Regret Is Not Enough: Teaching and Stability in Non-Stationary Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Standard treatments of non-stationary reinforcement learning cast it as a tracking problem, tacitly accepting any policy that keeps pace with a drifting optimum and relegating instability to a minor algorithmic concern. Yet in safety-critical, value-laden domains, decisions answer to external stakeholders, and the central question becomes not just how fast we track non-stationarity, but whether the learner is teachable under drift without sacrificing performance or stability. We formalize this question in what we call the *Teaching–Regret–Stability (TRS) Principle* for *Teachable Non-stationary RL (TNRL)*. Under standard variation-budget assumptions and a Lipschitz policy-update condition, we prove a high-level theorem showing that a bounded-budget teacher can simultaneously drive the teaching error to an arbitrarily small target, keep dynamic regret sublinear, and ensure that the policy sequence remains stable on average.

1 Introduction

Reinforcement learning (RL) in non-stationary environments has become a central abstraction for systems that must operate under continual change: recommendation platforms facing shifting user populations, robotic controllers subjected to wear-and-tear, or online decision systems exposed to evolving markets. A large body of work formalizes non-stationarity via *variation budgets* on rewards and transitions, and evaluates algorithms through *dynamic regret* (Besbes et al., 2015; Cheung et al., 2020; Fei et al., 2020; Zhou et al., 2022; Feng et al., 2023; Wei et al., 2023; Cheng et al., 2023). In this view, the environment is exogenous, nature drifts arbitrarily within the budget, and the learner is rewarded for tracking the moving optimum as closely as possible.

This perspective hides two implicit and rarely questioned assumptions:

1. **Any low-regret behavior is acceptable.** If the algorithm attains small dynamic regret, we declare it "good" without asking what policy it is converging to, or whether that policy aligns with any external goal beyond cumulative reward.
2. **Instability is a technical artifact.** Large step-to-step policy changes are seen as an optimization nuisance—something to be controlled in proofs or smoothed in practice, but not as a first-class object of study.

Both assumptions are benign for abstract adversarial benchmarks, but they become problematic once we acknowledge the presence of *stakeholders*. In real deployments, there is almost always a designer, regulator, or user who has a *target policy* in mind: a safe driving style for an autonomous car, a fair treatment policy for a clinical decision system, or a conservative trading strategy for a financial agent. These stakeholders can intervene by shaping rewards, modifying transitions (e.g., through interfaces, safety layers, or constraints), or filtering data—but such interventions carry costs and may themselves interact with non-stationarity.

From tracking to teachability. This leads to a different foundational question: *is the learner teachable under non-stationarity*? More concretely:

> Given a non-stationary environment and a target policy $\pi^\dagger$, can a teacher with bounded ability to poison the environment steer a standard RL algorithm toward $\pi^\dagger$, while preserving low dynamic regret in the true environment and maintaining a stable policy trajectory?

If the answer is "no", low dynamic regret is a dangerously incomplete certificate: the algorithm may track a moving optimum that is catastrophically misaligned with any reasonable target, and it may do so in a violently unstable way. Conversely, if the answer is "yes" under transparent structural conditions, then *teachability* becomes a new lens on the design of non-stationary RL algorithms: we can ask not only "how fast do they learn?" but "how gracefully can they be taught?"

Three literatures, one missing bridge. Pieces of this picture exist in three separate lines of work. Non-stationary RL with variation budgets provides dynamic-regret guarantees under drifting environments (Cheung et al., 2020; Fei et al., 2020; Zhou et al., 2022; Feng et al., 2023; Wei et al., 2023; Cheng et al., 2023). Policy teaching and environment poisoning study how an attacker or teacher can manipulate transitions and rewards to induce a desired policy, typically in stationary MDPs (Rakhsha et al., 2020). Algorithmic stability quantifies how smoothly updates react to perturbations, and has been linked to generalization in supervised learning (Hardt et al., 2016). However, there is currently no framework that *jointly* reasons about:

- the *teaching error* between the learned policy and an externally specified target policy;
- the *dynamic regret* in the *true* non-stationary environment, rather than in the manipulated one; and
- the *stability* of the policy sequence under non-stationarity and poisoning.

As a result, the field lacks a principled answer to the question of whether non-stationary RL algorithms are fundamentally teachable, or whether dynamic-regret guarantees can mask structurally misaligned behavior.

The TRS Principle: Teachable Non-stationary RL. In this work we propose a unifying viewpoint that we call the *Teaching–Regret–Stability (TRS) Principle* for *Teachable Non-stationary Reinforcement Learning (TNRL)*. We consider an episodic non-stationary MDP with variation budgets on rewards and transitions (§2), and introduce a teacher that can *poison* the environment prior to each episode by perturbing transitions and rewards at a per-episode cost. The teacher has a total poisoning budget C, and aims to teach a fixed target policy $\pi^\dagger$ by running a standard RL algorithm on the poisoned environments. We then focus on three metrics:

1. **Teaching error** $\mathrm{Mismatch}_K$ (equation 9): the average distance between the learner's policy and the target policy.
2. **Dynamic regret** DynReg_K (equation 10): the cumulative regret measured in the *true* drifting environments, not in the poisoned ones.
3. **Policy stability** Stab_K (equation 11): the average step-to-step change of the learner's policy.

Our main theorem (Theorem 1) shows that, under natural assumptions on the non-stationary MDP, the poisoning budget, the dynamic-regret guarantee of the base algorithm, and a Lipschitz-type policy update, there exists a teacher strategy and an RL algorithm such that:

- the teaching error $\mathrm{Mismatch}_K$ can be driven arbitrarily close to a target ε,
- the dynamic regret DynReg_K remains sublinear in the number of episodes, and
- the policy sequence is stable on average, with Stab_K controlled by the total environment drift and the poisoning budget.

