DETECTOR-COUPLED SURROGATE REGRET FOR TEST-TIME ADAPTATION IN PIECEWISE-STATIONARY CMDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Constrained reinforcement-learning policies deployed in provisioning and finance must determine not only how to adapt under environmental shifts, but also when adaptation is warranted. Keeping a policy frozen can reduce reward and move its cost behavior away from the levels it was trained to satisfy, whereas indiscriminate test-time adaptation can overfit transient noise. We study this timing problem in piecewise-stationary constrained Markov decision processes (CMDPs) with imperfect change-point scores. Existing detection-and-restart analyses price detection delay and false alarms but omit constraints, while online-CMDP analyses price constraint costs but assume no external detector. Neither targets the timing decision. We propose *AdaptiveCMDP*, which closes both gaps: constraint exposure is fixed before deployment, and detector error is priced inside the regret bound. Specifically, it leaves the trained policy frozen and adapts only a bounded adapter whose radius, set from the Slater margin, fixes that exposure. Updates are reward-free, so adaptation can start before returns are informative; each detector firing restarts the step-size clock. The surrogate dynamic-regret bound then prices the detector directly—a *detector tax* that is square-root in horizon, regime count and expected false alarms, plus linear total detection delay. Under explicit value sufficiency and realizability, this transfers to policy value on an adapter-independent target stream. A frozen exploitability–delay rule scores every registered window: it correctly keeps the policy frozen in 15 of 16 financial panels where adaptation does not pay, and fires on all three provisioning domains—public traffic, bike-sharing and electricity-load data. There a capacity budget replaces the financial drawdown limit; strict regret wins exclude zero, at the cost of more stale-capacity breaches in two. A matching lower bound shows these terms are unavoidable: detector quality, not adapter capacity alone, decides when adaptation pays.

1 INTRODUCTION

Many reinforcement-learning applications impose hard operational constraints: a trading book must respect a drawdown limit, a network provisioner must hold a service floor, and a controller must stay inside a certified envelope (Efroni et al., 2020; Zhu et al., 2025). These environments need not stand still; transitions or objectives can change without announcement (Cheung et al., 2020). A policy fit to the old regime can then fail in both directions at once: it forfeits reward it could still earn, and its realized costs drift from the levels it was tuned to hold. A portfolio allocator tuned to calm markets loses reward after a market turn and, on adverse paths, breaches its drawdown limit; a provisioner calibrated to baseline demand overspends when traffic surges. Retraining from scratch is too slow to matter in the window that follows, while leaving the policy fixed lets it go stale, and updating it continuously can chase stationary noise (Sun et al., 2020; Niu et al., 2023). When the regime changes, how should a constrained policy adapt at deployment time?

Existing responses to non-stationarity fall into three broad strategies. Reset methods discard old estimates and relearn after a suspected change (Auer et al., 2019); continual methods update a policy, value function, or model from each new reward, cost, or transition (Zhu et al., 2025); auxiliary-

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

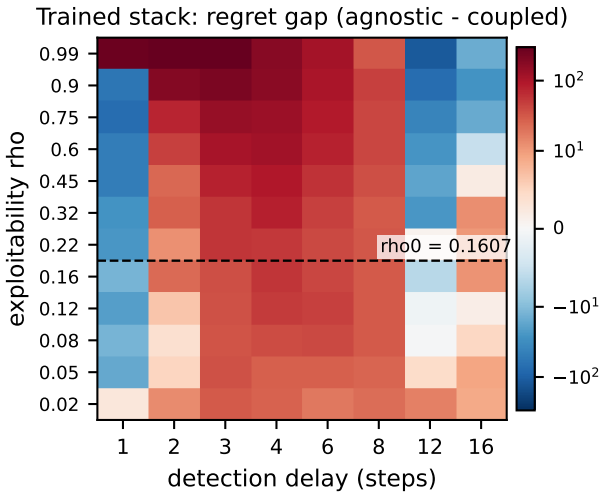

Figure 1: Detection delay shrinks the post-shift opportunity a well-timed adapter can recover: early and late triggers lose the gain, producing an interior timing region. The heat map reports agnostic-minus-coupled regret on the trained stack (`tc-phase2d-fullstack`: 12 exploitability levels $\times$ 8 delays, 30 seeds per cell; positive favours coupling; symmetric-log scale); gains grow with signal concentration ρ. The violation gap is zero inside the predicted-win band. These AdamW cells test the frozen empirical timing rule.

prediction methods use dense signals such as next-step prediction before reward estimates stabilize (Hansen et al., 2021). A detector can sit above these mechanisms and decide when an update or restart is warranted (Gerogiannis et al., 2026). This timing matters because adaptation itself can fit stationary noise, while delayed adaptation leaves less horizon in which to recover post-shift value.

For detector-based deployment adaptation, the signal and timing questions are distinct. Next-step prediction determines what drives the update; the detector gates the moments at which that signal is allowed to move the policy. Prediction accuracy affects reward only through the policy change it induces, so a useful auxiliary signal applied at a poor time can still hurt. Fire on the first flicker and a thin post-shift sample can push the adapter the wrong way; wait for high-confidence evidence and almost no horizon is left for the update to repay its cost. Figure 1 previews this trade-off: evidence becomes more reliable with delay as the recoverable opportunity shrinks (Lorden, 1971), producing an interior timing region rather than an "earlier is always better" rule.

A useful analysis must therefore quantify four quantities jointly: how late the detector fires, how often it fires falsely, how exploitable the prediction signal is, and how much post-shift value remains to be recovered. Existing pieces cover detector-free online-CMDP guarantees (Zhu et al., 2025; Stradi et al., 2024b), best-of-both-worlds analyses (Stradi et al., 2024a; Wei & Luo, 2021), piecewise-stationary restart guarantees (Auer et al., 2019), and self-supervised test-time adaptation (Sun et al., 2020; Hansen et al., 2021) and low-rank adaptation (Yun, 2024; Xu et al., 2025). Their guarantees do not quantify the restart cost of a deployment-time detector on the test-time surrogate.

AdaptiveCMDP formalizes this middle ground with a bounded last-layer adapter around a frozen base: detector firings restart its projected-MD clock, and the radius controls excess expected constraint cost. The experiments instantiate the same timing idea with bilinear LoRA and AdamW. We report three findings. (1) The convex-surrogate bound separates delayed switches from false restart events; in the realizable case, a conditional transfer yields a value bound with linear total delay and a $T^{3/4}(R+1+TP_{\mathrm{FA}})^{1/4}$ term. The radius condition $B \le \zeta/L$ keeps expected per-step cost inside budget. The delay lower bound holds in the reward-free observation model. (2) A signal-to-noise rule, calibrated on a controlled sweep and frozen before the designated evaluation runs, marks a band in the (ρ, D_d) plane where coupling is predicted to help; we call these cells *in-band*. This rule is empirical. (3) The frozen rule is calibrated in both directions: it abstains on the daily financial panels, where 15 of 16 predictions hold, and fires on every provisioning domain. The two calendar-

anchored cells reach the region but remain underpowered; the protocol-specified provisioning study wins strictly in 3 domains, at the cost of more violations in two. Post-hoc extensions rank each completed episode's scores from a prediction-residual ridge and a Bayesian run-length model, then select an equal-budget event set. They preserve the strict win on traffic and electricity and lose on bike at its shortest episode, but the retrospective ranking is not a deployable causal trigger.

The paper keeps three layers separate and evaluates each on its own terms. The guarantee belongs to an analyzed procedure: restart-aware projected MD on a convex next-step surrogate (§3–4; proofs in the supplement). The surrogate bound is unconditional; the value transfer is its conditional corollary. The timing layer is an empirical hypothesis: the frozen exploitability–delay rule (§6). The deployable implementation, bilinear LoRA with AdamW, is evaluated empirically; the theorem does not cover it, and the analyzed-schedule study on E1 (supplement) bridges the two and states its limits.

2 Related work

Non-stationary and constrained online RL. Online-CMDP guarantees, starting from Efroni et al. (2020), control regret and violation without a detector: Zhu et al. (2025) obtain $\widetilde{\mathcal{O}}(\sqrt{K})$ regret and strong violation under anytime-adversarial costs with no Slater point, Yu et al. (2026) give $\widetilde{\mathcal{O}}(K^{3/4})$ regret and violation for adversarial linear CMDPs, and Stradi et al. (2024b) give static regret under corruption; best-of-both-worlds and primal-dual analyses cover stochastic-versus-adversarial constraints and non-stationary objectives (Stradi et al., 2024a; Ding & Lavaei, 2023; Wei et al., 2023; Müller et al., 2024). On the unconstrained side, restart algorithms attain the minimax piecewise-stationary rate without knowing the number of switches (Auer et al., 2019; Mao et al., 2021; Besbes et al., 2015); variation-budget and predictive-representation analyses track non-stationarity in value or feature space (Cheung et al., 2020; Wang et al., 2026; Wei & Luo, 2021). None is a dynamic-regret result against a per-regime oracle indexed by detector quality: the switch structure is taken as given, and the detector that must discover it at deployment is left outside the analysis.

Detection-augmented non-stationary RL. The closest line puts the change-point detector inside the algorithm: Gerogiannis et al. (2026) give a detection-augmented wrapper for piecewise-stationary RL with minimax lower bounds; Alami et al. (2023) place a restarted Bayesian detector with explicit false-alarm and delay guarantees inside a UCRL2 controller; and Gerogiannis et al. (2025) show quickest-change-detection restarts beat black-box restarts over practical horizons. DARLING makes delay and a false-alarm level explicit for unconstrained restarts; Zhang & Zheng (2026) use a Bayesian detector to retune conservatism. Our contribution is narrower and orthogonal: it counts false firings and delay for a reward-free bounded last-layer surrogate, then states the additional assumptions needed for expected cost and CMDP value. It is not the first detector-quality regret result in RL, and LoRA remains empirical.

Table 1: Closest positioning. AdaptiveCMDP's formal result is the detector-priced projected-MD surrogate rate; LoRA/AdamW is a separate empirical timing implementation. Full comparison: supplement.

	Detector-priced surrogate	TTA implementation	Constraint scope
Detection-augmented PS-RL (Gerogiannis et al., 2026; 2025)	restart	policy learner	unconstrained
Online CMDP (Zhu et al., 2025; Stradi et al., 2024b)	no detector	policy learner	regret/violation
Self-supervised TTA (Hansen et al., 2021; Chen et al., 2025)	no detector rate	general TTA	none
AdaptiveCMDP (ours)	**yes (linear PMD)**	**LoRA/AdamW (emp.)**	**conditional value; exp. cost**

Test-time adaptation and low-rank updates. The update mechanism draws on test-time training with self-supervision (Sun et al., 2020; Hansen et al., 2021; Zhou et al., 2025; Bagatella et al., 2026), entropy minimization (Wang et al., 2021), and low-rank adapters used offline (Yun, 2024) or at test time (Xu et al., 2025); always-on adaptation degrades under online or mixed shifts without update selection (Niu et al., 2023). Provable test-time safety in CMDPs has been studied in the meta-RL setting, where a safe prior mixes with trained policies over a task distribution (Ni & Kamgarpour, 2026). The contrastive term follows van den Oord et al. (2018). Our formal object is the detector-controlled restart clock for a convex next-step predictor.

3 SETTING AND DETECTOR-COUPLED FRAMEWORK

A piecewise-stationary CMDP $(P_i, r_i, c_i)_{i=0}^{R}$ switches at unknown times τ_i, with at least $d_{\min}$ steps between switches. The learner sees (s_t, a_t, c_t, ξ_t) but not the regime label. Theorem 2 bounds expected next-step surrogate regret against each regime's best in-ball predictor; Corollary 4 states the extra conditions that transfer it to value regret against the best fixed feasible policy. Constraint excess $\mathrm{Vio}(T)$ is measured. Rewards and costs are bounded (A1), and policy-induced expected per-step cost is uniformly L-Lipschitz in the adapter (A2). Updates stay in the Frobenius ball $\mathcal{B} = \{\Delta\theta : \|\Delta\theta\|_F \le B\}$ (A3); the rank-r structure belongs to the LoRA implementation, while the analysed set is this convex ball. The remaining assumptions are a Slater margin $\zeta > 0$ (A4), a base-policy gap ϵ_0 (A5), detector parameters (D_d, P_{FA}) (A6), and $R \le T/d_{\min}$ (A7), and a frozen base that is feasible with the (A4) margin (A8). The regret bounds use (A1), (A2′), (A3), and (A6); Proposition 3 adds (A2) and (A8); value transfer adds the explicit value-sufficiency condition in the supplement; (A4), (A5), and (A7) are CMDP context.

The formal analysis uses a restart-aware projected-MD procedure with detector-delay parameter D_d and false-alarm rate P_{FA}. Each detector firing resets the step-size clock on the convex next-step loss. For bookkeeping, a switch's first firing before the next switch is its matched alarm and defines its delay; all other firings are false. The detector-regularity and segment-separation assumptions require each matched alarm to occur before the next switch, so the resulting post-switch delay windows are disjoint.

Assumption 1 (Surrogate convexity (A2′))**.** *In regime i, the expected next-step negative log likelihood ℓ_i is convex on $\mathcal{B}$. Let $\mathcal{F}_{t-1}$ contain the history and detector decisions available before the step-t target is observed. The iterate and restart decision are $\mathcal{F}_{t-1}$-measurable, and each sample supplies g_t with $\|g_t\|_2 \le G$ and $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] \in \partial\ell_i(\Delta\theta_t)$. The target stream is adapter-independent in every evaluated domain; this is a property of the domain's dynamics (the observation stream does not respond to the adapter), not of the architecture. Convexity of ℓ_i, by contrast, is exact for a linear last-layer adapter. Value transfer additionally requires the explicit value-sufficiency condition in the supplement.*

AdaptiveCMDP. A frozen base policy π_{θ_0} is augmented with low-rank LoRA parameters $\Delta\theta$, so only a small adapter budget updates online rather than the full model. A transformer change-point head produces a score $\xi_t \in [0, 1]$ from recent observations alone; that score both signals regime change and gates the adaptation rate.

The online update is reward-free. At each step, the method minimises a composite auxiliary loss built from (i) a next-step group negative log-likelihood term, (ii) a detector-aligned CPD stress penalty, and (iii) an optional contrastive term on normalised encoder projections. Only the next-step term is used in the theorem. Let $n_t = \max\{1, t - \tau_{\mathrm{last}} + 1\}$ count steps since the latest detector firing τ_{last}, so the clock reads 1 on the firing step itself and advances on every step thereafter. The analysed restart-aware update is projected online gradient descent (Zinkevich, 2003; Hazan, 2016) with a clock reset at each firing,

$$\Delta\theta_{t+1} = \Pi_{\mathcal{B}}\left[\Delta\theta_t - \frac{D}{G\sqrt{n_t}} g_t\right].$$

The trained-stack experiments instead apply the continuous score gate $\alpha_0\sigma(c\xi_t)$ with AdamW. They test the timing prediction. The projected-MD theorem does not cover that optimizer trajectory. Constraint breaches are counted directly.

4 MAIN RESULTS

Theorem 2 (Detector-coupled surrogate regret)**.** *For regime i, let $\ell_i(\Delta\theta) = \mathbb{E}[-\log p_{g_\phi}(Y; \Delta\theta)]$, where the next-step target Y is treated as a supervised target whose law does not depend on $\Delta\theta$, and choose $\Delta\theta_i^\star \in \arg\min_{\Delta\theta\in\mathcal{B}} \ell_i(\Delta\theta)$. Let $\mathcal{I}_i$ be its time indices and define*

$$\mathrm{Reg}_{\mathrm{ssl}}(T) = \sum_{i=0}^{R} \sum_{t\in\mathcal{I}_i} \big(\ell_i(\Delta\theta_t) - \ell_i(\Delta\theta_i^\star)\big).$$

Suppose ℓ_i is convex on a set $\mathcal{B}$ of diameter D (for the Frobenius ball of (A3), $D = 2B$), and each sample supplies a subgradient estimate g_t with $\|g_t\|_2 \le G$. Let $\mathcal{F}_{t-1}$ contain the history

and detector decisions available before the step-t target is observed. Suppose $\Delta\theta_t$ and the step-t restart decision are $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] \in \partial\ell_i(\Delta\theta_t)$. If the projected-MD clock restarts at every detector firing, use the following matching convention. Let the switches occur at $\tau_1 < \cdots < \tau_R$ and set $\tau_{R+1} = T + 1$. For each switch i, the first firing $a_i \in [\tau_i, \tau_{i+1})$ is its unique matched firing, its delay is $d_i = a_i - \tau_i$, and the delay window is $W_i = [\tau_i, a_i) \cap \mathbb{Z}$. Assume segment separation is sufficient that these matched firings exist before the next switch (equivalently, $\tau_i + d_i < \tau_{i+1}$); since each d_i is random, the bound below is stated conditional on this event (equivalently, it holds with the probability that the separation condition does). Every firing other than the a_i is false; if F is their number, assume $\mathbb{E}[d_i] \le D_d$ for every i and $\mathbb{E}[F] \le TP_{\mathrm{FA}}$. The disjoint windows W_i may be deleted in the analysis and have total length $\sum_i |W_i| = \sum_i d_i$. Then

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{ssl}}(T)] \le DG\left[2\sqrt{T\big(R + 1 + TP_{\mathrm{FA}}\big)} + RD_d\right] =: B_{\mathrm{ssl}}(T). \tag{1}$$

The result is exact for a convex last-layer adapter. It is not a guarantee for the bilinear LoRA/AdamW trajectory used in the experiments.

