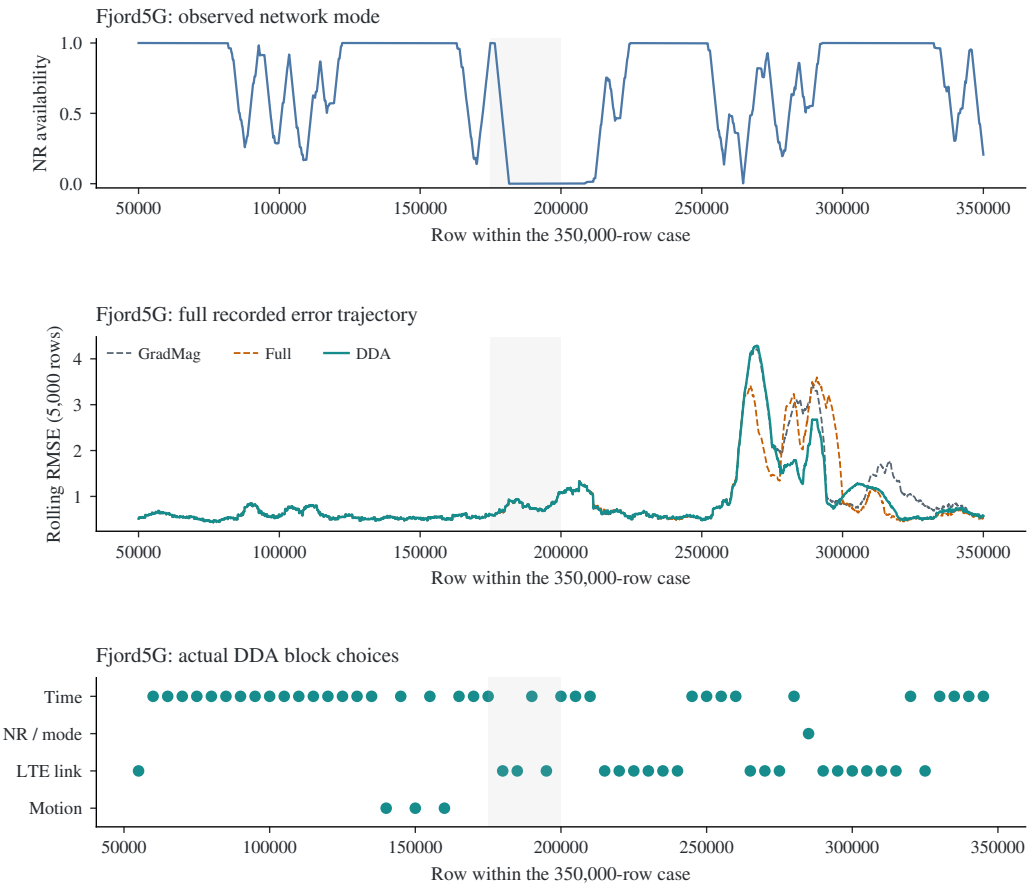


Figure 16: **Mode, prediction error, and selected blocks supply context, not causal attribution.** The first two panels retain the original 5,000-row rolling statistics on the complete fixed Fjord5G case; the third uses actual saved DDA update times and choices. Identical rolling operations are used across methods. Choices do not reveal the unobserved current-risk oracle. Each panel remains an independently reproducible figure file.

H.7 Current numeric ledgers

The tables below are generated from supplied CSVs, with display rounding only. Full-precision values and raw runs remain in the experiment archive. They supplement, rather than strengthen, the scope of the main-text comparisons.

Table 8: appliances current linear results.

Method	RMSE	Seed SD	MAE	Blocks/1k	Coords/1k
static	0.779926	0.000000	0.623076	0.000000	0.000000
random	0.541795	0.005632	0.391267	13.844588	70.885646
gradmag	0.534490	0.000000	0.381024	13.844588	79.877842
gain	0.534395	0.000000	0.384536	13.844588	78.520529
dda	0.535388	0.000000	0.386258	13.844588	82.253139
full	0.537274	0.000000	0.387978	41.533763	214.591110

Table 9: appliances full saved budget grid.

Method	K	RMSE	MAE	Blocks/1k
gradmag	1	0.542019	0.386718	6.922294
gain	1	0.543735	0.395748	6.922294
dda	1	0.544014	0.389853	6.922294
gradmag	2	0.534490	0.381024	13.844588
gain	2	0.534395	0.384536	13.844588
dda	2	0.535388	0.386258	13.844588
gradmag	3	0.537640	0.386658	20.766882
gain	3	0.541952	0.390765	20.766882
dda	3	0.543042	0.388872	20.766882
gradmag	4	0.534306	0.383643	27.689175
gain	4	0.539436	0.388780	27.689175
dda	4	0.539253	0.387939	27.689175
gradmag	5	0.536868	0.387316	34.611469
gain	5	0.537321	0.387905	34.611469
dda	5	0.537171	0.387703	34.611469
gradmag	6	0.537274	0.387978	41.533763
gain	6	0.537274	0.387978	41.533763
dda	6	0.537274	0.387978	41.533763
static	0	0.779926	0.623076	0.000000
full	6	0.537274	0.387978	41.533763

Table 10: appliances current sensitivity grid.

Axis	Value	RMSE	Blocks/1k
delay	0	0.535744	13.844588
delay	72	0.535388	13.844588
delay	144	0.545562	13.844588
delay	288	0.552919	13.844588
recent_window	504	0.744744	13.844588
recent_window	1008	0.600820	13.844588
recent_window	2016	0.535388	13.844588
recent_window	4032	0.533063	12.894469
lam	0.01	0.566105	13.844588
lam	0.05	0.535388	13.844588
lam	0.2	0.531434	13.844588

Table 11: appliances current replay grid.