We interpret this as a *three-dimensional trade-off frontier*—the TRS Principle—that any *teachable* non-stationary RL system must inhabit.

A structural failure mode of regret-only evaluation. The TRS viewpoint also exposes a disaster scenario for the standard paradigm. It is easy to construct non-stationary environments in which two algorithms achieve indistinguishable dynamic regret, yet one converges to a target policy that satisfies safety or fairness desiderata, while the other converges to a policy that is catastrophically misaligned. Dynamic regret alone cannot distinguish these runs; nor can a stability measure that ignores *what* is being stabilized. In our synthetic contextual-bandit experiments (§4.3), a no-teacher baseline attains dynamic regret comparable to that of a budgeted teacher, but exhibits substantially larger teaching error. In any application where $\pi^\dagger$ encodes safety or compliance constraints, this is the difference between a benign and a disastrous system—a difference that is completely invisible to regret-only metrics.

Contributions. Formally and conceptually, our contributions are as follows:

- We introduce **Teachable Non-stationary RL (TNRL)** and the **Teaching–Regret–Stability (TRS) Principle**, which jointly reason about teaching error, dynamic regret in the true environment, and policy stability under environment poisoning.
- We prove a **Teaching–Regret–Stability theorem** (Theorem 1) in episodic non-stationary MDPs with variation budgets and a Lipschitz policy-update assumption. The theorem shows that a bounded-budget teacher can align the learner with a target policy while keeping dynamic regret sublinear and the policy sequence stable.
- We instantiate the framework in a **non-stationary contextual bandit** with a synthetic generator and a Discounted LinUCB learner, and empirically probe the TRS frontier. The experiments demonstrate that modest poisoning budgets can significantly reduce teaching error at a mild regret cost, while preserving or even improving stability, in line with the theoretical scaling laws.

Taken together, these results invite a shift in how we think about non-stationary RL. Rather than asking only whether algorithms track the environment well, the TRS Principle asks whether they are *teachable*: can they be steered, at bounded cost, to stable policies that embody the values and constraints of their users, even as the world drifts?

2 Preliminaries

We consider an episodic, non-stationary Markov decision process (MDP) with finite state and action spaces

$$\mathcal{S},\ \mathcal{A}, \qquad |\mathcal{S}| = S,\ |\mathcal{A}| = A.$$

Time is partitioned into K episodes, each of horizon H, so that the total number of interaction steps is $T = KH$.

Non-stationary MDP sequence. Episode $k \in \{1, \ldots, K\}$ is associated with an MDP

$$M_k = (\mathcal{S}, \mathcal{A}, P_k, r_k, \rho_1, \gamma),$$

where

- $P_k(\cdot \mid s, a)$ is the transition kernel,
- $r_k(s, a) \in [0, 1]$ is the reward function,
- ρ_1 is a fixed initial-state distribution, and
- $\gamma \in (0, 1]$ is a discount factor (for finite-horizon problems one may set $\gamma = 1$).

The non-stationarity of the environment is captured by a variation budget over episodes:

$$V_P := \sum_{k=2}^{K} \sup_{s,a} \big\| P_k(\cdot \mid s,a) - P_{k-1}(\cdot \mid s,a) \big\|_1, \tag{1}$$

$$V_r := \sum_{k=2}^{K} \sup_{s,a} \big| r_k(s,a) - r_{k-1}(s,a) \big|, \tag{2}$$

and we write the total environment drift as

$$V_{\mathrm{env}} \;:=\; V_P + V_r. \tag{3}$$

Teacher and environment poisoning. We introduce a teacher (or attacker) that can modify the environment prior to each episode by applying environment poisoning. Concretely, before episode k begins, the teacher chooses a modified MDP

$$\tilde{M}_k = (\mathcal{S}, \mathcal{A}, \tilde{P}_k, \tilde{r}_k, \rho_1, \gamma),$$

where $\tilde{P}_k$ and $\tilde{r}_k$ may differ from the true P_k and r_k.

We measure the poisoning cost in episode k by

$$c_k := \sup_{s,a} \big\| \tilde{P}_k(\cdot \mid s,a) - P_k(\cdot \mid s,a) \big\|_1 + \sup_{s,a} \big| \tilde{r}_k(s,a) - r_k(s,a) \big|. \tag{4}$$

The teacher has a total poisoning budget

$$C_{\mathrm{tot}} := \sum_{k=1}^{K} c_k \;\le\; C. \tag{5}$$

The modified MDP sequence $\{\tilde{M}_k\}_{k=1}^{K}$ has its own variation budget

$$V_{\mathrm{eff}} := \sum_{k=2}^{K} \sup_{s,a} \big\| \tilde{P}_k(\cdot \mid s,a) - \tilde{P}_{k-1}(\cdot \mid s,a) \big\|_1 + \sum_{k=2}^{K} \sup_{s,a} \big| \tilde{r}_k(s,a) - \tilde{r}_{k-1}(s,a) \big|. \tag{6}$$

By construction, we always have

$$V_{\mathrm{eff}} \;\lesssim\; V_{\mathrm{env}} + C, \tag{7}$$

up to universal constants.

Learner and target policy. A reinforcement learning algorithm $\mathcal{A}$ interacts with the poisoned environments $\{\tilde{M}_k\}_{k=1}^{K}$. At the beginning of episode k, the learner selects a (possibly stochastic) policy $\pi_k(\cdot \mid s)$ based on its past observations.