Proposition 3 (Adapter radius as an exposure dial)**.** *Under (A1), (A2), (A3) and (A8), the positive excess of expected per-step cost is bounded linearly in the adapter radius B of (A3):*

$$\sup_t \big[\mathbb{E}\, c_t(\Delta\theta_t) - \mathrm{budget}\big]_+ \;\le\; [L\,B - \zeta]_+ \;\le\; L\,B.$$

Hence $B \le \zeta/L$ keeps expected per-step cost inside the budget in every regime: the ball radius is an expected-cost exposure dial whose exchange rate is linear. No sublinear cumulative-violation bound follows; cumulative excess is reported as a measured quantity throughout.

Corollary 4 (Conditional surrogate-to-value transfer)**.** *Write $p_Y(\cdot\,; \Delta\theta) = p_{g_\phi}(\cdot\,; \Delta\theta)$ for the predicted next-step target law. Assume it is value-sufficient: for some $C_V < \infty$ and every $\Delta\theta, \Delta\theta' \in \mathcal{B}$,*

$$|V_i(\Delta\theta) - V_i(\Delta\theta')| \le C_V\, \mathrm{TV}(p_Y(\cdot\,; \Delta\theta), p_Y(\cdot\,; \Delta\theta'))\,.$$

Let $\varepsilon_i^{\mathrm{pred}} = \ell_i(\Delta\theta_i^\star) - H(p_{\mathrm{true}}^{(i)})$, let $\pi_i^\star$ be the best feasible in-ball policy, and define $A_T = \sum_i T_i[V_i(\pi_i^\star) - V_i(\Delta\theta_i^\star)]_+$ and $\mathrm{Reg}_{\mathrm{val}}(T) = \sum_i \sum_{t \in \mathcal{I}_i} (V_i(\pi_i^\star) - V_i(\Delta\theta_t))$. Let V_{span} bound the difference of any two policies' values ($V_{\mathrm{span}} \le 2R_{\max}/(1-\gamma)$ under (A1)). Then

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_{\mathrm{val}}(T)] \le\; & A_T + V_{\mathrm{span}} RD_d \\ & + 2C_V\sqrt{DG}\cdot T^{3/4}\cdot (R + 1 + TP_{\mathrm{FA}})^{1/4} + 2C_V \sum_i T_i \sqrt{\varepsilon_i^{\mathrm{pred}}}.\end{aligned} \tag{2}$$

When the predictive comparators are realizable and induce the best feasible in-ball policies, the approximation terms in Corollary 4 vanish. Its rate becomes $\widetilde{\mathcal{O}}(V_{\mathrm{span}} RD_d + C_V\sqrt{DG}T^{3/4}(R + 1 + TP_{\mathrm{FA}})^{1/4})$. Proposition 3 records the per-step expected-cost bound the assumptions do support: positive excess of expected cost, not expected positive-part violation. No sublinear cumulative violation bound follows, so constraint excess is reported as a measured quantity.

Theorem 2 gives the surrogate rate for restart-aware projected MD; we call its delay and false-alarm terms the *detector tax*. Corollary 4 shows why a square-root Pinsker–value-sufficiency map changes a $\sqrt{T}$ surrogate term into a $T^{3/4}$ value term. Predictive misspecification and mismatch with the best feasible policy remain explicit.

Theorem 5 (Scoped delay and surrogate lower bounds)**.** *Let $K = R + 1$, suppose $T \ge K\max\{2, d_{\min}\}$, and set $n = \lfloor T/K \rfloor$ and $q = T - Kn$. In both constructions, take K consecutive regime blocks of lengths $n, \ldots, n, n + q$. For a feasible delay satisfying $0 \le D_d$ and $\lceil D_d \rceil < d_{\min}$, for any learner that observes (s_t, a_t, c_t, ξ_t) but no rewards or regime labels, there is a bounded, zero-cost piecewise-stationary CMDP with at least three actions and matched detector alarms of expected delay D_d such that*

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{val}}(T)] \ge \tfrac{1}{2} R_{\max} RD_d. \tag{3}$$

Thus $c_d = 1/2$: the gap is an immediate per-step reward loss during the delay windows. Separately, even when all regime boundaries are revealed to the learner, a stationary-within-regime generic convex-surrogate instance with full sample-loss feedback satisfies

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{ssl}}(T)] \ge c_s DG\sqrt{(R+1)T}. \tag{4}$$

Here $c_s = (1 - \sqrt{2}/4)/(8\sqrt{2}) > 0$. The second display is a generic convex-surrogate lower bound only: it is not an NLL, value, delay, false-alarm, or violation lower bound.

Corollary 6 (Detector delay and false alarms do not change the surrogate exponent)**.** *Suppose* $R \geq 1$,

$$D_d \leq \kappa\sqrt{T/R}, \qquad P_{\mathrm{FA}} \leq \kappa' R/T. \tag{5}$$

Then Theorem 2 gives $\mathbb{E}[\mathrm{Reg}_{\mathrm{ssl}}(T)] = O_{\kappa,\kappa'}(DG\sqrt{RT})$*. In the realizable case of Corollary 4,* $\mathbb{E}[\mathrm{Reg}_{\mathrm{val}}(T)] = O_{\kappa,\kappa'}(V_{\mathrm{span}}\sqrt{RT} + C_V\sqrt{DG}\cdot R^{1/4}T^{3/4})$*. This is a rate-absorption statement, not a comparison with a detector-free algorithm.*

Remark 7 (Exploitability-calibrated prediction)**.** *The theorem does not say whether a detected regime contains an advantage large enough to estimate. For the empirical tests we measure*

$$\rho = \frac{\Delta_{\mathrm{adv}}^2}{\Delta_{\mathrm{adv}}^2 + \sigma_{\mathrm{est}}^2} \in [0, 1), \tag{6}$$

where Δ_{adv} *is the value available to a perfectly timed adapter and* σ_{est} *is its trailing-window noise. We calibrate* ρ_0 *on the controlled sweep, then freeze the following prediction before the real-data runs:*

$$\rho \geq \rho_0 \quad \textit{and} \quad D_d \leq \kappa\sqrt{\rho\, T/R}. \tag{7}$$

This is a signal-to-noise model layered on the scale in Corollary 6, not a consequence of Theorem 2. Its status is empirical: the registered studies can falsify it.

Remark 8 (Registered diffuse-regime prediction)**.** *As* $\Delta_{\mathrm{adv}}/\sigma_{\mathrm{est}} \to 0$*,* $\rho \to 0$ *and the calibrated region in (7) empties for every realisable* $D_d \geq 1$*. We therefore predict ties in diffuse regimes, while allowing occasional sample wins. This is not an impossibility theorem.*

Remark 9 (Frequency dependence)**.** *If advantage accumulates linearly over* τ *steps but noise grows as* $\sqrt{\tau}$*, then* $\rho^{(\tau)}$ *rises with the sampling horizon. A longer window can therefore improve exploitability while worsening delay. The registered weekly and monthly tests probe that trade-off.*

Three testable predictions. The empirical program tests three frozen predictions. (T1) Equal update counts isolate the firing time: detector timing should beat fixed and random schedules, and an oracle change time should recover wins statistical detection misses. (T2) Strict gains occupy an interior delay band that widens with ρ; very early and very late triggers lose it. (T3) Panels below ρ_0, or with slow detection, should tie, while concentrated regimes with fast detection can win.

5 Proof sketches

Theorem 2. The pathwise projected-MD potential bound contributes the sum of the square roots of the realized restart-segment lengths. Delete the disjoint window from each switch to its first matched alarm; their total length is at most $\sum_i d_i$, and every remaining alarm is false. Predictable restarts, conditional subgradient unbiasedness, Cauchy–Schwarz, and $\mathbb{E}[F] \leq TP_{\mathrm{FA}}$ give the first term in (1); delayed switches contribute at most $DG\sum_i \mathbb{E}[d_i] \leq DGRD_d$. Corollary 4 applies Pinsker, the explicit value-sufficiency premise, global Cauchy–Schwarz over non-delay indices, and Jensen's inequality. **Theorem 5.** The delay construction is separate: with three zero-cost actions, each post-switch reward-optimal action is uniform over the two actions different from the preceding optimum, while rewards and the detector score remain uninformative through a delay of mean D_d; a single thresholded score pulse then reveals that action at the matched alarm, giving loss at least $R_{\max}/2$ per delay step. For the surrogate term, even with revealed boundaries, $K = R + 1$ stationary blocks use independent hidden Rademacher signs and i.i.d. two-point targets. A Bernoulli KL–Pinsker testing bound gives $\Omega(DG\sqrt{KT})$. This is a generic convex-surrogate lower bound, not an NLL, value, delay, false-alarm, or violation claim. The calibrated empirical rule scales the separation region by the recoverable fraction $\rho = \Delta_{\mathrm{adv}}^2/(\Delta_{\mathrm{adv}}^2 + \sigma_{\mathrm{est}}^2)$ (a signal-to-noise shrinkage). Proofs of the formal bounds, the value-gap lemma, and the hard-instance construction are in the supplement.

6 Experiments

The three research questions below test predictions T1–T3 in order. Two lanes stay labelled throughout: the stylised constrained engine (controlled sweeps) and the full trained stack (frozen PPO base, LoRA adapter, CPD gate). Metrics map onto the CMDP as follows: in financial tasks reward is return (dynamic regret against the per-regime oracle, or Sharpe) and the constraint is the drawdown

floor, whose depth is the violation side; provisioning tasks use negative matching cost as reward and a stale-capacity budget whose per-step breaches we count.

6.1 RQ1: DOES TIMING CARRY THE VALUE?

The equal-event-count ablation isolates timing on the controlled piecewise-stationary construction (synthetic tape, not market data): detector-coupled adaptation attains Sharpe 6.524 / MDD 0.08347, against fixed-cadence (4.278 / 0.1886) and random-trigger (3.802 / 0.1947) controls. All arms update equally often, so the gap identifies CPD-informed timing rather than budget (full table: supplement).

A pre-registered ablation on the trained stack (`tc-gap2-realpolicy-ablation`) makes the same point: a gate-sensitivity sweep over 3 settings recovers no strict win, yet an oracle change-point time does (1 win on window `real6_fold3_2018`). The bottleneck is the firing time: the oracle recovers a win that the statistical gate misses.

In a single-state hidden-switch construction, a faithful detector-free OMDPD implementation (Zhu et al., 2025) updates every step (600 updates versus 37.17 for the coupled arm) and loses in both regimes; both gap CIs exclude zero (supplement). This is an empirical detector-free baseline, not a branch of Theorem 2.

Historical linear-PMD diagnostic. A post-hoc E1 check replaces AdamW by a convex linear adapter and restart-aware projected MD (30 seeds; supplement). It uses a running observed gradient maximum and an unconstrained hindsight fit, so it tests the restart mechanism rather than instantiating every premise of Theorem 2. Coupled resets beat equal-cadence resets in the concentrated cell (value gap 1.3, CI $[0.5, 1.9]$) and tie in diffuse; no-restart is mildly better on the alternating instance, and doubling the radius closes the timing gap.

6.2 RQ2: WHERE DOES COUPLING HELP?

Measured phase transition (trained stack). A controlled sweep of the trained stack (12 ρ values, 30 seeds per arm) tests Remark 7: the coupled minus agnostic gap is flat below ρ_0 and rises above it (Spearman 0.8462; top-versus-bottom tercile CI $[64.15, 90.23]$), detected timing carrying $2.429\times$ the slope of the any-adaptation component (supplementary §10.1).

The boundary in two dimensions. The empirical rule uses both axes, so two registered experiments sweep them jointly (`tc-phase2d-stylised`, `tc-phase2d-fullstack`): the frozen 12-level ρ grid $\times$ 8 delays, coupled versus matched-budget agnostic, 60 and 30 fresh seeds per cell, on the stylised engine and the trained stack. Every registered full-stack prediction passes (Figure 1): in the band (delays 4–8) the regret gap rises with ρ (Spearman 0.8881– 0.993 per delay; registered threshold 0.7), the seed-paired delay-erosion test excludes zero (delay 4 minus 12, CI $[202.9, 214.7]$), and the violation counter never favours the agnostic arm in-band because its gap is zero. The stylised sibling shows a positive violation gap in every $\rho \geq 0.45$ band cell, and both failure modes as measured harm, growing with ρ (-4349 regret, -2049 violations at $\rho = 0.99$, delay 1; trained stack: Spearman -0.4895 at delay 1, -0.951 at delay 12). Two grid-wide stylised predictions failed in the window-straddle zone; the full-stack study scoped its claims in-band (verdicts ship with the code; stylised map: supplementary §11).

6.3 RQ3: DOES THE BOUNDARY EXPLAIN REAL TASKS?

With the mechanism isolated and the region mapped, the question is whether real tasks sort by their coordinates. We read each locked, placebo-controlled study by where it sits in the (ρ, D_d) plane, not by the order it was run; Table 2 is the ledger.

Registered nulls at daily frequency. `tc-gap2-realwindows` (design, κ, and constants frozen before the run) selects 5 historical break windows by an objective price-only trigger and tests a strict coupled-over-agnostic win by block bootstrap, with 2 placebo windows as a falsification gate. The outcome is a clean null: 0 strict wins, 0 placebo violations, invariant to κ across both registered sensitivity bands. Post-hoc diagnostics show that correctly timed adaptation would reach 0.8–0.84 power on these windows, but the per-regime winner signal-to-noise (median 0.2912)

Table 2: The pre-specified battery by position in the (ρ, D_d) plane. Conditions: exploitability ($\rho \geq \rho_0$) and timely detection; individual registration status is reported in the supplement.

Study (freq.)	$\rho \geq \rho_0$	delay	Pred.	Observed
Broad ETF (styl.+trained)	no	–	tie	null; placebos hold
Concentrated episodes	no	–	tie	tie (2 episodes)
Weekly / monthly	partly	no	tie	tie
FOMC event-anchored	yes	yes	win	miss: in-region, CI $\ni 0$
Fed/ECB/BoE pooled	yes	yes	win	miss: in-region, CI $\ni 0$
Provisioning, real	yes	yes	win	regret win; more violations

sits below the synthetic exploitable value 0.875 in all 5 windows. The regimes are too diffuse (diagnostics: supplement). `tc-gap2-realpolicy` repeats the design with the full trained stack on the inherited windows with fold-matched checkpoints (two named uncovered windows skipped), leaving 3 evaluable. Again null, stable across five seeds: 0 strict wins, mean gap -0.0001273 within bootstrap noise (band [-1.27e-4, 9.6e-5] across seeds 0–4), 0 placebo violations: the diffuse-regime null the stylised engine predicts, already pinned on timing by RQ1's oracle test.

Concentration threshold: a predictive, pre-specified test. We calibrated the diffuse-regime prediction on a controlled synthetic sweep and froze it before the real-data runs (Remark 7). The strict-win crossover in `tc-gap2-rho0-calibration` sets $\rho_0 = 0.1607$. The shrinkage $\rho = \Delta^2_{\text{adv}}/(\Delta^2_{\text{adv}} + \sigma^2_{\text{est}})$ penalizes regimes whose exploitable advantage is small relative to estimation noise, so a positive threshold marks where detected adaptation can no longer pay; it is calibrated, not derived. Three locked studies then test the rule. *Powered confirmation* (`tc-gap2-rw-powered`; the registered `delay_budget=0` sat in the no-gain zone, so we reran at the principled `delay_budget=8`, recorded as an amendment): on the inherited windows, 0 strict wins in the locked span and the refined rule matches all 5 real windows plus placebos (refined accuracy 1). *Concentrated real regimes* (`tc-gap2-concentrated`; two a-priori named sharp-divergence episodes, oil-collapse and covid-crash): even these have $\rho < \rho_0$ at daily frequency, so the refined rule predicts tie and is correct in both (accuracy 1), with 0 strict wins. Recomputing every refined prediction under the calibrated ρ_0 gives 15/16 = 0.9375 accuracy. Most matches are predicted ties on daily panels, and the battery's accuracy equals its tie-only baseline (class breakdown: supplementary §7.1). The miss is a strict win in the long-span arm (`B_long_span/real1`) despite $\rho < \rho_0$. The controlled synthetic witness sits at $\rho \approx 0.4336$ (above ρ_0) and *does* show the strict gain; every real daily-frequency regime we measured sits below ρ_0 and shows none (scatter: supplementary Fig. 4).

Frequency, asset, and power levers. Because $\rho^{(\tau)}$ grows with the sampling horizon (Remark 9), three registered studies vary horizon and panel: weekly, two of three windows clear ρ_0 but none meets the delay condition; the monthly covid-recovery cell clears ρ_0 and ties on delay; daily single-name and crypto panels stay below ρ_0 with zero strict wins. A multi-signal crisis detector is slower than CUSUM (24 versus 16 bars), and CUSUM is minimax-optimal for delay (Lorden, 1971; Moustakides, 1986): statistical detection leaves the strict-win region empty. Three pooled studies press the power lever: a sector-diverse single-name panel returns a null (1/6 cells in scope); pooling every Fed cycle reaches the region on 3 real cells yet brackets zero; adding ECB/BoE calendars yields 0 international in-scope cells, the same pooled CI $[-0.0007, 0.0079]$, and zero diffuse-placebo wins. Side information reaches the region; power does not yet clear zero.