Method	Sweep	Old mass	Old length	RMSE	Blocks/1k
dda	weight	0.00	0	0.535388	13.844588
dda	weight	0.25	2016	0.523708	13.844588
dda	weight	0.50	2016	0.529361	13.844588
dda	weight	0.75	2016	0.538643	13.844588
dda	window	0.50	1008	0.530078	13.844588
dda	window	0.50	2016	0.529361	13.844588
dda	window	0.50	4032	0.527326	13.844588
full	weight	0.00	0	0.537274	41.533763
full	weight	0.25	2016	0.522679	41.533763
full	weight	0.50	2016	0.524118	41.533763
full	weight	0.75	2016	0.534255	41.533763
full	window	0.50	1008	0.535703	41.533763

Method	Sweep	Old mass	Old length	RMSE	Blocks/1k
full	window	0.50	2016	0.524118	41.533763
full	window	0.50	4032	0.523073	41.533763

Table 12: appliances neural three-seed aggregate (different protocol from linear headline).

Method	RMSE	SD	MAE	Blocks/1k
dda	0.547869	0.007065	0.396094	6.922294
full	0.526064	0.005339	0.379353	20.766882
gala	0.564713	0.009314	0.415128	6.922294
gradmag	0.534471	0.008880	0.383923	6.922294
random	0.537130	0.005826	0.385743	6.922294
static	0.947907	0.068714	0.773588	0.000000

Table 13: fjord5g current linear results.

Method	RMSE	Seed SD	MAE	Blocks/1k	Coords/1k
static	1.229760	0.000000	0.586661	0.000000	0.000000
random	1.360140	0.168585	0.741036	0.196667	1.516000
gradmag	1.213114	0.000000	0.621382	0.196667	1.286667
gain	1.222727	0.000000	0.647682	0.196667	0.953333
dda	1.090513	0.000000	0.539046	0.196667	1.153333
full	1.136232	0.000000	0.567879	0.786667	6.293333

Table 14: fjord5g full saved budget grid.

Method	K	RMSE	MAE	Blocks/1k
gradmag	1	1.213114	0.621382	0.196667
gain	1	1.222727	0.647682	0.196667
dda	1	1.090513	0.539046	0.196667
gradmag	2	1.220220	0.592474	0.393333
gain	2	1.160310	0.579412	0.393333
dda	2	1.339492	0.688807	0.393333
gradmag	3	1.108199	0.558126	0.590000
gain	3	1.119040	0.561134	0.590000
dda	3	1.149519	0.574897	0.590000
gradmag	4	1.136232	0.567879	0.786667
gain	4	1.136232	0.567879	0.786667
dda	4	1.136232	0.567879	0.786667
static	0	1.229760	0.586661	0.000000
full	4	1.136232	0.567879	0.786667

Table 15: fjord5g current sensitivity grid.

Axis	Value	RMSE	Blocks/1k
delay	0	1.014537	0.196667
delay	600	1.090513	0.196667
delay	1200	1.119029	0.196667
delay	3000	1.166370	0.196667

Axis	Value	RMSE	Blocks/1k
window	10000	1.136196	0.196667
window	20000	1.090513	0.196667
window	40000	1.010724	0.176667
window	60000	1.096633	0.150000
lam	0.01	1.381208	0.196667
lam	0.05	1.090513	0.196667
lam	0.2	1.108606	0.196667

Table 16: fjord5g current replay grid.

Method	Sweep	Old mass	Old length	RMSE	Blocks/1k
dda	weight	0.00	0	1.090513	0.196667
dda	weight	0.25	20000	1.103313	0.196667
dda	weight	0.50	20000	1.114024	0.196667
dda	weight	0.75	20000	1.120404	0.196667
dda	window	0.50	10000	1.106733	0.196667
dda	window	0.50	20000	1.114024	0.196667
dda	window	0.50	40000	1.101739	0.196667
full	weight	0.00	0	1.136232	0.786667
full	weight	0.25	20000	1.096618	0.786667
full	weight	0.50	20000	1.099946	0.786667
full	weight	0.75	20000	1.112332	0.786667
full	window	0.50	10000	1.125005	0.786667
full	window	0.50	20000	1.099946	0.786667
full	window	0.50	40000	1.119308	0.786667

Table 17: fjord5g neural three-seed aggregate (different protocol from linear headline).

Method	RMSE	SD	MAE	Blocks/1k
dda	0.727936	0.060773	0.402438	0.155556
full	0.717593	0.058643	0.386588	0.311111
gala	0.736395	0.066729	0.413423	0.155556
gradmag	0.728475	0.061093	0.403843	0.155556
random	0.742587	0.084656	0.423185	0.155556
static	0.800640	0.157605	0.487710	0.000000

Table 18: All four cross-segment RMSEs.

Offset	Static	GradMag	Gain	DDA	Full
0	1.315917	1.021830	1.175740	1.031648	1.221238
2000000	0.810468	0.798417	0.796845	0.797253	0.802189
4000000	0.961364	0.969626	1.017136	0.979802	0.999339
8000000	0.594078	0.578837	0.577703	0.577928	0.577762

Table 19: Current controlled timing and touch counts (four coordinates per block).

M	DDA ms	Full ms	DDA coords	Full coords	Fraction
8	0.350727	0.318590	8	32	0.250000
16	0.664099	0.805437	8	64	0.125000
32	2.436310	2.987452	8	128	0.062500
64	6.556191	6.865658	8	256	0.031250
128	17.824839	17.886699	8	512	0.015625