We fix a target policy $\pi^\dagger$ (e.g., stationary) that the teacher aims to teach. We measure the distance between two policies π and π' via

$$d(\pi, \pi') := \sup_{s \in \mathcal{S}} \big\| \pi(\cdot \mid s) - \pi'(\cdot \mid s) \big\|_1. \tag{8}$$

The average teaching error (or policy mismatch) over K episodes is defined as

$$\mathrm{Mismatch}_K := \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\big[d(\pi_k, \pi^\dagger) \big], \tag{9}$$

where the expectation is taken over the randomness of the learning algorithm and the environment.

Dynamic regret in the true environment. All learning happens in the poisoned environments $\tilde{M}_k$, but performance is ultimately evaluated in the true environments M_k.

Let $V_k(\pi)$ denote the expected total return of policy π in the true environment M_k, starting from the initial distribution ρ_1:

$$V_k(\pi) := \mathbb{E}_{M_k,\pi}\Big[\sum_{t=1}^{H} \gamma^{t-1} r_k(s_t, a_t)\Big].$$

We define the per-episode optimal policy

$$\pi_k^\star \in \operatorname{argmax}_\pi V_k(\pi),$$

and the dynamic regret of the learner as

$$\mathrm{DynReg}_K := \sum_{k=1}^{K} \big(V_k(\pi_k^\star) - V_k(\pi_k)\big). \tag{10}$$

Policy stability. We measure the stability of the learner's policy sequence by the average step-to-step change:

$$\mathrm{Stab}_K := \frac{1}{K-1} \sum_{k=2}^{K} \mathbb{E}\big[d(\pi_k, \pi_{k-1})\big]. \tag{11}$$

Low Stab_K indicates that the learner's policy evolves smoothly over time, whereas a large value suggests frequent drastic changes.

Goal. The central question is: under bounded environment drift (V_{env} bounded) and bounded poisoning budget ($C_{\mathrm{tot}} \leq C$), can one design a teacher strategy and a learning algorithm such that

- the teaching error $\mathrm{Mismatch}_K$ is small (successful teaching),
- the dynamic regret DynReg_K in the true environments is controlled (no catastrophic loss in performance), and
- the policy sequence is stable, as quantified by Stab_K.

We state the assumptions used in our main result. They are chosen so as to be compatible with existing work on non-stationary RL and policy teaching.

Assumption 1 (Non-stationary MDP with bounded variation)**.** *The true environments $\{M_k\}_{k=1}^K$ satisfy the variation budget constraints equation 1–equation 3 with $V_{\mathrm{env}} \leq B_{\mathrm{env}}$ for some known constant B_{env}. Moreover, each M_k is communicating and has bounded diameter $D < \infty$.*

Assumption 2 (Bounded poisoning budget)**.** *The teacher generates a sequence of poisoned environments $\{\tilde{M}_k\}_{k=1}^K$ satisfying the per-episode cost equation 4 and total budget constraint equation 5 with $C_{\mathrm{tot}} \leq C$.*

Assumption 3 (RL algorithm with dynamic regret guarantee)**.** *There exists a reinforcement learning algorithm $\mathcal{A}$ such that for any environment sequence with effective variation budget V_{eff} (cf. equation 6), the dynamic regret with respect to $\{\tilde{M}_k\}$ satisfies*

$$\mathrm{DynReg}_K^{(\tilde{M})} \;\leq\; \tilde{\mathcal{O}}\big(K^{2/3} V_{\mathrm{eff}}^{1/3}\big), \tag{12}$$

where $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors. This scaling is known to be minimax-optimal in non-stationary online learning with variation budgets[1] *and has been achieved up to logarithmic and problem-dependent factors in several non-stationary RL settings.*[2]

[1] see, e.g., Besbes et al. (2015)

[2] see, e.g., Gajane et al. (2018), Fei et al. (2020), Mao et al. (2021), and Zhao et al. (2022).

Assumption 4 (Teachability and convergence in the canonical MDP)**.** *There exist a canonical MDP M^{c} and a target policy $\pi^\dagger$ such that:*

(i) ε-robust optimality. *There exists $\varepsilon > 0$ such that*

$$V^{\mathrm{c}}(\pi^\dagger) \;\geq\; V^{\mathrm{c}}(\pi) + 2\varepsilon \qquad \textit{for all policies } \pi \neq \pi^\dagger,$$

where $V^{\mathrm{c}}(\pi)$ denotes the value of π in M^{c}.

(ii) Convergence of $\mathcal{A}$ in M^{c}. *When run on the fixed environment M^{c}, the RL algorithm $\mathcal{A}$ achieves vanishing average regret with respect to $\pi^\dagger$:*

$$\frac{1}{K}\sum_{k=1}^{K}\Big(V^{\mathrm{c}}(\pi^\dagger) - V^{\mathrm{c}}(\pi_k)\Big) \;\to\; 0 \qquad \textit{as } K \to \infty.$$

(iii) Teachability via bounded poisoning. *For the true environment sequence $\{M_k\}_{k=1}^K$, there exists a sequence of poisoned environments $\{\tilde{M}_k\}_{k=1}^K$ with total poisoning budget*

$$C_{\mathrm{tot}} \;:=\; \sum_{k=1}^{K} c_k \;\leq\; C,$$

such that, when $\mathcal{A}$ is run on $\{\tilde{M}_k\}_{k=1}^K$, the induced policy sequence $\{\pi_k\}_{k=1}^K$ satisfies, for some constant $c_{\mathrm{teach}} > 0$ independent of K,

$$\mathrm{Mismatch}_K := \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\big[d(\pi_k, \pi^\dagger)\big] \;\leq\; \varepsilon + c_{\mathrm{teach}}\frac{C}{K}, \qquad \forall K \geq 1. \tag{13}$$