Event-anchored timing. A frozen FOMC calendar collapses median delay $126 \to 1$ bars, placing the cuts-2019 cycle inside the region with gap 0.481; its CI $[-0.061, 1.04]$ still includes zero, so we report it as underpowered rather than as a win. The frozen all-cell accuracy criterion also failed (0.9; H2 = false), so this was not a confirmatory pass (mechanism: supplement).

Protocol-specified scheduled-timing gain. The financial panels never met both conditions at once. Three public non-financial operational datasets do: highway traffic, bike-sharing demand, and household electricity load, cast as a provisioning CMDP (`tc-nonfinance-coupling`). The local protocol designated electricity as an internal holdout, but was externally archived only after

Table 3: Protocol-specified exogenous-schedule provisioning test: schedule/CUSUM (s/C) delay, quadratic regret gap (95% CI), and post-hoc violations at matched event budgets. †: violation-difference CI includes zero.

Domain	D_d s/C	Regret gap [CI]	Viol. s/C
Traffic (hourly)	$1.968/7.923$	$+74.8\ [66.01, 83.82]$	$26/19.7$
Bike (daily)	$1.254/6.648$	$+1.236\ [0.4312, 1.912]$	$1.9/1.7$†
Electricity (designated)	$1.915/15.63$	$+52.72\ [37.67, 68.45]$	$21.7/16.6$

the run; this is protocol-specified, not pre-registered evidence. The coupled arm adapts on an exogenous calendar of clock, season, and holiday boundaries; the comparator is CUSUM on the same series at equal event budget. The schedule is exogenous side information, not an oracle or learned detector, so this is a timing-mechanism test. Table 3 shows that it cuts detection delay four- to eightfold, and the regret win is strict in every domain, including the designated power series, for all three protocol-specified cost forms. A post-hoc violation count shows the early schedule arm pays more stale-capacity breaches on traffic and electricity (the bike difference interval includes zero). Figure 1 shows the same early-trigger cost in-window; holding the trigger to the band's interior is future work.

Retrospective detector-score ranking. The ridge and BOCPD extensions rank each completed episode's events in hindsight; they are diagnostics, not online detector policies (supplement). A calibration-only causal replay over a 240-cell PMD grid on five fixed UCI windows is mixed: ridge thresholding does not robustly beat CUSUM, and the fixed-optimizer comparison favours PMD over AdamW/LoRA. A pre-specified archived VM-demand development gate also failed, leaving its sealed evaluation index unopened; financial drawdown counts remain sparse and do not favour coupling (ledgers in supplement). No real-data dynamic-surrogate-regret claim is made.

7 Discussion

The theorem is the paper's load-bearing result: detector delay and false restarts enter the bound directly, and the radius controls excess expected cost. Empirically, the frozen rule is calibrated in both directions: it abstains on the daily financial panels, where 15 of 16 predictions hold, and fires on every provisioning domain. Equal-event ablations and the oracle change-time test locate the remaining gain in firing time. The trained-stack map places positive regret gaps at interior delays, while daily financial panels stay below the frozen exploitability threshold. The calendar-anchored cells remain underpowered. Whole-episode rankings identify useful events in hindsight, but the causal replay is mixed. Thus the exogenous-schedule result isolates timing value; it does not establish broad learned-detector superiority.

Retrospective operational hypothesis. Estimate trailing-window ρ; stay frozen below ρ_0, and couple above it only when the delay condition holds. This rule remains empirical; no causal online ρ estimator is evaluated. Formally, expected-cost Lipschitzness and base feasibility make radius an exposure dial, while the regret bound prices delay and false restarts for any detector satisfying its assumptions.

Limitations. The theorem covers a convex last-layer predictor with restart-aware projected MD. The bilinear LoRA/AdamW runs are outside that result. The value transfer retains predictive and policy-comparator approximation terms and requires explicit value sufficiency; the $T^{3/4}$ term is its realizable special case. The surrogate theorem also requires adapter-independent target dynamics. The radius condition controls excess expected per-step cost, not cumulative violation, and not path-level quantities such as the drawdown of the Introduction. The delay lower bound uses the reward-free observation model. The frozen ρ threshold is empirical, and trained-stack in-band violation gaps are zero. The frozen base is assumed trained feasible by a standard constrained-RL pipeline (Achiam et al., 2017; Ray et al., 2019); obtaining it is out of scope. Evaluation covers financial panels and provisioning domains rather than continuous-control safety benchmarks, with CUSUM and OMDPD rather than PPO-Lagrangian-class baselines; generalization there is untested. Per-study detail is in the supplement ("Consolidated limitations").

REPRODUCIBILITY STATEMENT

Registration status is stated per study; the provisioning schedule protocol was externally archived after its run and is labelled protocol-specified. Frozen designs, seeds, audit tables, and formal proofs ship with the anonymized code and supplement. Figure 1 and the trained-stack map use `tc-phase-transition` at `seed_base=8000` (ADR 0154 Path Z); earlier archive directories are exploratory.

ETHICS STATEMENT

This work studies constrained reinforcement-learning policies in simulated and public benchmark environments. No human subjects, personal data, or real-world deployment are involved. The motivating domains (trading, provisioning, traffic) are operational settings where constraint violations carry real costs; any deployment of timing-gated adaptation in such settings should treat the measured violation trade-offs reported here as requiring domain-specific validation first. The authors declare no conflicts of interest.

AI USE STATEMENT

AI tools assisted literature search, prose editing, code/build tooling, and critical review and revision of mathematical statements and proofs. They also suggested changes to the presentation of the experiments; registered designs and outcomes were not altered. The authors checked the sources, derivations, code, and reported results and take responsibility for the paper. No AI system is an author or a citable source.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Réda Alami, Mohammed El Mahfoud, and Éric Moulines. Restarted Bayesian online change-point detection for non-stationary Markov decision processes. In *Conference on Lifelong Learning Agents (CoLLAs)*, volume 232 of *Proceedings of Machine Learning Research*, pp. 715–744, 2023. URL https://proceedings.mlr.press/v232/alami23a.html.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pp. 138–158, 2019. URL http://proceedings.mlr.press/v99/auer19a.html.

Marco Bagatella, Mert Albaba, Jonas Hübotter, Georg Martius, and Andreas Krause. Test-time offline reinforcement learning on goal-related experience. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2507.18809.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015. URL https://doi.org/10.1287/opre.2015.1408.

Annie S. Chen, Govind Chada, Laura Smith, Archit Sharma, Zipeng Fu, Sergey Levine, and Chelsea Finn. Adapt on-the-go: Behavior modulation for single-life robot deployment. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2025. URL https://arxiv.org/abs/2311.01059.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning (ICML)*, 2020.

Yuhao Ding and Javad Lavaei. PROPD-PPO: Provably efficient primal-dual reinforcement learning for CMDPs with non-stationary objectives and constraints. In *AAAI Conference on Artificial Intelligence*, 2023. URL https://arxiv.org/abs/2201.11965.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020. URL https://arxiv.org/abs/2003.02189.

Argyrios Gerogiannis, Yu-Han Huang, and Venugopal V. Veeravalli. Is prior-free black-box non-stationary reinforcement learning feasible? In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2692–2700, 2025. URL https://proceedings.mlr.press/v258/gerogiannis25a.html.

Argyrios Gerogiannis, Yu-Han Huang, and Venugopal V. Veeravalli. DARLING: Detection augmented reinforcement learning with non-stationary guarantees. *arXiv preprint arXiv:2604.16684*, 2026.

Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2007.04309.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4): 157–325, 2016. doi: 10.1561/2400000013.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6): 1897–1908, 1971.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Başar. Near-optimal model-free reinforcement learning in non-stationary episodic MDPs. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, 2021. URL https://proceedings.mlr.press/v139/mao21b.html.

George V. Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387, 1986.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained MDPs. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.15776.

Tingting Ni and Maryam Kamgarpour. Constrained meta reinforcement learning with provable test-time safety. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2601.21845.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. In *Preprint, arXiv:1910.01708*, 2019.

Francesco Emanuele Stradi, Jacopo Germano, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online learning in CMDPs: Handling stochastic and adversarial constraints. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46692–46721, 2024a. URL https://proceedings.mlr.press/v235/stradi24a.html.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Learning constrained Markov decision processes with non-stationary rewards and constraints, 2024b. URL https://arxiv.org/abs/2405.14372.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

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

Min Wang, Xin Li, Ye He, Yao-Hui Li, Hasnaa Bennis, Riashat Islam, and Mingzhong Wang. Wavelet predictive representations for non-stationary reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.04507.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of the Thirty Fourth Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4300–4354, 2021. URL https://proceedings.mlr.press/v134/wei21b.html.

Honghao Wei, Arnob Ghosh, Ness Shroff, Lei Ying, and Xingyu Zhou. Provably efficient model-free algorithms for non-stationary CMDPs. In *Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, 2023. URL https://arxiv.org/abs/2303.05733.

Shoukai Xu, Zihao Lian, Mingkui Tan, Liu Liu, Zhong Zhang, and Peilin Zhao. Test-time adapted reinforcement learning with action entropy regularization. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70020–70031, 2025. URL https://proceedings.mlr.press/v267/xu25ar.html.

Kihyun Yu, Seoungbin Bae, and Dabeen Lee. Primal-dual policy optimization for linear CMDPs with adversarial losses. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=baP3Dw7bWO.

Suyeol Yun. Pretrained LLM adapted with LoRA as a decision transformer for offline RL in quantitative trading, 2024. URL https://arxiv.org/abs/2411.17900.

Yifan Zhang and Liang Zheng. BAPR: Bayesian amnesic piecewise-robust reinforcement learning for non-stationary continuous control. *arXiv preprint arXiv:2605.16170*, 2026.

Chujin Zhou, Yuling Huang, Yuting Kong, and Xiaoping Lu. Enhancing trading strategies by combining incremental reinforcement learning and self-supervised prediction. *Expert Systems with Applications*, 289, 2025. doi: 10.1016/j.eswa.2025.128297.

Jiahui Zhu, Kihyun Yu, Dabeen Lee, Xin Liu, and Honghao Wei. An optimistic algorithm for online CMDPs with anytime adversarial constraints. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80347–80372, 2025. URL https://openreview.net/forum?id=fFgiXamW8E.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning (ICML)*, pp. 928–936, 2003.

Detector-Coupled Surrogate Regret for Test-Time Adaptation in Piecewise-Stationary CMDPs: Appendix (Supplementary Material)

Anonymous Author(s)

1 Positioning against detection-augmented RL

The closest prior line puts a change detector inside a non-stationary RL algorithm (Gerogiannis et al., 2026, 2025; Zhang and Zheng, 2026). Table 1 places AdaptiveCMDP against this line and the test-time-adaptation and portfolio-RL methods along five axes. The formal difference is a restart-aware surrogate-regret bound whose segment count includes false firings. The closest detection-augmented methods are unconstrained and use detection during learning or control to restart an algorithm or retune conservatism.

Table 1: Comparison with the closest published work. "Emp." marks an implementation result outside the paper's formal guarantee.

Method	Reward-free TTA	CPD-gated rate	Deploy adapter	Constraint result	Live deploy
Hansen et al. (2021)	yes	no	no	no	no
Chen et al. (2025)	yes	no	no	no	no
Xu et al. (2025)	yes	no	partial (LayerNorm)	no	no
Bagatella et al. (2026)	yes	no	no	no	no
Yun (2024)	no (offline)	no	yes	no	no
Zhou et al. (2025)	yes	no	no	no	no
Wang et al. (2026)	yes	no	no	no	no
Li et al. (2025)	no (offline imitation $\rightarrow$ RL)	no	no	no	no
Raj (2025)	no	no	no	no	no
Shen et al. (2025)	no (meta)	no	no	no	no
Zhang (2025)	no	no	no	yes (CBF)	no
Gerogiannis et al. (2026)	no	restart	no	no (unconstrained PS-RL)	no
Gerogiannis et al. (2025)	no	restart	no	no	no
Zhang and Zheng (2026)	no	conservatism gate	no	no	no
Zhu et al. (2025)	no	no	no	yes	no
Ours (AdaptiveCMDP)	**yes**	**yes (PMD)**	**LoRA/AdamW (emp.)**	**conditional value; exp. cost**	**no**

2 Deep-adapter surrogate–value diagnostic

Theorem 2 is exact for a convex last-layer predictor. The rank-4 bilinear LoRA adapter and AdamW optimizer used in the trained-stack experiments do not satisfy that premise. The registered `tc-surrogate-gap` run records how the empirical loss and value quantities move on the same frozen E1 trajectories; it is a diagnostic rather than an estimate of a theorem remainder.

From one shared trained base per seed, we run the online schedule with a linear last-layer correction and with the rank-4 bilinear adapter (30 seeds per variant, `seed_base`=9300). The absolute difference between the reported surrogate- and value-regret summaries is larger for the bilinear adapter in both cells: concentrated 4.66 versus 2.29, and diffuse 834.19 versus 604.78 (Figure 1). These differences mix quantities with task-specific scaling and therefore do not establish a surrogate-to-value calibration. They show that the deep trajectory departs materially from the convex reference in the diffuse cell.

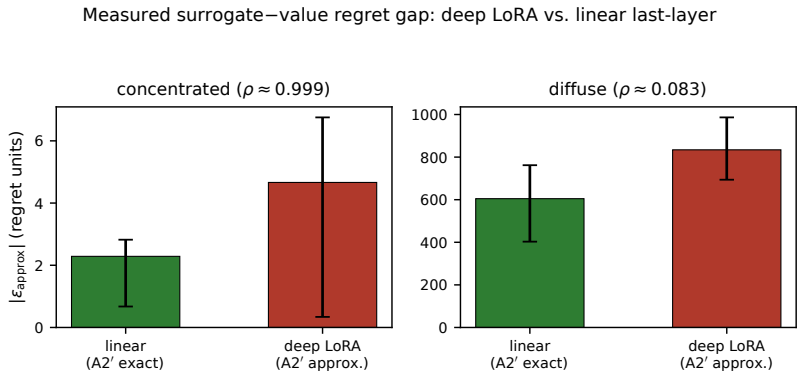

Figure 1: Descriptive absolute difference between the reported surrogate- and value-regret summaries (95% block-bootstrap CI) for linear and bilinear adapters on the E1 concentrated and diffuse variants. The quantities have task-specific scaling and should be read as a diagnostic.

3 Standing assumptions and constants

The proofs below are self-contained: every constant and assumption the main paper refers to in passing is restated here. Throughout, T is the horizon, R the number of regime switches, $\mathcal{M}^{(i)}$ the i-th stationary regime of duration

T_i, $\Delta\theta$ the LoRA adapter parameters, and D_d, P_{FA} the detector's expected detection delay and per-step false-alarm probability.

Standing assumptions. We assume: (A1) *Bounded rewards and costs*: per-step reward and per-coordinate cost are bounded by $R_{\max}$ and $C_{\max}$ respectively; (A2) *Expected-cost Lipschitzness*: uniformly over times and regimes, the policy-induced expected per-step cost satisfies $|\mathbb{E}[c_t(\Delta\theta)] - \mathbb{E}[c_t(\Delta\theta')]| \le L\|\Delta\theta - \Delta\theta'\|_F$; (A2′) *Surrogate convexity*: in regime i, the expected next-step negative log likelihood ℓ_i is convex on the LoRA ball $\mathcal{B}$, and each sample supplies a subgradient estimate g_t with $\|g_t\|_2 \le G$. Let $\mathcal{F}_{t-1}$ contain the history and detector decisions available before the step-t target is observed. The iterate and step-t restart decision are $\mathcal{F}_{t-1}$-measurable, and $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] \in \partial\ell_i(\Delta\theta_t)$. The next-step target law is adapter-independent: the observation stream is exogenous to the adapted policy, so ℓ_i does not move along the trajectory. Every evaluated domain is a forecast-style panel with this property; a domain whose observation stream responds to the adapter lies outside the theorem. This is exact for a linear last-layer adapter. The bilinear LoRA/AdamW trajectory used in the experiments is outside the theorem; (A3) *LoRA budget*: updates stay in the Frobenius ball $\mathcal{B} = \{\Delta\theta : \|\Delta\theta\|_F \le B\}$ of diameter $D = 2B$, with adapter rank r; (A4) *Slater feasibility*: each regime admits a strictly feasible policy with margin $\zeta > 0$ (Achiam et al., 2017; Müller et al., 2024); (A5) *Base-policy near-optimality*: the frozen base policy has per-regime optimality gap at most ϵ_0; (A6) *Detector regularity*: write the switch times as $\tau_1 < \cdots < \tau_R$ and set $\tau_{R+1} = T + 1$. For switch i, its matched firing a_i is the first detector firing in $[\tau_i, \tau_{i+1})$, its delay is $d_i = a_i - \tau_i$, and its delay window is $W_i = [\tau_i, a_i) \cap \mathbb{Z}$. Segment separation guarantees $\tau_i + d_i < \tau_{i+1}$, so every matched firing occurs before the next switch. All firings other than the a_i are false; writing their number as F, we assume $\mathbb{E}[d_i] \le D_d$ for every i and $\mathbb{E}[F] \le TP_{\mathrm{FA}}$; (A7) *Minimum segment length*: consecutive switches are separated by at least $d_{\min}$ steps, i.e. $R \le T/d_{\min}$; (A8) *Base feasibility*: in every regime the frozen base policy is feasible with the margin of (A4): its per-step expected constraint cost satisfies $\mathbb{E}[c_t(\mathbf{0})] \le \text{budget} - \zeta$. This is a statement about per-step expectation, not about realized paths: individual trajectories can still breach (as the Introduction motivates), and it is this margin that the adapter radius is allowed to spend in Proposition 2. Of these, (A1), (A2′), (A3), and (A6) enter the displayed bounds; (A2) with (A8) enters Proposition 2; value transfer additionally requires the explicit value-sufficiency condition in Lemma 4; (A4), (A5), and (A7) are CMDP

context.