Assumption 5 (Lipschitz policy update)**.** *There exist constants $L_P, L_r \geq 0$ and $\alpha \in [0, 1)$ such that for all $k \geq 2$,*

$$\begin{aligned} d(\pi_k, \pi_{k-1}) \leq \alpha\, d(\pi_{k-1}, \pi_{k-2}) + L_P \sup_{s,a}\big\|\tilde{P}_k(\cdot \mid s,a) - \tilde{P}_{k-1}(\cdot \mid s,a)\big\|_1 \\ + L_r \sup_{s,a}\big|\tilde{r}_k(s,a) - \tilde{r}_{k-1}(s,a)\big|. \end{aligned} \tag{14}$$

Remark 1 (On Assumptions 4 and 5)**.** *Assumption 4 postulates the existence of a canonical environment in which the target policy $\pi^\dagger$ is separated from all competing policies by a fixed value margin (ε-robust optimality), and in which the learning algorithm $\mathcal{A}$ enjoys vanishing average regret. This is satisfied by many standard tabular RL algorithms (e.g., Q-learning with suitable step sizes, optimistic model-based methods), and ensures that $\pi^\dagger$ can in principle be learned in a benign stationary environment. The third item further requires that a teacher with bounded poisoning budget C can make the true non-stationary environment "look like" M^{c} from the learner's perspective so that the average teaching error obeys equation 13.*

Assumption 5 captures two properties of the policy update. First, in a fixed environment ($\tilde{P}_k = \tilde{P}_{k-1}$ and $\tilde{r}_k = \tilde{r}_{k-1}$), the update is contractive with factor $\alpha < 1$, so that policy changes decay over time. Second, when the environment drifts between episodes, the induced policy change is Lipschitz in the size of the drift, with sensitivities L_P and L_r to changes in the transition kernel and reward function, respectively. Together, these properties allow us to control the cumulative policy variation in terms of the total environment variation and the poisoning budget.

3 Main Theorem

We now state a high-level theorem that captures the trade-off between teaching success, dynamic regret, and policy stability in non-stationary MDPs under bounded poisoning.

The first step is to show that, there exists a sequence of poisoned environments $\{\tilde{M}_k\}$ with total cost $C_{\mathrm{tot}} \leq C$ such that the learner's policies $\{\pi_k\}$ converge towards the target policy $\pi^\dagger$ when running $\mathcal{A}$ on $\{\tilde{M}_k\}$. This can be formalized as follows.

Lemma 1 (Teaching feasibility)**.** *Under Assumption 4, there exists a sequence of poisoned environments* $\{\tilde{M}_k\}$ *satisfying* $C_{\text{tot}} \le C$ *and an RL algorithm* $\mathcal{A}$ *such that*

$$\text{Mismatch}_K = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\big[d(\pi_k, \pi^\dagger)\big] \;\le\; \varepsilon + \tilde{\mathcal{O}}\Big(\frac{1}{K}\Big). \tag{15}$$

The proof follows standard arguments in the environment poisoning literature. One constructs a canonical MDP M^{c} in which $\pi^\dagger$ is robustly optimal, and then defines the poisoned environments $\tilde{M}_k$ to gradually steer the learner's observations and rewards towards those induced by M^{c}. Robust optimality of $\pi^\dagger$ ensures that small deviations in transitions and rewards (controlled by the poisoning budget) do not change the identity of the optimal policy. The convergence guarantee of $\mathcal{A}$ in the canonical environment then implies that the sequence $\{\pi_k\}$ approaches $\pi^\dagger$, which yields equation 15.

Next, we analyze the dynamic regret of the learner with respect to the poisoned environment sequence $\{\tilde{M}_k\}$. By equation 6–equation 7 and Assumptions 1–2, the effective variation budget satisfies

$$V_{\text{eff}} \;\lesssim\; V_{\text{env}} + C.$$

Lemma 2 (Dynamic regret in the poisoned environment)**.** *Under Assumption 3, the dynamic regret of* $\mathcal{A}$ *with respect to the poisoned environment sequence* $\{\tilde{M}_k\}$ *satisfies*

$$\text{DynReg}_K^{(\tilde{M})} \;\le\; \tilde{\mathcal{O}}\big(K^{2/3}(V_{\text{env}} + C)^{1/3}\big). \tag{16}$$

This is a direct consequence of the assumed dynamic regret bound equation 12 and the upper bound $V_{\text{eff}} \lesssim V_{\text{env}} + C$ on the effective variation budget of the poisoned environment sequence.

We then relate $\text{DynReg}_K^{(\tilde{M})}$ to the dynamic regret DynReg_K in the true environment sequence $\{M_k\}$.

Lemma 3 (Regret transfer to the true environment)**.** *Under Assumptions 1–2, we have*

$$\text{DynReg}_K \;\le\; \text{DynReg}_K^{(\tilde{M})} + \mathcal{O}(C). \tag{17}$$

For each episode k and policy π, the difference between the value functions in M_k and $\tilde{M}_k$ can be bounded using standard perturbation arguments for MDPs:

$$\big|V_k(\pi) - \tilde{V}_k(\pi)\big| \;\le\; \mathcal{O}(c_k),$$

where $\tilde{V}_k(\pi)$ is the value in $\tilde{M}_k$. Summing over episodes and applying triangle inequalities yields equation 17.

Combining Lemmas 2 and 3 yields the bound equation 20 in Theorem 1.

Finally, we analyze the stability of the policy sequence using the Lipschitz update property in Assumption 5.

Lemma 4 (Stability bound)**.** *Under Assumptions 2 and 5, the average policy change satisfies*

$$\text{Stab}_K \;\le\; \frac{L_P + L_r}{1-\alpha} \cdot \frac{V_{\text{env}} + C}{K} + \mathcal{O}\Big(\frac{1}{K}\Big). \tag{18}$$

Unrolling the recursion equation 14 yields

$$d(\pi_k, \pi_{k-1}) \le \alpha^{k-2} d(\pi_2, \pi_1) + \sum_{j=2}^{k} \alpha^{k-j}\Big(L_P \Delta_j^P + L_r \Delta_j^r\Big),$$

where $\Delta_j^P := \sup_{s,a} \|\tilde{P}_j(\cdot \mid s,a) - \tilde{P}_{j-1}(\cdot \mid s,a)\|_1$ and $\Delta_j^r := \sup_{s,a} |\tilde{r}_j(s,a) - \tilde{r}_{j-1}(s,a)|$. Averaging over k, using the geometric series bound $\sum_{k \ge j} \alpha^{k-j} \le 1/(1-\alpha)$ and the fact that $\sum_j (\Delta_j^P + \Delta_j^r) \lesssim V_{\text{env}} + C$ yields equation 18. The $\mathcal{O}(1/K)$ term comes from the initial transient.

Theorem 1 (Teaching–Regret–Stability Trade-off in Non-stationary MDPs)**.** *Suppose Assumptions 1–5 hold. Then there exists a teacher strategy $\{\tilde{M}_k\}_{k=1}^K$ with total poisoning budget $C_{\text{tot}} \le C$ and a reinforcement learning algorithm $\mathcal{A}$ such that the following properties hold for any $\varepsilon > 0$, provided C is larger than a problem-dependent threshold $C_{\min}(\varepsilon)$:*

(i) ***Teaching success.*** *The average mismatch between the learner's policy and the target policy satisfies*

$$\text{Mismatch}_K \;\le\; \varepsilon + \tilde{\mathcal{O}}\Big(\frac{1}{K}\Big). \tag{19}$$

(ii) ***Dynamic regret in the true environment.*** *The dynamic regret measured with respect to the true environment sequence $\{M_k\}_{k=1}^K$ satisfies*

$$\text{DynReg}_K \;\le\; \tilde{\mathcal{O}}\Big(K^{2/3}(V_{\text{env}} + C)^{1/3} \;+\; C\Big), \tag{20}$$

where V_{env} is defined in equation 3. In particular, if $V_{\text{env}} = o(K)$ and $C = o(K)$, then $\text{DynReg}_K = o(K)$ and hence the average regret $\text{DynReg}_K/K \to 0$ as $K \to \infty$.

(iii) ***Policy stability.*** *The average policy change satisfies*

$$\text{Stab}_K \;\le\; \frac{L_P + L_r}{1-\alpha} \cdot \frac{V_{\text{env}} + C}{K} + \mathcal{O}\Big(\frac{1}{K}\Big). \tag{21}$$

The theorem states that under bounded environment drift and bounded poisoning budget, one can (i) successfully teach a target policy (up to an arbitrarily small ε), while (ii) keeping the dynamic regret in the true environment sublinear in K, and (iii) ensuring that the resulting policy sequence is stable on average.

4 The TRS Principle

We now instantiate our framework in a controlled synthetic environment and empirically probe the joint teaching–regret–stability trade-off predicted by Theorem 1. All experiments are conducted in a non-stationary contextual bandit model, which can be viewed as a horizon-one special case of a non-stationary MDP.

4.1 Non-stationary Contextual Bandit Environment

Model. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a context (feature) space and $\mathcal{A} = \{1, \dots, A\}$ a finite action set. At each round $t = 1, \dots, T$,

1. the environment draws a context $x_t \in \mathcal{X}$,
2. the learner selects an action $a_t \in \mathcal{A}$ according to a policy $\pi_t(\cdot \mid x_t)$,
3. the learner observes a stochastic reward $r_t(a_t) \in [0, 1]$ for the chosen action only.

There is no state transition beyond a single step; each round is an independent episode of length one. We write $\bar{r}_t(x, a) := \mathbb{E}[r_t(a) \mid x_t = x]$ for the (possibly time-varying) mean reward of action a at context x and time t.

Definition 1 (Non-stationary contextual bandit)**.** *A non-stationary contextual bandit environment is a sequence*

$$\mathcal{E} = \big\{\mathcal{D}_t,\, \bar{r}_t : \mathcal{X} \times \mathcal{A} \to [0, 1]\big\}_{t=1}^T,$$

where $\mathcal{D}_t$ is a distribution over contexts $x_t \in \mathcal{X}$ and $\bar{r}_t$ is the mean reward function at round t. Non-stationarity is captured by the fact that either $\mathcal{D}_t$ or $\bar{r}_t$ (or both) may change with t.

In this subsection we focus on non-stationarity in the reward mechanism and keep the marginal context distribution fixed, i.e., $\mathcal{D}_t \equiv \mathcal{D}$ for all t.

Assumption 6 (Bounded contexts and rewards)**.** *There exists $R_x > 0$ such that $\|x_t\|_2 \le R_x$ almost surely for all t, and rewards are bounded in $[0,1]$, i.e., $r_t(a) \in [0,1]$ almost surely for all t and a. Moreover, the noise is conditionally σ^2-sub-Gaussian (Boucheron et al. (2003); Vershynin (2018); Lattimore & Szepesvári (2020)):*

$$r_t(a) = \bar{r}_t(x_t, a) + \xi_t, \qquad \mathbb{E}[\xi_t \mid x_t] = 0, \quad \mathbb{E}\big[e^{\lambda \xi_t} \mid x_t\big] \le \exp\big(\tfrac{\sigma^2 \lambda^2}{2}\big)$$

for all $\lambda \in \mathbb{R}$.

To quantify non-stationarity we impose a variation budget on the sequence of mean reward functions.