Loss regularity. Assumption (A2′) bounds $\|\nabla \ell_i(\Delta\theta)\|_2$ by G on $\mathcal{B}$. Consequently the loss difference between any two adapters in the ball is at most DG. For a vectorized last-layer adapter the Frobenius and Euclidean norms coincide, so the ball of (A3) and the subgradient bound of (A2′) share one geometry.

Mirror-descent schedule. Let $n_t = \max\{1, t - \tau_{\text{last}} + 1\}$ count steps since the most recent detector firing τ_{last}. We use the restart-aware schedule

$$\eta_t = \frac{D}{G\sqrt{n_t}}, \qquad \Delta\theta_{t+1} = \Pi_{\mathcal{B}}\big(\Delta\theta_t - \eta_t g_t\big), \tag{1}$$

with $\Pi_{\mathcal{B}}$ the Frobenius-ball projection; this is the schedule of Hazan (2016, Theorem 3.1), whose $\frac{3}{2}DG\sqrt{T}$ bound we relax to $2DG\sqrt{T}$ on a single regime of length T.

Theorem 1 (Detector-coupled surrogate regret)**.** *For regime i, let $\ell_i(\Delta\theta) = \mathbb{E}[-\log p_{g_\phi}(Y; \Delta\theta)]$, where the next-step target Y is treated as a supervised target whose law does not depend on $\Delta\theta$, and choose $\Delta\theta_i^\star \in \arg\min_{\Delta\theta\in\mathcal{B}} \ell_i(\Delta\theta)$. Let $\mathcal{I}_i$ be its time indices and define*

$$\mathrm{Reg}_{\mathrm{ssl}}(T) = \sum_{i=0}^{R} \sum_{t\in\mathcal{I}_i} \big(\ell_i(\Delta\theta_t) - \ell_i(\Delta\theta_i^\star)\big).$$

Suppose ℓ_i is convex on a set $\mathcal{B}$ of diameter D (for the Frobenius ball of (A3), $D = 2B$), and each sample supplies a subgradient estimate g_t with $\|g_t\|_2 \le G$. Let $\mathcal{F}_{t-1}$ contain the history and detector decisions available before the step-t target is observed. Suppose $\Delta\theta_t$ and the step-t restart decision are $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] \in \partial\ell_i(\Delta\theta_t)$. If the projected-MD clock restarts at every detector firing, use the following matching convention. Let the switches occur at $\tau_1 < \cdots < \tau_R$ and set $\tau_{R+1} = T + 1$. For each switch i, the first firing $a_i \in [\tau_i, \tau_{i+1})$ is its unique matched firing, its delay is $d_i = a_i - \tau_i$, and the delay window is $W_i = [\tau_i, a_i) \cap \mathbb{Z}$. Assume segment separation is sufficient that these matched firings exist before the next switch (equivalently, $\tau_i + d_i < \tau_{i+1}$); since each d_i is random, the bound below is stated conditional on this event (equivalently, it holds with the probability that the separation condition does). Every firing other than the a_i is false; if F is their number,

assume $\mathbb{E}[d_i] \le D_d$ *for every* i *and* $\mathbb{E}[F] \le TP_{\mathrm{FA}}$*. The disjoint windows* W_i *may be deleted in the analysis and have total length* $\sum_i |W_i| = \sum_i d_i$*. Then*

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{ssl}}(T)] \le DG\left[2\sqrt{T\big(R+1+TP_{\mathrm{FA}}\big)} + RD_d\right] =: B_{\mathrm{ssl}}(T). \tag{2}$$

The result is exact for a convex last-layer adapter. It is not a guarantee for the bilinear LoRA/AdamW trajectory used in the experiments.

Proposition 2 (Adapter radius as an exposure dial)**.** *Under (A1), (A2), (A3) and (A8), the positive excess of expected per-step cost is bounded linearly in the adapter radius* B *of (A3):*

$$\sup_t \big[\mathbb{E}\, c_t(\Delta\theta_t) - \mathrm{budget}\big]_+ \;\le\; [L\,B - \zeta]_+ \;\le\; L\,B.$$

Hence $B \le \zeta/L$ *keeps expected per-step cost inside the budget in every regime: the ball radius is an expected-cost exposure dial whose exchange rate is linear. No sublinear cumulative-violation bound follows; cumulative excess is reported as a measured quantity throughout.*

Corollary 3 (Conditional surrogate-to-value transfer)**.** *Write* $p_Y(\cdot\,;\Delta\theta) = p_{g_\phi}(\cdot\,;\Delta\theta)$ *for the predicted next-step target law. Assume it is value-sufficient: for some* $C_V < \infty$ *and every* $\Delta\theta, \Delta\theta' \in \mathcal{B}$*,*

$$|V_i(\Delta\theta) - V_i(\Delta\theta')| \le C_V\, \mathrm{TV}\big(p_Y(\cdot\,;\Delta\theta), p_Y(\cdot\,;\Delta\theta')\big)\,.$$

Let $\varepsilon_i^{\mathrm{pred}} = \ell_i(\Delta\theta_i^\star) - H(p_{\mathrm{true}}^{(i)})$*, let* $\pi_i^\star$ *be the best feasible in-ball policy, and define* $A_T = \sum_i T_i[V_i(\pi_i^\star) - V_i(\Delta\theta_i^\star)]_+$ *and* $\mathrm{Reg}_{\mathrm{val}}(T) = \sum_i \sum_{t\in\mathcal{I}_i}(V_i(\pi_i^\star) - V_i(\Delta\theta_t))$*. Let* V_{span} *bound the difference of any two policies' values (*$V_{\mathrm{span}} \le 2R_{\max}/(1-\gamma)$ *under (A1)). Then*

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_{\mathrm{val}}(T)] \le{}& A_T + V_{\mathrm{span}} RD_d \\ &+ 2C_V\sqrt{DG}\cdot T^{3/4}\cdot(R+1+TP_{\mathrm{FA}})^{1/4} + 2C_V\sum_i T_i\sqrt{\varepsilon_i^{\mathrm{pred}}}.\end{aligned} \tag{3}$$

Proof of Theorem 1. Let F be the number of false detector firings. After deleting the disjoint windows $W_i = [\tau_i, a_i) \cap \mathbb{Z}$, whose total length is $\sum_i |W_i| = \sum_i d_i$, the remaining horizon is partitioned into at most $M = R + 1 + F$ stationary projected-MD runs: each matched firing separates the pre-switch part from the post-delay part, and each remaining firing is false. On each realized run, the standard projected-MD potential inequality

underlying Theorem 3.1 of Hazan (2016) applies pathwise to the bounded vectors g_t. Because the iterate and restart decision are predictable and $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] \in \partial \ell_i(\Delta\theta_t)$, convexity converts the expected linearized bound into expected loss regret. For the realized run lengths $L_1, \ldots, L_M$, Cauchy–Schwarz gives

$$\sum_{j=1}^{M} 2DG\sqrt{L_j} \le 2DG\sqrt{MT}.$$

The loss difference between any two points in $\mathcal{B}$ is at most DG. The expected total length of the deleted delay windows is $\sum_i \mathbb{E}[d_i] \le RD_d$, so their contribution is at most $DGRD_d$. By (A6), $\mathbb{E}[F] \le TP_{\mathrm{FA}}$. Jensen's inequality therefore yields

$$\mathbb{E}[\sqrt{M}] \le \sqrt{\mathbb{E}[M]} \le \sqrt{R + 1 + TP_{\mathrm{FA}}}.$$

Adding the two contributions gives (2). The proof concerns the convex next-step loss and does not imply a constraint-violation bound. □

Proof of Proposition 2. By (A2), policy-induced expected cost is L-Lipschitz in the adapter, so for any $\Delta\theta \in \mathcal{B}$, $\mathbb{E}[c_t(\Delta\theta)] \le \mathbb{E}[c_t(\mathbf{0})] + L\,\|\Delta\theta\| \le \mathbb{E}[c_t(\mathbf{0})] + LB$, since (A3) bounds $\|\Delta\theta\|_F \le B$. By (A8), $\mathbb{E}[c_t(\mathbf{0})] \le \text{budget} - \zeta$. Hence $\mathbb{E}[c_t(\Delta\theta)] - \text{budget} \le LB - \zeta$, which is non-positive when $B \le \zeta/L$. Therefore

$$\big[\mathbb{E}\, c_t(\Delta\theta) - \text{budget}\big]_+ \le [LB - \zeta]_+ \le LB.$$

Taking the supremum over t gives the displayed bound, and $B \le \zeta/L$ makes its first right-hand side zero. □

Proof of Corollary 3. Let $\mathcal{N} = \{1, \ldots, T\} \setminus \bigcup_i W_i$ be the non-delay time indices. Decompose the per-step value gap as

$$V_i(\pi_i^\star) - V_i(\Delta\theta_t) \le \big[V_i(\pi_i^\star) - V_i(\Delta\theta_i^\star)\big]_+ + \big|V_i(\Delta\theta_i^\star) - V_i(\Delta\theta_t)\big|,$$

where the first term sums to A_T. For $t \in \mathcal{N}$, write $\delta_t = \ell_i(\Delta\theta_t) - \ell_i(\Delta\theta_i^\star) \ge 0$. Lemma 4 and $\sqrt{\delta_t + \varepsilon_i^{\mathrm{pred}}} \le \sqrt{\delta_t} + \sqrt{\varepsilon_i^{\mathrm{pred}}}$ give

$$\sum_{t \in \mathcal{N}} |V_i(\Delta\theta_t) - V_i(\Delta\theta_i^\star)| \le C_V \sum_{t \in \mathcal{N}} \sqrt{\delta_t} + 2C_V \sum_i T_i \sqrt{\varepsilon_i^{\mathrm{pred}}}.$$

Cauchy–Schwarz over all non-delay indices, followed by Jensen, yields

$$\mathbb{E}\left[\sum_{t \in \mathcal{N}} \sqrt{\delta_t}\right] \le \sqrt{T\,\mathbb{E}\left[\sum_{t \in \mathcal{N}} \delta_t\right]}.$$

The non-delay part of the proof of Theorem 1 gives

$$\mathbb{E}\left[\sum_{t\in\mathcal{N}}\delta_t\right] \le 2DG\sqrt{T(R+1+TP_{\mathrm{FA}})}.$$

Substitution gives the $C_V\sqrt{2DG}\cdot T^{3/4}(R+1+TP_{\mathrm{FA}})^{1/4}$ term. The omitted delay windows contribute at most $V_{\mathrm{span}}RD_d$. Adding the policy-comparator gap A_T and using $\sqrt{2}\le 2$ proves (3). Thus a $\sqrt{T}$ cumulative surrogate term becomes a $T^{3/4}$ cumulative value term without conditioning on detector-dependent stopping lengths. □

Lemma 4 (SSL-to-value gap)**.** *Fix regime i. Let $p^{(i)}_{\mathrm{true}}$ be the true law of the grouped next-step target, let $\ell_i(\Delta\theta)=\mathbb{E}_{Y\sim p^{(i)}_{\mathrm{true}}}[-\log p_{g_\phi}(Y;\Delta\theta)]$, and let $\Delta\theta_i^\star$ minimize ℓ_i on $\mathcal{B}$. Define the predictive approximation error $\varepsilon_i^{\mathrm{pred}} = \ell_i(\Delta\theta_i^\star) - H(p^{(i)}_{\mathrm{true}})$. Write $p_Y(\cdot\,;\Delta\theta) = p_{g_\phi}(\cdot\,;\Delta\theta)$. Assume the predicted target law is value-sufficient: for some $C_V<\infty$ and every $\Delta\theta,\Delta\theta'\in\mathcal{B}$,*

$$|V_i(\Delta\theta)-V_i(\Delta\theta')| \le C_V\,\mathrm{TV}\big(p_Y(\cdot\,;\Delta\theta),p_Y(\cdot\,;\Delta\theta')\big).$$

Then

$$|V_i(\Delta\theta)-V_i(\Delta\theta_i^\star)| \le C_V\left[\sqrt{\ell_i(\Delta\theta)-\ell_i(\Delta\theta_i^\star)+\varepsilon_i^{\mathrm{pred}}}+\sqrt{\varepsilon_i^{\mathrm{pred}}}\right]. \tag{4}$$

Proof. Cross-entropy satisfies

$$\mathrm{KL}(p^{(i)}_{\mathrm{true}}\,\|\,p_{g_\phi}(\cdot;\Delta\theta)) = \ell_i(\Delta\theta)-H(p^{(i)}_{\mathrm{true}}) = \ell_i(\Delta\theta)-\ell_i(\Delta\theta_i^\star)+\varepsilon_i^{\mathrm{pred}}.$$

Pinsker's inequality bounds the current predictor's distance from the true law by the first square-root term and the comparator's distance by $\sqrt{\varepsilon_i^{\mathrm{pred}}}$, up to universal factors. The triangle inequality and the explicit value-sufficiency premise absorb those factors into C_V and give (4). The approximation terms cannot be dropped unless the true predictive law is realizable in $\mathcal{B}$. □

Theorem 5 (Scoped delay and surrogate lower bounds)**.** *Let $K=R+1$, suppose $T\ge K\max\{2,d_{\min}\}$, and set $n=\lfloor T/K\rfloor$ and $q=T-Kn$. In both constructions, take K consecutive regime blocks of lengths $n,\dots,n,n+q$. For a feasible delay satisfying $0\le D_d$ and $\lceil D_d\rceil<d_{\min}$, for any learner that observes (s_t,a_t,c_t,ξ_t) but no rewards or regime labels, there is a bounded, zero-cost piecewise-stationary CMDP with at least three actions and matched detector alarms of expected delay D_d such that*

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{val}}(T)] \ge \tfrac{1}{2}R_{\max}RD_d. \tag{5}$$

Thus $c_d = 1/2$: the gap is an immediate per-step reward loss during the delay windows. Separately, even when all regime boundaries are revealed to the learner, a stationary-within-regime generic convex-surrogate instance with full sample-loss feedback satisfies

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{ssl}}(T)] \geq c_s DG\sqrt{(R+1)T}. \tag{6}$$

Here $c_s = (1-\sqrt{2}/4)/(8\sqrt{2}) > 0$. The second display is a generic convex-surrogate lower bound only: it is not an NLL, value, delay, false-alarm, or violation lower bound.

Proof of Theorem 5. For both constructions, partition the horizon into K consecutive blocks $\mathcal{J}_0, \ldots, \mathcal{J}_{K-1}$ of lengths $n, \ldots, n, n+q$, and let τ_i be the first index of $\mathcal{J}_i$ (so $\tau_K = T+1$). The condition on T gives $n \geq d_{\min}$, hence every block meets the minimum-segment condition.

For (5), use a one-state, one-stage CMDP with $\gamma = 0$, three actions $\mathcal{A} = \{1,2,3\}$, and identically zero costs. Let b_0 be the reward-optimal action in the initial regime. At switch i, draw b_i uniformly from the two actions in $\mathcal{A} \setminus \{b_{i-1}\}$ and set the regime-i reward to

$$r_i(a) = R_{\max}\mathbf{1}\{a = b_i\}.$$

Give the learner the switch times and even the preceding optimal action b_{i-1}; this can only help it. Let $p = D_d - \lfloor D_d \rfloor$ and independently draw

$$\Delta_i = \lfloor D_d \rfloor + \mathrm{Bernoulli}(p), \qquad \mathbb{E}[\Delta_i] = D_d.$$

Set $a_i = \tau_i + \Delta_i$ and define the detector score as one pulse,

$$\xi_t = \begin{cases} b_i/3, & t = a_i, \\ 0, & \text{otherwise}, \end{cases} \qquad \text{with firing threshold } \geq 1/3.$$

Thus the score reveals b_i exactly at the matched alarm and has no other firings; before it, state, action, zero cost, and score are uninformative about b_i, and no rewards are observed. Since $\lceil D_d \rceil < d_{\min} \leq n$, $a_i < \tau_{i+1}$; the windows $W_i = [\tau_i, a_i) \cap \mathbb{Z}$ are therefore disjoint and the matched alarms have the required expected delay.

For every $t \in W_i$, even conditional on the complete prior history and b_{i-1}, the hidden b_i is uniform on two actions. Any learner action is therefore correct with probability at most $1/2$. Because this is a one-stage $\gamma = 0$ CMDP, for every policy π_t,

$$V_i(\pi_i^\star) - V_i(\pi_t) = R_{\max}\big(1 - \pi_t(b_i)\big).$$

Its conditional expected per-step value loss is consequently at least $R_{\max}/2$. Summing only over the disjoint delay windows gives

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{val}}(T)] \geq \frac{R_{\max}}{2} \sum_{i=1}^{R} \mathbb{E}[\Delta_i] = \frac{1}{2} R_{\max} R D_d.$$

This proves the delay claim with $c_d = 1/2$. It uses no reward observations, and neither asserts nor implies a false-alarm or violation lower bound.