Assumption 7 (Variation budget on mean rewards)**.** *Let $\|\cdot\|_\infty$ denote the supremum norm over $\mathcal{X} \times \mathcal{A}$. The environment satisfies a reward-variation budget $B_r \ge 0$ if*

$$\sum_{t=1}^{T-1} \big\|\bar{r}_{t+1} - \bar{r}_t\big\|_\infty \;\le\; B_r. \tag{22}$$

Assumption 7 is the contextual-bandit analogue of the variation-budget conditions commonly used for non-stationary MDPs. It allows abrupt or gradual changes in the reward structure, as long as the total drift over time is bounded by B_r.

Linear parametrization. For concreteness in our experiments, we instantiate $\bar{r}_t$ via a time-varying linear model. We fix a feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ and a sequence of parameter vectors $\theta_t \in \mathbb{R}^d$, and set

$$\bar{r}_t(x, a) \;=\; \sigma\big(\langle \theta_t, \phi(x, a)\rangle\big), \tag{23}$$

where $\sigma(\cdot)$ is a 1-Lipschitz squashing function such as the logistic sigmoid or a clipped identity. In this case the variation budget is controlled by the path length of $\{\theta_t\}_{t=1}^T$,

$$\sum_{t=1}^{T-1} \|\theta_{t+1} - \theta_t\|_2 \;\le\; B_\theta,$$

which implies equation 22 up to the Lipschitz constants of ϕ and σ.

From parameter path length to reward variation. The variation budget B_r in Assumption 7 is imposed directly in the reward space via

$$\sum_{t=1}^{T-1} \big\|\bar{r}_{t+1} - \bar{r}_t\big\|_\infty \;\le\; B_r.$$

Under the linear parametrization equation 23, this budget can be controlled by the path length of the parameter sequence $\{\theta_t\}_{t=1}^T$.

We assume that the squashing function $\sigma : \mathbb{R} \to [0,1]$ is L_σ-Lipschitz and that the feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ is uniformly bounded in norm:

$$\big|\sigma(u) - \sigma(v)\big| \;\le\; L_\sigma |u - v| \quad \forall u, v \in \mathbb{R}, \qquad \sup_{x \in \mathcal{X},\, a \in \mathcal{A}} \|\phi(x, a)\|_2 \;\le\; L_\phi < \infty. \tag{24}$$

For instance, the logistic sigmoid is 1/4-Lipschitz, and linear or one-hot feature maps are uniformly bounded after rescaling.[3]

Define the parameter path length

$$B_\theta \;:=\; \sum_{t=1}^{T-1} \big\|\theta_{t+1} - \theta_t\big\|_2. \tag{25}$$

We then have the following simple control of the reward variation by B_θ.

[3] See, e.g., Vershynin (2018); Lattimore & Szepesvári (2020) for standard Lipschitz and boundedness assumptions in linear bandit models.

Algorithm 1 Synthetic non-stationary contextual bandit generator

```
Require: Dimension $d$, number of actions $A$, horizon $T$, number of segments $M$, drift scale $\eta > 0$, context
    covariance $\Sigma_x$ (typically $I_d$).
Set segment length $L \leftarrow \lfloor T/M \rfloor$.
Initialize $\theta^{(1)} \sim \mathcal{N}(0, \sigma_\theta^2 I_d)$.
for $m = 2$ to $M$ do
    Draw a drift vector $\Delta^{(m)} \sim \mathcal{N}(0, \eta^2 I_d)$.
    Set $\theta^{(m)} \leftarrow \theta^{(m-1)} + \Delta^{(m)}$.
end for
Fix a feature map $\phi : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$ (e.g., $\phi(x, a)$ concatenates $x$ with a one-hot encoding of $a$).
for $t = 1$ to $T$ do
    Let $m \leftarrow 1 + \lfloor (t-1)/L \rfloor$ be the current segment index.
    Sample a context $x_t \sim \mathcal{N}(0, \Sigma_x)$.
    for each action $a \in \mathcal{A}$ do
        Compute the mean reward
            $\bar{r}_t(x_t, a) \;\leftarrow\; \sigma\big(\langle \theta^{(m)}, \phi(x_t, a) \rangle\big).$
        Draw noise $\xi_t(a)$ (e.g., $\mathcal{N}(0, \sigma^2)$) and set
            $r_t(a) \;\leftarrow\; \mathrm{clip}\big(\bar{r}_t(x_t, a) + \xi_t(a),\, 0,\, 1\big).$
    end for
end for
```

Lemma 5 (Lipschitz control of reward variation)**.** *Under equation 23 and equation 24, the reward variation budget satisfies*

$$B_r \;\le\; L_\sigma L_\phi B_\theta. \tag{26}$$

Thus, in our linear contextual bandit model the reward variation budget B_r is controlled by the parameter path length B_θ up to the Lipschitz constants L_σ and L_ϕ, a standard pattern in path-length analyses of non-stationary online learning (Besbes et al., 2015; Hazan et al., 2016).

Synthetic generator for experiments. We now specify a concrete synthetic generator that we will use in the experiments.

By construction, the mean reward function $\bar{r}_t$ is piecewise-stationary with M segments, and the total variation in equation 22 is controlled (via Lemma 5) by the path length B_θ of the parameter sequence. In our generator, the parameter vector is constant within each segment and evolves as

$$\theta^{(m)} = \theta^{(m-1)} + \Delta^{(m)}, \qquad \Delta^{(m)} \sim \mathcal{N}(0, \eta^2 I_d),$$

so that B_θ only accumulates at segment boundaries. Writing $Z \sim \mathcal{N}(0, I_d)$, we have

$$\mathbb{E}\big[B_\theta\big] = \sum_{m=2}^{M} \mathbb{E}\big\|\Delta^{(m)}\big\|_2 = (M-1)\,\eta\,\mathbb{E}\big\|Z\big\|_2 \;\le\; c_d\, M\eta,$$

where $c_d := \mathbb{E}\|Z\|_2$ depends only on the dimension d. In particular, for fixed d the expected path length $\mathbb{E}[B_\theta]$ grows *on the order of* $M\eta$, so varying (M, η) induces different effective variation budgets B_r through the Lipschitz relation equation 26.

4.2 Learners and Teaching Strategies

Learner. Throughout the experiments, the base learner A is instantiated as Discounted LinUCB [4], which satisfies Assumption 3 under a standard variation-budget condition on the non-stationarity; Concretely, A

[4] see, e.g., Russac et al. (2019) for formal guarantees.

can be instantiated by any standard non-stationary linear contextual bandit algorithm with $O\big(T^{2/3}V_{\text{eff}}^{1/3}\big)$ dynamic regret.

Teacher and poisoning budget. We consider two types of teacher strategies:

- **No-teacher baseline.** A reference setting with poisoning budget $C = 0$, where the learner interacts with the true non-stationary environment sequence $\{(D_t, \bar{r}_t)\}_{t=1}^{T}$ without any intervention. This baseline illustrates the dynamic regret and stability behavior of A under pure non-stationarity.
- **Budgeted teacher.** A teacher with total poisoning budget $C > 0$ constructs a sequence of poisoned environments $\{\tilde{\mathcal{E}}_t\}_{t=1}^{T}$ that stays within the budget constraint and gradually morphs the observed rewards towards those of a canonical environment in which the target policy $\pi^\dagger$ is robustly optimal (cf. Assumption 4). The construction follows the environment-poisoning view used in the analysis: the teacher perturbs rewards in a way that remains indistinguishable at small scales but consistently biases the learner towards $\pi^\dagger$.

In all experiments we vary the budget C while keeping the underlying non-stationarity (M, η) fixed, in order to disentangle the effect of environment drift from the effect of teacher interventions.

4.3 Experimental protocol

We report the three metrics defined in equation 9, equation 10, and equation 11: the teaching error Mismatch_K, the dynamic regret DynReg_K (and its per-round version DynReg_K/K), and the stability measure Stab_K. These are exactly the quantities that appear in the trade-off of Theorem 1, whose bounds are given in equation 19–equation 21.

We use the synthetic generator in Algorithm 1 with horizon $T = K$, segment counts $M \in \{1, 5, 20\}$, drift scales $\eta \in \{0, 0.1, 0.3\}$, and a fixed linear feature map ϕ as in the parametrization equation 23. The base learner A is instantiated as Discounted LinUCB, which satisfies the dynamic regret condition in Assumption 3 under a standard variation-budget condition on the non-stationarity; in particular, its regret matches the $K^{2/3}V_{\text{eff}}^{1/3}$ scaling in equation 12, where the effective variation budget V_{eff} for the poisoned environments satisfies equation 6–equation 7. For each choice of (M, η) we vary the total poisoning budget C through the normalized fraction $C_{\text{frac}} := C/T \in \{0, 0.05, 0.10, 0.20\}$. The case $C_{\text{frac}} = 0$ corresponds to the no-teacher baseline, while $C_{\text{frac}} > 0$ activates the budgeted teacher described in Section 4.1 and Assumptions 2 and 4. All results are averaged over 5 random seeds; we report means and standard deviations.

4.4 Global Teaching–Regret–Stability Trade-off

By Theorem 1, the three metrics Mismatch_K, DynReg_K, and Stab_K defined in equation 9, equation 10, and equation 11 satisfy the bounds equation 19–equation 21: for fixed $T = K$ we expect (i) teaching error that can be driven down to ε up to an $O(1/K)$ term, (ii) average regret DynReg_K/K that grows sublinearly in C through the $(V_{\text{env}} + C)^{1/3}$ factor in equation 20, and (iii) stability that is controlled by $(V_{\text{env}} + C)/K$ as in equation 21, where V_{env} is the environment drift defined in equation 3.

Table 1 summarizes the global behavior of the three metrics as we vary the normalized budget C_{frac}, averaging over all non-stationarity configurations (M, η) and seeds. We report means ± standard deviations over all (M, η) and seeds. Dynamic regret grows mildly with the budget, while teaching error decreases and stability remains essentially unchanged, in line with equation 19–equation 21.

Two patterns stand out and mirror the structure of equation 19–equation 21.

(1) Teaching buys alignment at a modest regret cost. As we increase the normalized budget from $C_{\text{frac}} = 0$ to 0.20, the average teaching error Mismatch_K (defined in equation 9) drops from approximately 1.46 to 1.26, a relative reduction of about 13%. Over the same range, the cumulative dynamic regret DynReg_K (defined in equation 10) increases from roughly 1.8×10^3 to 2.0×10^3, corresponding to only a $\sim 9\%$ increase in DynReg_K and a similarly mild increase in the average regret DynReg_K/K. In other words,

Table 1: Global teaching–regret–stability trade-off as the normalized poisoning budget $C_{\text{frac}} = C/T$ increases.

C_{frac}	DynReg_K	DynReg_K/K	Mismatch_K	Stab_K
0.00	1793 ± 162	$3.59\times10^{-2} \pm 3.2\times10^{-3}$	1.46 ± 0.48	0.194 ± 0.062
0.05	1812 ± 203	$3.62\times10^{-2} \pm 4.1\times10^{-3}$	1.41 ± 0.50	0.193 ± 0.056
0.10	1844 ± 272	$3.69\times10^{-2} \pm 5.4\times10^{-3}$	1.36 ± 0.52	0.192 ± 0.050
0.20	1962 ± 456	$3.92\times10^{-2} \pm 9.1\times10^{-3}$	1.26 ± 0.52	0.192 ± 0.045

a small but well-structured amount of poisoning significantly improves how closely the learner tracks the target policy $\pi^\dagger$, while keeping the overall regret very close to the no-teacher baseline.