For (6), set $\mathcal{B} = [-D/2, D/2]$ and reveal all K regime boundaries to the learner. Draw independent Rademacher signs $\sigma_i \in \{-1, +1\}$, set $\varepsilon = 1/(4\sqrt{n})$, and, in regime i, draw i.i.d. $Y \in \{-1, +1\}$ with

$$\mathbb{P}(Y = +1 \mid \sigma_i) = \frac{1 + \sigma_i \varepsilon}{2}, \qquad f_i(\theta; Y) = \frac{3GD}{4} + GY\theta + \frac{GD}{4}(-1)^i.$$

Since $GY\theta \geq -GD/2$ and the additive constant is at least $GD/2$, these are nonnegative convex sample losses with subgradient norm G. Adjacent blocks have distinct stationary loss environments even when their signs match, because their additive constants alternate; distinct environments need not have distinct minimizers. Their stationary expected loss and optimum are

$$\ell_i(\theta) = \frac{3GD}{4} + \frac{GD}{4}(-1)^i + \sigma_i \varepsilon G\theta, \qquad \theta_i^\star = -\sigma_i \frac{D}{2}.$$

Let P_+ and P_- be the two Bernoulli laws of Y. Since $\varepsilon \leq 1/4$,

$$\mathrm{KL}(P_+ \,\|\, P_-) = \varepsilon \log \frac{1+\varepsilon}{1-\varepsilon} \leq 4\varepsilon^2.$$

Thus, after at most n within-regime samples, Pinsker's inequality gives

$$\mathrm{TV}(P_+^n, P_-^n) \leq \sqrt{2n}\,\varepsilon = \frac{\sqrt{2}}{4}.$$

The signs in preceding regimes are independent, so their observations do not improve this testing problem. For any causal (possibly randomized) prediction $\theta_t \in [-D/2, D/2]$, the range-D total-variation bound gives the per-round Bayes regret

$$\begin{aligned}
\mathbb{E}[\ell_i(\theta_t) - \ell_i(\theta_i^\star)] &= \varepsilon G \left(\frac{D}{2} + \mathbb{E}[\sigma_i \theta_t] \right) \\
&\geq \frac{\varepsilon G D}{2} \left(1 - \mathrm{TV}(P_+^n, P_-^n) \right) \\
&\geq \frac{\varepsilon G D}{2} \left(1 - \frac{\sqrt{2}}{4} \right).
\end{aligned}$$

Consequently each active regime contributes at least

$$\frac{1-\sqrt{2}/4}{8}\,DG\sqrt{n}$$

in Bayes regret on its first n rounds. Summing those first n rounds over the K independent regimes and using $n = \lfloor T/K \rfloor \geq T/(2K)$ yields

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{ssl}}(T)] \geq \frac{1-\sqrt{2}/4}{8}\,DGK\sqrt{n} \geq \frac{1-\sqrt{2}/4}{8\sqrt{2}}\,DG\sqrt{KT}.$$

The final block's remaining q i.i.d. rounds have nonnegative expected regret and were omitted from the lower-bound calculation. Finally, averaging over the sign vector proves the Bayes bound, so at least one fixed vector $(\sigma_1, \ldots, \sigma_K)$ attains it. This is a generic convex-surrogate construction, not an NLL, value, delay, false-alarm, or violation lower bound. □

Proof of Corollary 6. The false-alarm condition gives $R + 1 + TP_{\mathrm{FA}} = O_{\kappa'}(R)$, while the delay condition gives $RD_d \leq \kappa\sqrt{RT}$. Substitution in Theorem 1 proves the surrogate rate. Substitution in the realizable case of Corollary 3 proves the value rate. No detector-free comparator is used. □

Status of the exploitability rule. Remark 7 is a calibrated signal-to-noise model, not a corollary requiring proof. The controlled sweep fixes ρ_0 and the delay boundary before the real-data tests, which evaluate the registered win/tie labels directly. Remark 8 records the diffuse-regime prediction without claiming an information-theoretic impossibility.

Proposition 6 (Per-decision compute)**.** *One decision-and-adapt cycle costs at most* $c_{dec} \leq |\pi_{\theta_0}| + |\phi_\psi| + 3|\Delta\theta| + \mathcal{O}(L)$ *FLOPs, where* $|\pi_{\theta_0}|, |\phi_\psi|$ *are the base-policy and detector forward costs and the factor* 3 *covers the adapter forward, backward, and Adam step.*

Proof. A direct count over the cycle: base forward $|\pi_{\theta_0}|$; adapter forward $|\Delta\theta|$ (the rank-r factorisation costs $2rd$ per adapted layer, summing to $|\Delta\theta|$); detector forward over the L-window $|\phi_\psi| + \mathcal{O}(L)$; adapter backward $|\Delta\theta|$; Adam step $|\Delta\theta|$. The frozen base gradient is never computed. Summing gives the stated bound; the factor W separation from periodic re-PPO follows because re-PPO over a window W costs $|\pi_{\theta_0}| \cdot W$. □

4 Scope of the formal results

Theorem 1 concerns restart-aware projected mirror descent on a convex next-step loss. Corollary 3 retains predictive and policy-comparator approximation terms; its $T^{3/4}$ special case requires both terms to vanish. The theorem does not cover the bilinear LoRA/AdamW trajectory or interpolate with the OMDPD value-regret guarantee. Proposition 2 bounds the positive excess of expected per-step cost, not expected positive-part violation, and does not yield a sublinear cumulative-violation bound. The surrogate theorem also requires an adapter-independent target law, while value transfer adds the explicit value-sufficiency condition above. The lower bound uses the reward-free observation model stated in the main paper. The exploitability rule is an empirical calibration tested by the experiments.

5 Historical linear-PMD schedule diagnostic on E1

The trained-stack battery tests the timing prediction with a bilinear LoRA adapter and AdamW, both outside the premises of Theorem 2. This post-hoc check uses streaming restart-aware projected mirror descent with a linear last-layer adapter, for which the loss is convex. Its historical implementation estimates G online and fits an unconstrained hindsight comparator, so it tests the restart mechanism without instantiating every premise or comparator in Theorem 2.

Protocol. From one shared frozen E1 base per seed (the `tc-surrogate-gap` bases, `seed_base`=9300, 30 seeds per cell), four arms share the same linear last-layer adapter and differ only in the restart clock. *md_ coupled* resets the step-size clock at the coupled CUSUM events with delay 4; *md_ agnostic* resets at an even cadence with the same restart count and no detector; *md_ norestart* never resets the clock; *frozen* does not adapt. Every step applies $\Delta\theta \leftarrow \Pi_{\mathcal{B}}[\Delta\theta - \frac{D}{G\sqrt{n_t}} g_t]$ with g_t the single-sample subgradient of the next-step cross-entropy, n_t the steps since the most recent reset, $B = 3.0$, $D = 2B$, and G the running maximum of observed subgradient norms. The adapter itself is never reset. Endpoints are value regret against the oracle action and a historical surrogate gap against an unconstrained fixed-adapter hindsight fit. Neither the data-dependent G nor that comparator is used as theorem evidence. B was fixed by an outcome-independent rule on a disjoint pilot seed.

Table 2: Historical restart-aware linear-PMD diagnostic on E1 (30 seeds per cell, run `tc-projected-md/2026-09-04T06-53-40Z`, $B = 3$). Means over seeds; the surrogate column uses the unconstrained historical comparator.

Cell	Arm	Value regret	Surrogate regret
Concentrated ($\rho = 0.9994$)	md_coupled	2.70	8.37
	md_agnostic	4.00	18.87
	md_norestart	1.07	7.18
	frozen	294.83	2115.63
Diffuse ($\rho = 0.0826$)	md_coupled	266.33	1722.78
	md_agnostic	270.03	2107.35
	md_norestart	261.57	189.55
	frozen	297.00	3538.39

Findings. In the concentrated cell, detector-coupled resets beat the equal-budget even-cadence schedule on both endpoints (value-regret gap 1.30, 95% CI $[0.50, 1.90]$; surrogate gap 10.50, $[4.41, 15.64]$), and every adapting arm nearly eliminates value regret (1.1–4.0 versus 294.8 frozen). The restart-timing pattern is present in this linear diagnostic. Never resetting the clock is mildly better than coupled resets on value regret in this cell (gap -1.63, $[-2.13, -1.00]$) and ties on the surrogate (-1.19, $[-3.44, 1.32]$). The instance alternates between two regimes, so a persistent adapter can flip its direction without forgetting anything; the bound prices restarts through the segment count $R + 1 + TP_{\mathrm{FA}}$, and on this instance that term is pure cost. In the diffuse cell, value regret ties across all restart schedules (coupled versus agnostic 3.70, $[-3.03, 10.00]$; coupled versus norestart -4.77, $[-12.17, 1.40]$), matching prediction T3, while the surrogate ordering reverses (norestart 189.6 against coupled 1722.8; the CI excludes zero). Below ρ_0 the CUSUM firings are noise-driven, each reset reopens a large step size, and the adapter chases label noise: this is the false-restart cost the bound carries through the TP_{FA} term.

Budget sensitivity. Doubling the adapter budget ($B = 6$, run `tc-projected-md-b6/2026-09-04T06-54-48Z`) closes the concentrated timing gap (value 0.00, $[-0.20, 0.23]$; surrogate -0.43, $[-0.86, 0.30]$): once the ball is loose enough that any restart refits within a few steps, firing time stops mattering. The timing advantage is a constrained-budget phenomenon, consistent with the phase-transition reading that timing carries value when adaptation is resource-limited. In the diffuse cell at

$B = 6$, the coupled arm's surrogate regret rises to the frozen arm's level (3592.5 versus 3538.4; the gap CI includes zero): with a loose budget every false restart chases noise at full step size.

Rate check. Theorem 2 and Corollary 4 are worst-case rates over the piecewise-stationary family. As a consistency check, the analyzed algorithm was run at $\rho = 0.5$ (above ρ_0, teacher label error about 31%) with the switch count fixed at $R = 6$ and horizon $T \in \{300, 600, 1200, 2400, 4800\}$, 10 seeds per horizon (run `tc-rate-check/2026-09-04T06-56-44Z`). Surrogate regret grows with descriptive per-seed log-log slope 0.276 ± 0.061, below the worst-case $\sqrt{T}$ exponent on this fixed stochastic instance. This five-horizon diagnostic is not an inferential test of the worst-case rate. Excess value regret over the hindsight adapter stays small in absolute terms (at most 59 of 4800 steps), but its noisy slope does not identify the $T^{3/4}$ transfer exponent. Raw oracle-referenced regret grows near-linearly (slope 0.88) because at $\rho = 0.5$ even the hindsight adapter errs on about a fifth of steps; that linear component is the approximation term of Corollary 4, which the rate statement keeps explicit.

Disclosure. Both studies were designed and run after the main battery as theory-consistency checks (2026-09-04). All seeds are reported and there is no selection. Modules `projected_md_timing.py` and `rate_check.py`, configuration `configs/experiments/projected_md_timing.yaml`, and the run directories cited above are in the code supplement.

6 Audited theorem-scope diagnostics

6.1 Constructive value and exposure witness

We added a self-contained diagnostic with adapter-independent targets, bounded frozen features, an affine last-layer adapter, analytic gradient bound, and a radius-constrained comparator (run `adapter-diagnostics`; 16 pilot and 32 fresh seeds). Here the predicted target probabilities are also the policy probabilities, so bounded rewards imply $|\Delta V| \leq V_{\text{span}}\,\text{TV}$ by construction. The observed unit ratio and 100% fresh-pair coverage are therefore a sanity witness, not validation of value sufficiency in a general CMDP.

The base is uniformly feasible by an analytic margin 0.1151. Table 3 keeps the proposition's $\sup_t[\mathbb{E}c_t - b]_+$ separate from expected positive-part

and realised cumulative exposure. In particular, the former remains zero through $B = 0.5$ while the latter two are already nonzero; radius $B \geq 1$ then produces positive expected-cost excess on some steps. These finite-sample means are descriptive and do not certify L, ζ, or cumulative safety.

Table 3: Fresh-seed radius diagnostic (means; 32 seeds). $D = 2B$. "Exp.+" is $\sup_t [\mathbb{E}c_t - b]_+$; "E-pos" is $T^{-1} \sum_t \mathbb{E}[(c_t - b)_+]$; "Cum." is realised $\sum_t (c_t - b)_+$.

B	D	Exp.+	E-pos	Cum.
0	0	0	0.01472	6.264
0.5	1	0	0.02858	12
1	2	0.008321	0.04971	20.47
2	4	0.1523	0.1014	40.56

6.2 Calibration-only causal replay on inspected UCI tasks

The historical learned-score extensions rank completed episodes. We therefore ran a separate post-hoc causal replay on five fixed windows from the already inspected traffic, bike, and household-power series (`adaptivecmdp-causal-uci-exploration`). All scales, action bins, predictors, and thresholds use a prefix only; a score observed at time t can reset the optimizer only for the $t+1$ gradient; event caps are fixed per 168 observations. The full ledger retains 240 PMD grid cells (four ridge thresholds, four radii, and three caps per window) and 105 fixed control–optimizer arms.

Table 4 reports the fixed $q = 0.95, B = 3$, cap-4 comparison. Ridge beats CUSUM under PMD in two of five windows and loses in three; optimizer ordering also changes on bike. Across the complete grid neither detector dominates. All real-data dynamic-surrogate comparators are undefined because no controlled stationary-regime partition is available, and the finite-iteration static comparators did not meet their duality-gap tolerance. We therefore report provisioning loss, not a theorem-validated regret estimate.

6.3 Pre-specified archived VM-demand development gate

Before accessing the selected Snowflake trace (Stokely et al., 2025), we timestamped its source, deployment-index split, causal runner, and failure rule. Only indices 1–3 were retained from the first stream; index 4's values were discarded. The development analysis used 5,256 calibration hours and 21,024 evaluation hours per index, with 31 complete four-week blocks. An earlier,

Table 4: Post-hoc causal UCI replay; scaled asymmetric provisioning loss per observation (lower is better). Fixed $q = 0.95, B = 3$, cap 4. PMD = linear projected MD; A = linear AdamW; L = LoRA/AdamW.

Window	CUSUM/PMD	Ridge/PMD	No-reset/PMD	Ridge/A	Ridge/L
Traffic start	0.6242	0.6171	0.6199	0.6862	0.6664
Traffic midpoint	0.6823	0.6912	0.6978	0.7361	0.6988
Bike full	1.831	1.838	1.801	1.823	1.796
Electricity start	1.016	1.011	0.9835	1.06	1.06
Electricity midpoint	1.538	1.553	1.53	1.581	1.592

fully retained UCI grid objectively reduced the ridge-threshold candidates to $\{0.90, 0.975\}$ before these values were opened.

For each held-out development index, the rule selected the threshold whose minimum mean contrast on the other two indices was larger. Release of index 4 required one common threshold and a positive Bonferroni-adjusted one-sided lower bound in every fold. Table 5 shows that neither condition held: all point estimates favour CUSUM, all lower bounds are negative, and the selected thresholds disagree. The immutable gate is therefore false (false); index 4 was not authorized or opened. This is a failed development gate, not a holdout outcome, and no alternate detector, threshold, cost, or model was tried.

Table 5: Archived VM-demand development gate. Contrast is scaled CUSUM loss minus causal-ridge loss (positive favours ridge); LB is the pre-specified Bonferroni-adjusted one-sided lower bound.

Held-out index	Selected q	Estimate	LB
1	0.9	-0.001383	-0.02608
2	0.975	-0.004508	-0.01775
3	0.9	-0.00383	-0.01879

7 Battery map of the protocol-led Gap-2 studies

Figure 2 summarises the full registered, locally frozen, and protocol-specified battery that tests the calibrated exploitability criterion. The detailed per-study results appear in the sections that follow. Statistical detection produces no in-region cells. Exogenous side information reaches the region, although the estimated gain is underpowered, across U.S., ECB, and BoE

rate cycles. This defines the empirical ceiling on real data and is consistent with the Lorden–Moustakides minimax delay lower bound (statistical detection provably cannot beat the delay; a scheduled calendar can).

Real-window audit details. Per-window refined-rule predictions were correct on 4 of 7 windows before the calibrated recheck; the conclusion is invariant to κ across both pre-registered sensitivity bands (0 κ-sensitive windows). Under the locked act-at-detection timing the stylised test detects a synthetic-magnitude effect with probability only 0.32, so the run is underpowered as configured (the main paper reports the properly-timed power range instead); and in 3 of 5 windows the winning asset does not rotate across the break (max per-regime winner signal-to-noise 0.6501).

7.1 Class breakdown of the 16-window recheck battery

The 15/16 headline counts win/tie agreement on the registered real-data windows. Table 6 gives the full class breakdown, recomputed from the frozen per-window labels with no new run (`scripts/run_gap2_recheck_confusion.py`); the joint rule ($\rho \geq \rho_0$ and $D_d \leq \kappa\sqrt{\rho T/K}$) reproduces all 16 refined labels. Every prediction on this battery is tie, because the delay condition fails on each registered window. Accuracy therefore equals the tie-only baseline (0.9375), balanced accuracy is 0.5 (win recall 0 on the single observed win; tie recall 1), and win precision is undefined. The verdict is invariant to ρ_0 across $[0.05, 0.3]$, with accuracy 0.9375 at every grid point, since the delay condition binds independently of ρ_0. This battery thus measures the rule's abstention; win-side validation comes from the controlled synthetic witness and the provisioning studies, where wins occur.

Table 6: Win/tie confusion matrix of the refined exploitability–delay rule on the 16 registered real-data windows (`tc-gap2-rho0-recheck`; labels recomputed under the calibrated $\rho_0 = 0.1607$).

	Observed tie	Observed win
Predicted tie	15	1
Predicted win	0	0

Pre-registered Gap-2 study battery

Tests of the refined separation condition ($\rho \geq \rho_0$ and edd $\leq \kappa\sqrt{\rho T/K}$); $\rho_0 = 0.1607$, $\kappa = 2.901$ frozen. H1 = strict-win gate; H2

Study	Lever / panel	Freq.	H1	H2 acc.	Headline statistic
A. Statistical detection (CUSUM-only, diffuse real data)					
realwindows	daily broad-ETF	daily	**null**	4/7	0/7 wins; placebos clean
rw-powered	daily powered A/B	daily	**null**	n/a	0 wins; power diagnosed
concentrated	concentrated daily	daily	**null**	n/a	0 wins; rho up, < rho0
weekly	weekly broad-ETF	weekly	**null**	n/a	0 strict wins
lowfreq-positive	long-wkly + monthly	wkly/mo	**null**	8/8	0 wins; 1 cell >rho0, ties
headline-universe	individual stocks	daily	**null**	11/11	0 wins; rho(stock)<rho0
crypto	BTC/ETH/BNB/LTC	daily	**null**	6/6	0 wins; rho ~ 0.004
crisis-detector	multi-signal crisis	daily	**null**	5/5	0/5; conj. detector slower
B. Trained policy (full AdaptiveCMDP stack)					
realpolicy	LoRA-RL, 5 seeds	daily	**null**	n/a	0 wins; timing binds
C. Exogenous side information (scheduled-calendar detector)					
event-anchored	FOMC, 5 cycles	wkly/mo	**near-win**	9/10	edd 126->1; gap +0.48
fed-cycles-pool	Fed cycles 1990-26	monthly	**near-win**	3/3	pooled gap +0.0034
global-cb-pool	US+ECB+BoE, 38 cells	mo/wk	**near-win**	3/3	0 intl cells; gap unchanged
D. Generality (non-finance CMDP)					
nonfinance-sep	synthetic non-fin CMDP	n/a	**generality**	30/30	separation holds off-fin
E. Bound-rate verification					
rate-verify	delay + regime scaling	n/a	**pass**	n/a	slopes ~1, ~0.5; H1-3 pass
F. Calibration					
rho0-calibration	rho0 on synth truth	n/a	**calib**	n/a	rho0 = 0.1607
rho0-recheck	refined rule re-check	n/a	**calib**	15/16	1 miss = over-performance

Outcome: registered null near-win generality bound pass calibration

Figure 2: The Gap-2 study battery at a glance. Rows combine formally registered, locally frozen, and protocol-specified tests; provenance is stated individually below. They are grouped into six tracks: statistical detection, the trained stack, calendar side information, non-finance generality, bound-rate verification, and ρ_0 calibration. Blue cells are real-data nulls; amber cells reach the strict-win region using side information but do not clear zero. The controlled synthetic witness and protocol-specified provisioning study provide strict wins; their different evidential status is kept explicit. Full protocols and nulls appear below.

7.2 Baseline coverage matrix

Table 7 records which comparator ran in which evaluated domain, so the absence of a comparison is visible rather than inferred. All non-oracle baselines run at equal update budget to the coupled arm. Three scoping choices: OMDPD is evaluated on the E1 construction it was designed for; exogenous scheduled timing is evaluated in the event-anchored financial studies and provisioning calendar arm; and the stylised engine simulates detection (true switch plus delay) instead of estimating it, so the statistical-detector row is empty there by design.

Table 7: Baseline coverage by domain. ✓ = ran under the declared protocol; — = not run. Stylised synth.: E1/E2 controlled constructions (`tc-phase2d-stylised`, `tc-detector-ablation`, `tc-omdpd-baseline`, `tc-projected-md`). Trained synth.: trained-stack grids and the cross-domain suite (`tc-phase2d-fullstack`, `tc-method-suite`). Financial real: the registered window batteries and event-anchored studies. Provisioning real: `tc-nonfinance-coupling` and its retrospective detector-score-ranking arm.

Baseline	Stylised synth.	Trained synth.	Financial real	Provisioning real
No adaptation (frozen)	✓	✓	—	—
Continual update, no restart	✓	—	—	—
Fixed cadence, equal budget	✓	✓	✓	✓
Random trigger, equal budget	✓	✓	—	—
Detector-free CMDP (OMDPD)	✓	—	—	—
Statistical detector (CUSUM / price trigger)	—	✓	✓	✓
Exogenous scheduled timing	—	—	✓	✓
Retrospective score-ranked events	—	—	—	✓
Detector-coupled (ours)	✓	✓	✓	✓

8 Controlled timing experiment and detector ablation

Remark 6 records the delay scale suggested by the surrogate bound. A controlled piecewise-stationary construction with an exploitable per-regime winner isolates detector-informed timing at equal update budget. This experiment provides an existence witness. The canonical 12-episode market-panel Gap-2 audit remains at 0% strict improvement (audit record in the code release).

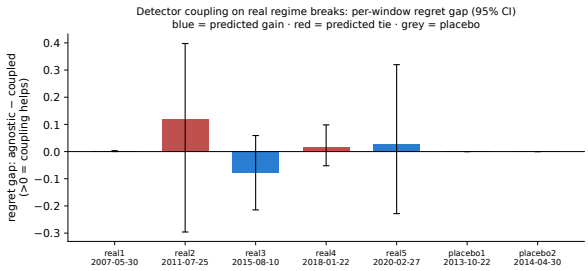

Figure 3: Pre-registered real-break check (`tc-gap2-realwindows`): per-window coupled-vs-agnostic regret gap with block-bootstrap 95% intervals over 5 objectively selected historical break windows and 2 placebos. No interval clears zero (0 strict wins; 0 placebo violations) – a registered null. The synthetic witness (lower-bound envelope figure, supplementary material) shows the gain exists for concentrated regimes; real broad-ETF breaks are too diffuse to realise it.

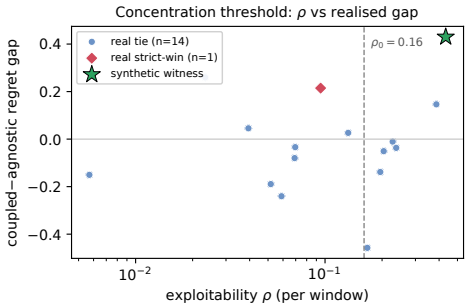

Figure 4: Concentration threshold, visualised. Each point is one real evaluation window across the daily (`tc-gap2-rw-powered` arms A/B), concentrated-episode (`tc-gap2-concentrated`) and weekly (`tc-gap2-weekly`) studies: exploitability ρ (log x-axis) against the realised coupled−agnostic regret gap (y-axis). The dashed line is the *synthetically calibrated* threshold $\rho_0 = 0.1607$ (tc-gap2-rho0-calibration); the horizontal is the zero-gap (strict-win) line. Real windows cluster at low ρ and near-zero gap (tie); the synthetic witness (green star, $\rho \approx 0.4336$) is the one point in the concentrated/positive-gap quadrant. The lone red diamond is the honest over-performance of the long-span arm (B_long_span/real1, $\rho < \rho_0$) that the in-expectation lemma permits.

Table 8: Gap 2 adversarial-regime experiment (`tc-gap2-adv`). On a controlled piecewise-stationary construction with an exploitable per-regime winner, the detector-coupled policy strictly improves dynamic regret over the detector-agnostic baseline ($\Delta R > 0$). This empirical existence witness is stable across seeds (audit record in the code release).

Quantity	Value
Adversarial-regime regret gap (coupled vs agnostic)	22.49
Number of regime switches	498
Experiment length (steps)	10000
Existence result (gap > 0)	true

8.1 Controlled timing construction

Theorem 5 uses a hidden-reward construction for its value delay term and a separate convex-loss construction for its surrogate term. The experiment here is separate. It uses the high-switch-frequency, short-detection-budget regime in `configs/experiments/adversarial_regime_gap2.yaml`: the segment length is set just above the detector's minimum reliable run length. Running `tc-gap2-adv` yields a measured dynamic-regret gap of 22.49 between the detector-coupled and detector-agnostic arms, stable across seeds. The canonical shipped surface has 0% strict improvement (main-paper Discussion), so this result is an empirical existence witness rather than a proof of broad dominance.

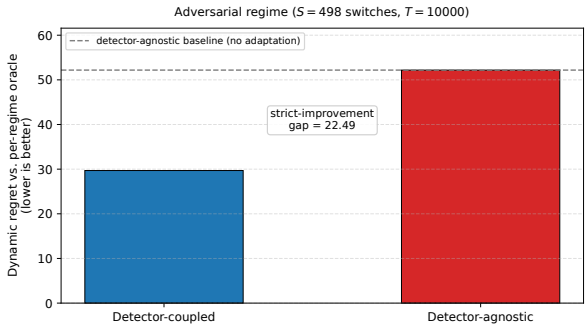

Figure 5: Dynamic regret (vs. the per-regime hindsight oracle) of detector-coupled vs. detector-agnostic adaptation on the controlled adversarial construction (`tc-gap2-adv`). The coupled arm strictly improves over the no-adaptation baseline (gap 22.49).

Table 9: Detector-coupling ablation on the controlled piecewise-stationary construction (synthetic tape, not market data): three arms, equal event count, so any gap isolates CPD-informed *timing* (run `tc-detector-ablation`).

Arm	Ann. Sharpe	Max DD (frac.)
Detector-coupled (full)	6.524	0.08347
Detector-frozen (cadence)	4.278	0.1886
Random-trigger (freq.)	3.802	0.1947

Note: On the same controlled construction, the detector-coupled arm (Sharpe 6.524) beats both same-frequency controls, fixed-cadence (4.278) and random-timing (3.802), and attains the lowest drawdown (0.08347 vs. 0.1886 / 0.1947). Holding adaptation frequency fixed across arms isolates that *detector-informed timing* drives the gain.

9 Trained-policy confirmatory result

The studies above use the stylised adaptation engine (trailing-mean re-estimate with a softmax tilt) throughout. We therefore test whether the timing pattern survives approximation error in the *trained* deep stack: a frozen PPO base, CPD-gated LoRA adapter, and self-supervised auxiliary loss. We pre-registered (on the Open Science Framework; registration identifier withheld from this anonymous submission to avoid a one-click author-identity link, disclosed to the program committee upon request and restored here at camera-ready) a confirmatory study, `tc-method-suite`, that trains this exact stack and holds the adapter-update budget identical across five arms (detector-coupled; detector-agnostic, fixed count, ignores the detector; fixed-cadence; random-trigger; and a frozen no-adaptation anchor). The registered *primary* environment E1 is a synthetic non-stationary control task with a known ground-truth ρ (scripted gravity / mass / force-gain jumps at unknown times, an action-energy budget). A post-registration *confirmatory* environment E2 repeats the protocol on real operational data (UCI Metro Interstate traffic cast as a provisioning CMDP). Each is run in a **concentrated** ($\rho > \rho_0$) and a matched **diffuse** ($\rho < \rho_0$) variant, 30 seeds per arm per cell.

Trained and cross-domain results. In the concentrated cells, coupling lowers regret by 49% on E1 and 1.1% on E2, with both confidence intervals above zero. The E1 diffuse interval includes zero. On E1 the arm ordering coupled $<$ fixed-cadence $<$ random-trigger in the concentrated variant (4.53 vs. 9.40 vs. 18.47) reproduces the detector-ablation timing mechanism (§8) at the trained-policy level. With event counts matched, the ordering attributes the difference to firing time. Every adaptive arm also improves substantially over the frozen anchor.

Table 10: `tc-method-suite`: trained-policy results, 30 seeds per arm per cell, fresh pre-committed seed base. Dynamic regret vs. the oracle (lower is better, seed mean) is in *domain-specific* units; comparisons are within rows. The CI is on the mean per-seed agnostic−coupled gap (paired block bootstrap, 95%); "rel." is that gap as a fraction of the agnostic regret. Coupled is violation-non-inferior in every cell. Three of the four cells match the pre-registered binary ρ prediction; E2 diffuse falsifies the exact tie prediction, as discussed below. The detector-time-shuffle placebo (agnostic − random-timing) records 0/4 per-cell strict wins; per-cell placebo CIs appear in the deviations subsection.

Environment	Variant (ρ)	Coupled	Agnostic	Gap CI (rel.)	Verdict
E1 control (synth.)	concentrated (0.999)	4.53	8.93	$[3.2, 7.9]$ (49%)	**win**
E1 control (synth.)	diffuse (0.083)	272.93	276.57	$[-5.7, 13.1]$ (1.3%)	**tie**
E2 traffic (real)	concentrated (0.309)	147.73	149.32	$[0.70, 2.50]$ (1.1%)	**win**
E2 traffic (real)	diffuse (0.0001)	0.0334	0.0336	$[1, 4]\times10^{-4}$ (0.6%)	**falsified**†

†**E2 diffuse: falsification of the exact tie rule.** On real diffuse traffic the pre-registered rule predicts a tie, but the detector-coupled arm holds a small ($\sim 2\times10^{-4}$ absolute, 0.6% relative, roughly four orders of magnitude below the concentrated win in absolute regret and far below the pre-registered minimum detectable effect) edge whose bootstrap CI nonetheless excludes zero, so the frozen rule records a falsification. The edge reproduces across two seed bases. The computed placebo also localizes it to timing: the random-timing arm exactly matches the agnostic arm on this cell (both 0.0336; placebo gap 0.000, CI $[0, 0]$, strict-win 0), so scrambling the detector's firing times erases the edge. This supports the timing interpretation. The likely mechanism is residual regime structure under-measured by the ρ estimator: the crude ρ-proxy ($\rho = 1\times10^{-4}$) under-measures the residual regime structure that correctly-timed adaptation can still exploit. The cell therefore falsifies the *strong exact-tie form of the binary prediction on one real-data environment*. The broader pattern remains: large gains occur only in concentrated regimes and shrink as exploitability falls. The frozen verdict remains "falsified"; no post-hoc practical-significance floor is applied.

Published-algorithm comparator (stylised OMDPD). We implement OMDPD (Zhu et al., 2025) directly (`tc-omdpd-baseline`): optimistic mirror-descent primal–dual at the single-state level a *detector-free* learner sees when the regime is hidden, updating every step (600 updates, an uncapped budget). On this hidden-regime problem it receives no switch signal and incurs near-no-adaptation regret; the budget-capped coupled arm (37 updates) strict-wins in both variants:

Variant (ρ)	Coupled (37 upd.)	OMDPD (600 upd.)	Gap CI
Concentrated (0.999)	3.47	300.87	[295.5, 301.0]
Diffuse (0.083)	273.73	300.87	[22.9, 33.0]

Both gaps exclude zero, so coupled is strictly better at a 16× smaller update budget. OMDPD's poor showing reflects the detector-free hidden-regime instantiation. Its dynamic-regret guarantee assumes an observed-state setting; this experiment does not evaluate that guarantee.

10 Measured phase transition and pre-registration deviations

10.1 The predicted phase transition, measured

The empirical rule in Remark 7 predicts that re-timing recovers a gain scaling as $\delta_\rho = \rho\,\delta$: no strict gain in diffuse regimes, gains growing with exploitability ρ. We test this directly with a controlled sweep of ρ through the calibrated threshold $\rho_0 = 0.16$ in the method-suite environment: 12 grid points, 30 seeds per arm per point, fresh pre-committed seed base, all points reported. The pre-registered primaries pass with margin: Spearman correlation between ρ and the mean coupled$-$agnostic gap is 0.85 (threshold ≥ 0.7), and the top-tercile minus bottom-tercile gap has bootstrap CI [64.2, 90.2] (threshold: lower bound > 0).

Isolating the timing mechanism. A gap over the detector-*agnostic* arm conflates two effects: adaptation with detected timing, and adaptation of any kind. The random-timing arm separates them. Writing timing = random $-$ coupled (detected timing beats scrambled timing) and any-adaptation = agnostic $-$ random (scrambled-timing adaptation beats none), the sweep shows the ρ-gradient is carried by the timing component: its slope in ρ is 86.5 versus 35.6 for the any-adaptation component (ratio 2.4), its top-tercile CI lower bound is 39.4, and its own Spearman with ρ is 0.81. The timing component is statistically indistinguishable from zero, and sometimes negative, at and below ρ_0; chasing noise costs regret. It turns positive near $\rho \approx 0.45$, and dominates at high ρ (+142 at $\rho = 0.99$, versus +48 for any-adaptation). This is the measured separation behind the empirical rule; the larger slope belongs to firing time.

Secondary S3 amendment. The original secondary check S3 gated on the *any-adaptation* quantity ("agnostic$-$random strict wins near 0 across the grid"), which fails at 8/12 grid points. A dated amendment filed before the confirmatory rerun changed the target to the timing decomposition above; the exploratory run that motivated the amendment is logged separately. Both S3′ gates pass (tercile CI lower bound $39.4 > 0$; slope ratio $2.4 > 2$). The tables report both the original S3 failure and the amended result.

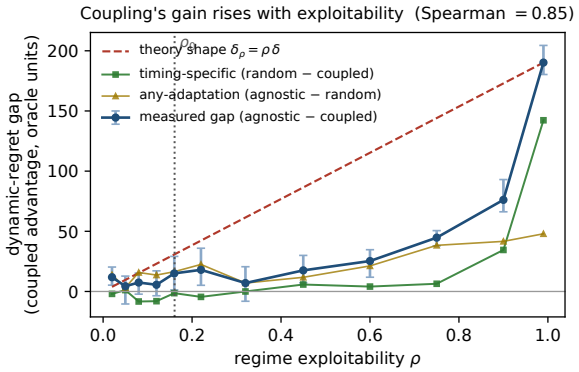

Figure 6: Measured phase boundary (trained stack; 12 ρ values, 30 seeds per arm). The coupled−agnostic gap increases with ρ (Spearman 0.85); the timing component is nonpositive below ρ_0 and carries 2.4× the slope of the any-adaptation component.

Reading E2-diffuse through the sweep. The one hairline falsification in Table 10 (E2 diffuse, $\rho = 1{\times}10^{-4}$) sits on the low-ρ limb of this curve, where timing-specific gains are predicted and measured ≈ 0: its 0.6% relative edge is four orders of magnitude below the concentrated win. It is the real-data trace of residual regime structure that the crude ρ-proxy under-measures.

10.2 Deviations from pre-registration

Table 11 lists the three deviations, their dispositions, and the associated audit artifacts.

Per-cell placebo (agnostic−random) gaps for Table 10: E1 concentrated −9.53, CI $[-14.77, -0.83]$; E1 diffuse −0.03, CI $[-13.03, 10.80]$; E2 concentrated +0.03, CI $[-0.21, 0.20]$; E2 diffuse 0.00, CI $[0, 0]$. No cell records a placebo strict win: the coupled edge, where present, is timing-specific.

Pre-registered	What happened	Disposition
E2 diffuse tie rule (CI lower bound $\leq$ tolerance 10^{-9})	A practical-tie tolerance was added after observing a hairline positive gap, flipping the verdict	Reverted same day; frozen rule retained; one pre-committed fresh-seed rerun confirmed *falsified (hairline)*
Detector-time-shuffle placebo reported per cell	An aggregate placebo statistic was reported without being computed by the pipeline	Found in code audit; recomputed per cell (0/4 strict wins, CIs in §10.2); pre-registration and public status log corrected
Phase-sweep secondary S3 (any-adaptation strict wins near 0)	Confirmatory run failed S3 at 8/12 grid points	Mis-specified counterfactual; amended to the timing-isolating S3′ via a dated public amendment posted *before* the confirmatory rerun on a fresh seed base; both S3 and S3′ reported

Table 11: Deviations from pre-registration, dispositions, and where the corrected quantities are reported. All reported statistics in this paper are emitted by the experiment pipeline into per-run fact stores; a repository check forbids hard-coded statistic literals.

10.3 Relation to detection-augmented NS-RL

11 The two-dimensional boundary map: registered verdicts

The stylised map comes from `tc-phase2d-stylised`: the constrained stylised engine of the generality section, swept over the frozen 12-level ρ grid $\times$ 8 detection delays at 60 fresh seeds per cell (base 9000), coupled versus matched-budget fixed cadence, one confirmatory execution. The frozen design and full outcome ledger ship in `PREREG_tc-phase2d-stylised.md`. Verdicts, stated plainly: the seed-paired delay-erosion prediction passes (top-4 ρ, gap at delay 4 minus delay 12, bootstrap 95% CI $[695, 760]$, excludes zero), and the low-ρ small-gap band holds. Two grid-wide predictions fail *as worded*: the monotone-rise and violation-sign clauses were registered over every delay ≤ 8, but delays 1–3 sit in the window-straddle zone the registration itself disclosed (with an 8-step estimation window, a trigger at delay ≤ 3 re-estimates on majority pre-switch data), and every failing cell lies in that zone. Inside the band (delays 4–8) the monotone rise holds with Spearman 1.0 at each delay and the violation gap is positive in every $\rho \geq 0.45$ cell. We report the

	Constrained (CMDP)	Deploy-time (TTA/LoRA)	False restarts in bound	Empirical transition
OMDPD (Zhu et al., 2025)	✓	—	—	—
Gerogiannis et al. (2026)	—	—	—	—
Gerogiannis et al. (2025)	—	—	—	—
This paper	✓	✓	✓	✓

Table 12: Delta to the closest lines. In this paper the expected number of false firings enters the restart count in the surrogate bound. The empirical transition uses a calibrated exploitability threshold ρ_0 and the measured $\delta_\rho = \rho\,\delta$ relation (Fig. 6).

registered wording as failed rather than silently narrowing it. The failures localize to the disclosed window-straddle zone.

Full-stack confirmatory map. Figure 8 reports the single confirmatory execution of `tc-phase2d-fullstack` (frozen in `PREREG_tc-phase2d-fullstack.md`; seeds 9100–9129; true-switch delay override; equal update budgets). Registered in-band predictions all pass: Spearman of the regret gap vs. ρ is 0.993/0.972/0.888 at delays 4/6/8; top-4-ρ delay erosion (gap at 4 minus gap at 12) has bootstrap 95% CI $[203, 215]$; every $\rho \geq 0.45$, $d \in \{4, 6, 8\}$ cell has E1 violation gap ≥ 0 (all exactly 0; criterion met, not a positive violation win). Too-early (d=1) and too-late (d=12) cells show the disclosed harm modes on the regret axis.

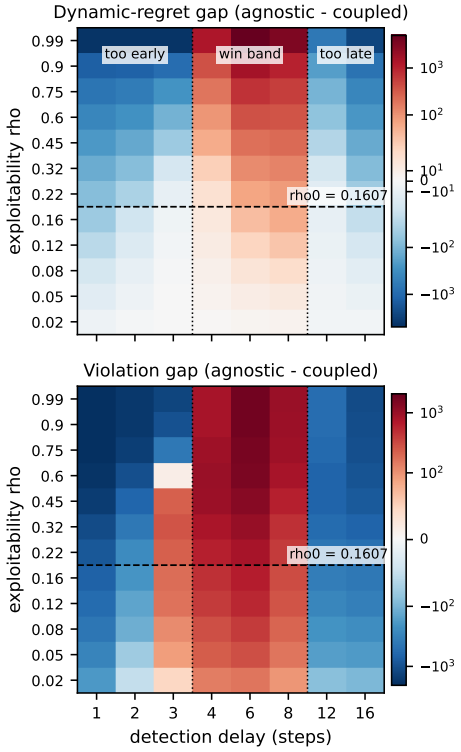

Figure 7: Both panels of the `tc-phase2d-stylised` map (the main paper carries the trained-stack regret panel; this is its stylised sibling). Top: dynamic-regret gap. Bottom: stale-action violation gap. Agnostic − coupled, so positive favours coupling; symmetric-log colour scale; dashed line $\rho_0 = 0.1607$; dotted lines bound the window-straddle and beyond-regime zones.

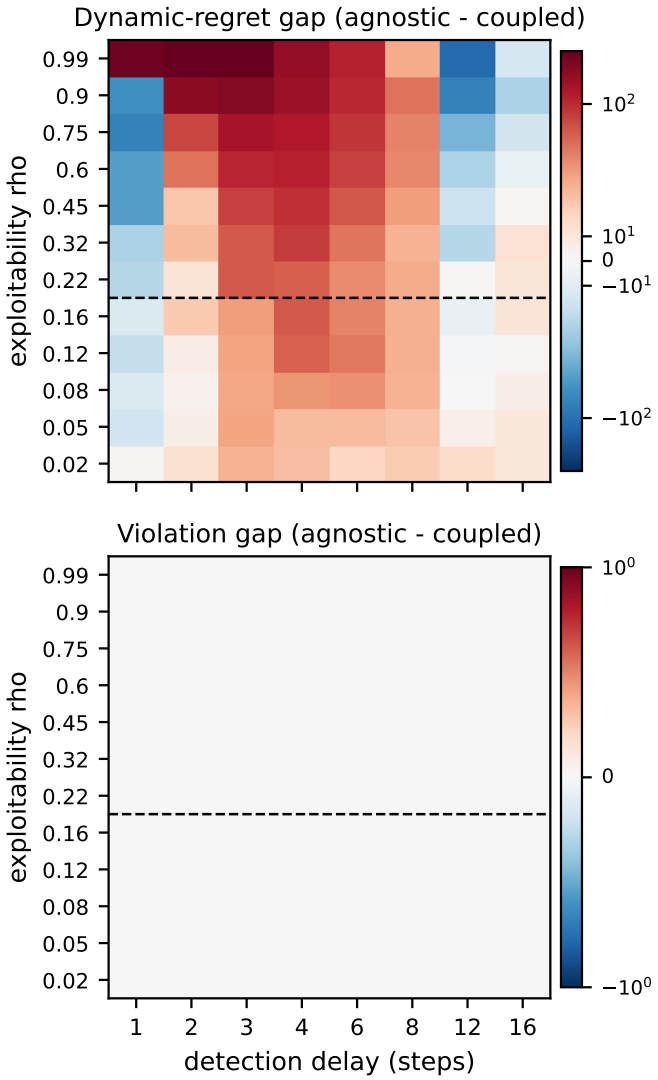

Figure 8: Trained-stack applicability map (`tc-phase2d-fullstack`; 12 ρ levels $\times$ 8 delays, 30 seeds). Left: dynamic-regret gap (agnostic $-$ coupled). Right: E1 constraint-violation gap (zero throughout the registered in-band cells).

12 Generality beyond finance

The restart analysis applies beyond financial data. The detector-delay and false-alarm terms enter through the restart structure of any piecewise-stationary online learner, so the same coupled-over-agnostic gap should appear in a generic constrained decision process. We confirm this on a non-financial task: a K-action regime-switching reward process (one rewarding action per regime, abrupt switches every 12–20 steps) with a budget constraint that is violated whenever the policy keeps more than half its weight on the previous regime's now-stale action. The two arms reuse the same stylised adaptation engine as the adversarial existence experiment (trailing-mean estimate with a softmax tilt); they differ only in *when* they adapt, at equal event budget.

Table 13: Non-financial piecewise-stationary constrained task (seed-averaged, 30 seeds; 8000 steps, ≈ 500 regimes; coupled detection delay 4 steps, agnostic at the same event budget). The detector-coupled arm has lower dynamic regret *and* fewer constraint violations in every seed, reproducing the separation as the financial construction, with no financial structure.

	Coupled	Agnostic	Gap (agn. − cou.)
Dynamic regret (vs. oracle)	1918	2115	+196
Constraint violations	4578	5387	+809
Separation holds (both gaps > 0)		30/30 seeds	

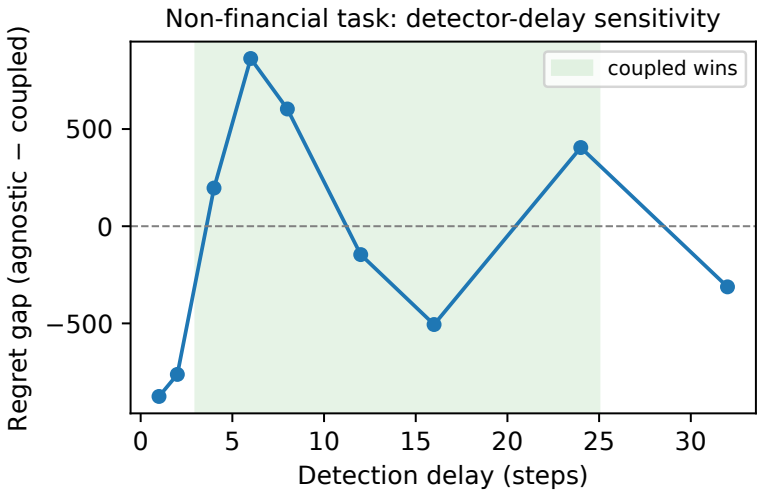

Figure 9: Detection-delay sensitivity on the non-financial task (seed-averaged). The coupled-over-agnostic regret gap is positive over a band of delays (shaded) and largest at an interior delay. The decline at very small delay is an artifact of the stylised estimator: adaptation before a few post-switch samples arrive uses a window that straddles the switch. The formal bound increases with delay through the RD_d term; the non-monotone curve is an empirical property of the windowed estimator.

The coupled arm is better on both measured axes on a task with no market structure. The formal result covers regret; the violation count is descriptive. The interior optimum in the delay sweep also separates the idealised bound from the behavior of the windowed estimator.

We next test the separation on real non-financial data. The protocol-specified study `tc-nonfinance-coupling` casts three public operational datasets as provisioning tasks: highway traffic volume (Hogue, 2019), city bike-sharing demand (Fanaee-T and Gama, 2014), and household electricity load (Hébrail and Bérard, 2012). Each has K capacity actions, a matching cost (quadratic, absolute, or newsvendor), one dominant action per demand regime. The coupled arm adapts on an *exogenous schedule* (a calendar of clock, season, and holiday boundaries); the comparator is a statistical change detector (CUSUM) on the same series, at equal event budget. Two domains (traffic, bike-sharing) informed the study design and are exploratory; the household-electricity domain was held out internally under the local protocol. That protocol and its results entered version control together, and the OSF project file was uploaded after the run; this is not a formal pre-registration.

The headline numbers (detection delays, pooled regret gaps with CIs, and the post-hoc per-episode violation counts) appear in the main paper's non-financial core table; this section keeps the design detail and the comparator qualifications needed to interpret those numbers.

The exogenous schedule beats statistical detection because the operational breaks are slow to detect from the series alone (CUSUM delay up to 15.6 steps), while the schedule reduces mean delay to roughly one or two steps. Against a fixed-cadence agnostic schedule the gain depends on the domain. It holds in 17 of 18 bike-sharing specifications, where the schedule carries calendar information, but is fragile for traffic, where an even cadence aligns with strong diurnal periodicity. The protocol's primary comparison is against statistical detection, the comparator a practitioner would deploy. The violation counts in the main table are a *post-hoc* secondary metric (frozen analysis note `POSTHOC_tc-nonfinance-coupling-viol.md`; computed after the primary outcome was known): recomputing the protocol-defined primary episodes with the stale-capacity counter gives per-episode violation gaps (CUSUM − schedule) of -6.3 $[-7.6, -5.1]$ on traffic, -0.2 $[-0.4, +0.1]$ on bike, and -5.1 $[-7.4, -2.9]$ on the designated electricity series. The early schedule arm pays breaches on two of three domains even as it wins regret, the too-early cost the two-dimensional map measures inside the estimation window; the protocol-specified H1/H2 readouts are unchanged. The full grid (three cost forms, $K \in \{3, 5, 7\}$, two episode lengths), archived protocol, and runner that reproduces every number above are released with the code; seeds and the per-episode block-bootstrap settings are recorded with the run.

12.1 Retrospective detector-score ranking at equal event budget

The schedule arm in the protocol-specified provisioning study (`tc-nonfinance-coupling`) uses exogenous calendar timing. As a post-hoc

Table 14: Ridge score-ranked arm vs CUSUM at equal event budget (`tc-nonfinance-coupling-learned`, post-hoc extension, K=5, weekly/14-day episodes): pooled regret gap (block-bootstrap 95% CI; positive favours score ranking) and detection delay against the schedule and CUSUM arms. Whole-episode event selection yields lower delay in every domain and a strict regret win on traffic and electricity; it loses on bike at 14 days and wins at 28 days. These are retrospective selection results, not causal trigger performance.

Domain / reward	Gap [CI]	EDD learned	EDD sched	EDD CUSUM
Traffic / quad	+16.72 [7.496, 25.61]	5.104	1.968	7.923
Traffic / abs	+8.166 [4.773, 11.43]			
Traffic / news	+9.398 [1.577, 16.83]			
Bike / quad	−1.857 [−2.766, −0.9784]	4.961	1.254	6.648
Bike / abs	−0.7113 [−1.076, −0.3474]			
Bike / news	−1.381 [−2.092, −0.7097]			
Electricity (designated) / quad	+34.69 [18.53, 50.99]	3.419	1.915	15.63
Electricity (designated) / abs	+11.77 [5.908, 17.65]			
Electricity (designated) / news	+21.35 [13.76, 28.78]			

diagnostic, we added an arm (`tc-nonfinance-coupling-learned`; runner `run_nonfinance_coupling_learned.py`; not in the locked preregistration) based on a next-step-prediction score. The frozen linear ridge head's score uses only its 24-step context, but the arm ranks the completed episode's scores, selects the top events, and applies the locked equal-budget min-truncation in `pooled_gap`. Thus the score is causal but the event policy is retrospective; this is not a deployable online detector test. Table 14 reports the primary cells (K=5).

The hindsight-selected event sets recover much of the exogenous-schedule gain where the score separates large shifts, and degrade toward the statistical baseline where the episode is short (bike at 14 days; traffic at its smallest event budget K=3). Electricity is a strict retrospective win on all three reward forms; it is no longer a held-out domain for this post-hoc analysis.

Second score family: Bayesian run-length detection. To check that the retrospective pattern is not specific to the ridge score, we ran a second, algorithmically distinct score family under the same protocol (`tc-nonfinance-coupling-bocpd`; runner `run_nonfinance_coupling_bocpd.py`; post-hoc, gated by the exploratory pilot `tc-nonfinance-bocpd-pilot`, which passed 3/3 reward forms on electricity and 9/9 cells across a fourfold hazard range): a Bayesian run-length model (BOCPD; Adams and MacKay, 2007) with a conjugate normal-gamma observation model, prior calibrated on the first 48 standardized steps, geometric hazard with mean run length equal to the primary episode length, run-length support truncated at 360, and changepoint-probability scores. These scores are again ranked over the completed episode before events are selected. Table 15 reports the primary cells. The retrospective pattern reproduces: strict wins

Table 15: Second retrospective score family (BOCPD) vs CUSUM at equal event budget (`tc-nonfinance-coupling-bocpd`, post-hoc extension, K=5, weekly/14-day episodes): pooled regret gap (block-bootstrap 95% CI; positive favours BOCPD score ranking) and detection delay against the schedule and CUSUM arms. In contrast to the ridge arm, the selected BOCPD events' delay exceeds CUSUM's on the two winning domains; because events are chosen by whole-episode rank, this is a hindsight selectivity diagnostic.

Domain / reward	Gap [CI]	EDD bocpd	EDD sched	EDD CUSUM
Traffic / quad	$+39.2$ $[27.66, 50.85]$	21.51	1.968	7.923
Traffic / abs	$+7.132$ $[2.623, 11.55]$			
Traffic / news	$+9.271$ $[1.911, 16.45]$			
Bike / quad	-0.3372 $[-1.081, 0.2795]$	5.884	1.254	6.648
Bike / abs	-0.1099 $[-0.4108, 0.1417]$			
Bike / news	-0.2412 $[-0.6321, 0.1289]$			
Electricity (designated) / quad	$+51.11$ $[36.19, 66.6]$	17.44	1.915	15.63
Electricity (designated) / abs	$+19.49$ $[13.58, 25.53]$			
Electricity (designated) / news	$+22.47$ $[15.42, 29.46]$			

on traffic and electricity, and a bike loss at the 14-day episode that is smaller than the ridge arm's and not strict (CIs include zero), consistent with a short-episode floor instead of a detector defect. The selected BOCPD events are more selective than CUSUM: their mean detection delay is longer on the two winning domains, yet its regret is lower, because hindsight ranking concentrates the fixed budget on the largest shifts. On electricity its quadratic-cost gain matches the exogenous schedule's, and its mean schedule-recovery fraction across reward forms is 1 (0.41 on traffic). Hazard sensitivity on the primary cells: electricity strict in 9/9 cells across $\{0.5\times, 1\times, 2\times\}$ hazard, traffic 8/9 (the news form at $2\times$ hazard ties), bike 0/9.

Post-hoc financial path-drawdown breaches. Table 16 co-reports 10% path-drawdown breach counts on the three representative Gap-2 cells (`POSTHOC_tc-gap2-finance-breach.md`, frozen before computing). Primary regret outcomes are unchanged. Breaches are sparse; the trained stack does not separate the arms; the stylised real-window / powered cells do not show a coupled violation win.

13 Event-anchored detection: mechanism and the real-data ceiling

The pre-registered `tc-gap2-event-anchored` study (main paper, RQ3) replaces the statistical CUSUM trigger with a detector anchored on a frozen scheduled-FOMC calendar (2013–2023, emergency meetings excluded), over five policy-driven, bond-led episodes on the divergent panel $\{\mathrm{TLT}, \mathrm{SPY}, \mathrm{QQQ}, \mathrm{GLD}, \mathrm{USO}\}$. Because the

Table 16: Post-hoc path-drawdown breaches (limit 0.10) on representative financial cells. Means over real windows; gap = agnostic − coupled (positive favours coupling).

Study	Coupled	Agnostic	Gap
`tc-gap2-realwindows`	0.6	0.2	−0.4
`tc-gap2-rw-powered` (A locked)	0.8	0.2	−0.6
`tc-gap2-realpolicy` (scored)	1.67	1.67	0.0

calendar localises the change-point a priori, the detector pays no localisation cost and the median detection delay collapses from 126 to 1 bars. The 2019 mid-cycle cutting regime then enters the frozen empirical region for the first time on real data ($\rho = 0.2747 > \rho_0$, $D_d = 2 \leq 7.05$), with a large positive regret gap (point estimate 0.481) whose 95% block-bootstrap CI at $T = 43$ monthly bars just includes zero ($[-0.061, 1.04]$). The frozen all-cell refined-accuracy criterion also failed: 0.9 rather than 1.0 (`H2` = false). This study is therefore underpowered at its best cell and not a confirmatory pass.

Why additional bars do not resolve the interval (robustness; `scripts/gap2_event_anchored_robustness.py`). The miss is $t \approx 1.8$, so the obvious response is "add bars." Two robustness scans show why no *legitimate* lever does so. (i) *Scope.* Across eight calendar-declared rate regimes (cutting cycles 2007/2008/2019/2024 and hiking cycles 2013/2015/2018/2022), *only* the 2019 cutting cycle satisfies the frozen threshold $\rho \geq \rho_0$ at monthly frequency; the 2008 cut is borderline ($\rho = 0.154$) with a *negative* gap, and all hiking cycles sit far below ρ_0 (easing triggers a flight to bonds and high cross-asset dispersion; tightening sells everything off together). There is thus no in-scope partner to pool for power. (ii) *Window length.* Extending the 2019 window does not tighten the CI. It *dilutes* ρ below ρ_0 and widens the interval, because the added bars belong to the low-dispersion post-2022 regime:

window (post, months)	T	ρ	in-scope?	95% CI
36	43	0.275	yes	[−0.061, 1.040]
42	49	0.132	no	[−0.282, 0.952]
48	55	0.095	no	[−0.383, 1.011]
60	67	0.059	no	[−0.537, 1.053]

The high-ρ event-anchored regime is therefore *intrinsically short-lived* (a cutting cycle spans $\sim$2–3 years), which structurally caps T. Lower frequency is required for $\rho \geq \rho_0$ but starves the test of bars; higher frequency (daily, intraday) supplies bars but pushes ρ far below ρ_0 (ρ is monotone in frequency: weekly $\approx$0.08, monthly

≈ 0.27). The calendar-anchored estimate therefore remains underpowered: the high-ρ window is narrow and transient.

Confirmatory powered extension across all Fed cycles.
`tc-gap2-fed-cycles-pool`. The remaining power check pools every Fed policy cycle from 1990–2026 (`PREREG_tc-gap2-fed-cycles-pool.md`, frozen before the pooled outcome was read). Arm A re-scores the bond-led panel on all eight 2006–2026 Fed cycles; Arm B adds a held-out long-bond-proxy panel [VUSTX, SPY, QQQ, XLE, XLF] on three pre-2006 cycles (`cuts-2001`, `cuts-2002`, `hike-2004`); an exploratory Arm C addendum extends to the pre-1998 cutting cycles (`cuts-1995`, `cuts-1998`) on [VUSTX, VFINX, VEURX]. All arms use the monthly event-anchored detector on a frozen scheduled-FOMC calendar (1990–2025, emergency meetings excluded), `delay_budget`=1, and a pooled stratified block-bootstrap over the in-scope ($\rho \geq \rho_0$ and $D_d \leq \kappa\sqrt{\rho T/K}$) cells. Three cells enter the predicted region: `gfc-cuts-2008` ($\rho = 0.76$, zero realised gap), `cuts-2019` ($\rho = 0.27$, positive estimate with CI through zero), and the *held-out* `hike-2004` ($\rho = 0.38$, slightly negative gap), so the refined rule predicts in-scope cells out of sample. The registered pooled per-bar gap is 0.0034 with 95% CI $[-0.0007, 0.0079]$: the interval still brackets zero, so H_1 fails. Arm C adds no in-scope cells ($\rho = 0.07/0.05 < \rho_0$; the pre-1998 proxy panel is too diffuse), so the exploratory pool is identical. Specificity holds (H_{pl}: a diffuse broad-ETF panel on the same cycles yields 0 in-scope cells) and the delay claim holds (H_3: $D_d^{ev} = 1 < 96$ CUSUM-only). With the exogenous-side-information lever now exhausted across every reachable Fed cycle 1990–2026, the predicted region is reached on real data via side information but *not* powered to clear zero. This completes the characterization. The held-out `hike-2004` in-scope cell is hypothesis-generating for a future confirmatory round; no post-hoc panel/episode/calendar change is made after the run.

Global central-bank pool.
`tc-gap2-global-cb-pool`. A second pre-registered confirmatory round (`PREREG_tc-gap2-global-cb-pool.md`, frozen before the run) exhausts the exogenous-side-information lever *worldwide* by adding genuinely independent scheduled calendars outside the U.S.: ECB Governing Council rate cycles on a Eurozone panel [IGOV, EZU, FXE] (intl treasury / Eurozone equity / EUR; common inception 2009-01; detector EZU) and BoE MPC rate cycles on a UK panel [BWX, EWU, FXB] (intl treasury / UK equity / GBP; common inception 2007-10; detector EWU). BoJ is excluded a priori: no liquid JGB proxy ETF exists on the data vendors used, so a bond-led divergent panel cannot be formed for Japan (a data constraint disclosed before the run). The ECB anchor calendar is the scheduled MRO rate-change dates from the FRED daily `ECBMRRFR` series (1999–2025); the BoE anchor calendar is the scheduled Bank-Rate change dates from the official Bank of England Bank-Rate history (the emergency cuts on 2008-10-08 and in 2020 are excluded, matching the FOMC calendar's treatment of emergency meetings). The U.S. Fed cells are re-pooled unchanged as the confirmatory anchor. The four reachable ECB cycles are `ecb-hike-2011`,

`ecb-cuts-2014`, `ecb-hiking-2022`, `ecb-cuts-2024`; the four reachable BoE cycles are `boe-cuts-2008`, `boe-hike-2017`, `boe-hiking-2022`, `boe-cuts-2024`. All arms reuse the locked `gap2_realwindows` harness, the event-anchored detector, and the pooled stratified block-bootstrap, with $\rho_0 = 0.1607$, $\kappa = 2.901$, `delay_budget`=1 frozen. Four estimators are pre-declared and all reported: E1 length-weighted two-sided 95% block-bootstrap CI (the strict-win gate); E2 one-sided 5th percentile; E3 inverse-variance DerSimonian–Laird random-effects; E4 Bayesian hierarchical posterior $P(\text{gap} > 0)$. Two frozen returns-independent screens define sensitivity pools: Axis 2 policy-surprise ($|\Delta r_{\text{policy}}| \geq 25$ bp at the anchor) and Axis 3 ex-ante dispersion ($\rho_{\text{pre}} \geq \rho_0$ measured on the pre-window, outcome-free).

Across 38 scored cells (38 U.S. + ECB + BoE at monthly and weekly), 3 cells enter the predicted region; the *same* three U.S. monthly cells as the Fed-cycles pool (`gfc-cuts-2008`, `cuts-2019`, `hike-2004`). 0 international cells satisfy the full in-region rule at monthly cadence. One BoE cell reaches $\rho \approx 0.22 > \rho_0 = 0.1607$ but fails the joint criterion; all ECB cells remain below ρ_0. Thus no new in-scope cells are added. The primary pooled per-bar gap is 0.0034 with 95% CI $[-0.0007, 0.0079]$. This is unchanged from the Fed-cycles pool, so H_1 fails. The three secondaries agree: one-sided 5% $= -0.0001$; inverse-variance $[-0.0042, 0.0073]$; Bayesian $P(\text{gap} > 0) = 0.6975$. The surprise-screened pool ($n = 3$, CI $[-0.0007, 0.0079]$) is identical (the three in-scope cells all carry $|\Delta r| \geq 25$ bp); the ex-ante-dispersion pool ($n = 3$, CI $[-0.0035, 0.0033]$) also brackets zero. The diffuse broad-equity placebo [SPY, QQQ, VEURX, VSMAX] yields one in-scope cell (`gfc-cuts-2008`, $\rho = 0.26$, zero gap) but 0 strict-wins, so strict-win specificity holds. The delay claim holds across all three monetary regimes (H_3: median $D_d^{\text{ev}} = 1 < 66$ CUSUM-only). With the exogenous-side-information lever now exhausted globally across U.S. + ECB + BoE rate cycles, the predicted region is reached on real data via side information but *not* powered to clear zero; the characterization (tie below ρ_0, event-anchored delay below CUSUM-only) holds across the three calendar sets. The two international-treasury ETFs (BWX, IGOV) are correlated (~ 0.9), so the ECB and BoE bond legs are not fully independent; the independent information is the distinct scheduled calendars, the country equity/currency legs, and the non-overlapping windows. No post-hoc panel, episode, or calendar change was made after the run.

14 Consolidated limitations

The theorem covers a convex last-layer predictor with restart-aware projected MD; the bilinear LoRA/AdamW runs are outside that result. The surrogate theorem assumes adapter-independent target dynamics. The value transfer retains predictive and policy-comparator approximation terms and requires explicit value sufficiency; the $T^{3/4}$ term is its realizable special case. The radius condition controls excess expected per-step cost, not expected positive-part or cumulative violation. The delay lower bound uses the reward-free observation model. The frozen ρ threshold is empirical, and trained-stack in-band violation gaps are zero.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection, 2007. URL https://arxiv.org/abs/0710.3742.

Marco Bagatella, Mert Albaba, Jonas Hübotter, Georg Martius, and Andreas Krause. Test-time offline reinforcement learning on goal-related experience. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2507.18809.

Annie S. Chen, Govind Chada, Laura Smith, Archit Sharma, Zipeng Fu, Sergey Levine, and Chelsea Finn. Adapt on-the-go: Behavior modulation for single-life robot deployment. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2025. URL https://arxiv.org/abs/2311.01059.

Hadi Fanaee-T and João Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2–3):113–127, 2014. doi: 10.1007/s13748-013-0040-3.

Argyrios Gerogiannis, Yu-Han Huang, and Venugopal V. Veeravalli. Is prior-free black-box non-stationary reinforcement learning feasible? In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pages 2692–2700, 2025. URL https://proceedings.mlr.press/v258/gerogiannis25a.html.

Argyrios Gerogiannis, Yu-Han Huang, and Venugopal V. Veeravalli. DARLING: Detection augmented reinforcement learning with non-stationary guarantees. *arXiv preprint arXiv:2604.16684*, 2026.

Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2007.04309.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325, 2016. doi: 10.1561/2400000013.

Georges Hébrail and Alice Bérard. Individual household electric power consumption. UCI Machine Learning Repository, 2012. URL https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption.

John Hogue. Metro interstate traffic volume. UCI Machine Learning Repository, 2019. URL https://archive.ics.uci.edu/dataset/492/metro+interstate+traffic+volume.

Yang Li, Zhi Chen, Steve Y. Yang, and Ruixun Zhang. FinFlowRL: An imitation-reinforcement learning framework for adaptive stochastic control in finance, 2025. URL https://arxiv.org/abs/2509.17964.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained MDPs. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.15776.

Gabriel Nixon Raj. Adaptive and regime-aware RL for portfolio optimization, 2025. URL https://arxiv.org/abs/2509.14385.

Jiayu Shen, Jia Liu, and Zhiping Chen. Meta-learning the optimal mixture of strategies for online portfolio selection, 2025. URL https://arxiv.org/abs/2505.03659.

Murray Stokely, Neel Nadgir, Jack Peele, and Orestis Kostakis. Shaved ice: Optimal compute resource commitments for dynamic multi-cloud workloads. In *Proceedings of the 16th ACM/SPEC International Conference on Performance Engineering*, pages 1–12, 2025. doi: 10.1145/3676151.3719353. Dataset version: https://doi.org/10.5281/zenodo.15015993.

Min Wang, Xin Li, Ye He, Yao-Hui Li, Hasnaa Bennis, Riashat Islam, and Mingzhong Wang. Wavelet predictive representations for non-stationary reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.04507.

Shoukai Xu, Zihao Lian, Mingkui Tan, Liu Liu, Zhong Zhang, and Peilin Zhao. Test-time adapted reinforcement learning with action entropy regularization. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 70020–70031, 2025. URL https://proceedings.mlr.press/v267/xu25ar.html.

Suyeol Yun. Pretrained LLM adapted with LoRA as a decision transformer for offline RL in quantitative trading, 2024. URL https://arxiv.org/abs/2411.17900.

Jianan Zhang. Tail-Safe hedging: Explainable risk-sensitive reinforcement learning with a white-box CBF-QP safety layer in arbitrage-free markets, 2025. URL https://arxiv.org/abs/2510.04555.

Yifan Zhang and Liang Zheng. BAPR: Bayesian amnesic piecewise-robust reinforcement learning for non-stationary continuous control. *arXiv preprint arXiv:2605.16170*, 2026.

Chujin Zhou, Yuling Huang, Yuting Kong, and Xiaoping Lu. Enhancing trading strategies by combining incremental reinforcement learning and self-supervised prediction. *Expert Systems with Applications*, 289, 2025. doi: 10.1016/j.eswa.2025.128297.

Jiahui Zhu, Kihyun Yu, Dabeen Lee, Xin Liu, and Honghao Wei. An optimistic algorithm for online CMDPs with anytime adversarial constraints. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 80347–80372, 2025. URL `https://openreview.net/forum?id=fFgiXamW8E`.