This behavior is consistent with equation 20: for fixed K and moderate budgets, the upper bound

$$\frac{\text{DynReg}_K}{K} \lesssim K^{-1/3}(V_{\text{env}} + C)^{1/3} + \frac{C}{K},$$

with V_{env} from equation 3, predicts that increasing C by a constant factor should only moderately increase the average regret, especially when the environment variation already dominates. Empirically, we see exactly this regime: the teacher can "spend" up to 20% of the horizon on poisoning without destroying the dynamic-regret guarantees of the base algorithm in Assumption 3.

(2) Stability is preserved—and often improved in practice. Perhaps surprisingly, the stability metric Stab_K defined in equation 11 remains essentially flat as we increase C_{frac}. Across all non-stationarity configurations, the average value of Stab_K stays around 0.19, and the standard deviation actually shrinks slightly when the budget increases. The worst-case bound in Theorem 1(iii), equation 21, only guarantees that Stab_K scales with $(V_{\text{env}}+C)/K$ through the Lipschitz update condition in Assumption 5; it does not rule out the possibility that teaching might *improve* stability by steering the learner toward a fixed target policy. Our experiments show that this benign behavior is typical in the synthetic contextual bandit: the teacher reduces large oscillations by pulling the policy sequence toward $\pi^\dagger$, so the contractive term $\alpha d(\pi_k, \pi_{k-1})$ in equation 14 dominates and smooths the trajectory.

4.5 Effect of Environment Non-stationarity

We next examine how the trade-off behaves as we vary the non-stationarity of the environment. Recall that in our generator the total variation budget on the mean rewards is controlled by the number of segments M and the drift scale η through Assumption 7 and equation 22: piecewise-stationary instances with larger M and η correspond to larger effective variation budgets B_r.

For the no-teacher baseline ($C_{\text{frac}} = 0$), we observe that DynReg_K and DynReg_K/K remain stable across all (M, η), with changes well below 10% even when we move from a stationary single-segment environment ($M = 1$, $\eta = 0$) to highly non-stationary cases ($M = 20$, $\eta = 0.3$). This matches the intuition behind the $K^{2/3}V_{\text{eff}}^{1/3}$ scaling in equation 12: in the finite-horizon regime we explore, the variation budgets induced by our choices of (M, η) are not large enough to dominate the $K^{2/3}$ term, so the regret curves are relatively flat.

When we fix a non-zero budget (e.g., $C_{\text{frac}} = 0.10$) and vary (M, η), we again see an essentially stable dynamic regret and stability, while the teaching error Mismatch_K shows only mild dependence on the drift level. This suggests that, in this regime, the teacher's cost C is the dominant contribution to the effective variation budget $V_{\text{eff}} \lesssim V_{\text{env}}+C$ in equation 7: once C is fixed, moderate changes in V_{env} do not qualitatively change the trade-off.

4.6 Summary and Implications

Overall, the synthetic experiments give a clean empirical picture of the teaching–regret–stability frontier predicted by Theorem 1.

- A small poisoning budget C is enough to substantially improve alignment with a target policy $\pi^{\dagger}$ (small $\mathrm{Mismatch}_K$ in equation 9), while keeping dynamic regret close to that of a strong non-stationary bandit baseline (sublinear DynReg_K as in equation 20).

- The learner's policy sequence remains stable on average; in fact, the teacher can make the policy *smoother* by suppressing large, purely non-stationary-driven jumps, consistently with the Lipschitz stability guarantee equation 21.

- The qualitative behavior is robust across a range of non-stationarity levels, controlled by (M, η) and Assumption 7, indicating that the guarantees of Theorem 1 are not an artifact of a particular environment, but capture a genuine phenomenon in non-stationary RL with poisoning.

From a higher-level perspective, this highlights the importance of our framework: dynamic regret alone (via equation 10) is blind to *what* policy is being learned, and stability alone (via equation 11) does not protect against converging to a bad policy. By explicitly coupling teaching error, regret, and stability through equation 19–equation 21, we obtain a genuinely three-dimensional view of non-stationary RL. The experiments show that this joint control is quantitatively achievable with standard algorithms and a simple teacher, which we view as a first step toward *teachable* non-stationary RL systems in the wild.

5 Related Work

Non-stationary RL with variation budgets has been studied in tabular MDPs (Cheung et al., 2020; Fei et al., 2020; Mao et al., 2025), linear and structured settings (Zhou et al., 2022; Feng et al., 2023; Wei et al., 2023; Cheng et al., 2023), and risk-sensitive or constrained formulations (Ding et al., 2023; Wei et al., 2023). Our formulation follows this line by adopting variation budgets on rewards and transitions.

Policy teaching and environment poisoning against RL were formalized in (Rakhsha et al., 2020), which characterized the feasibility and cost of teaching arbitrary target policies by manipulating rewards and transitions. Our work combines such teaching mechanisms with non-stationary RL dynamic-regret bounds.

Finally, our stability condition is inspired by the algorithmic stability literature (Hardt et al., 2016), where Lipschitz-type update rules are used to control generalization error. Here, a similar idea quantifies the smoothness of policy updates in the face of non-stationarity and poisoning.

6 Discussion

Experiments in non-stationary contextual bandits confirm that this triple control is not a vacuous worst-case guarantee. With a simple Discounted LinUCB learner and a budgeted teacher, we observe that modest poisoning budgets lead to substantial reductions in teaching error at a mild regret cost, while preserving—and sometimes improving—stability. At the same time, the no-teacher baseline exposes a structural failure mode of regret-only evaluation: it achieves similar dynamic regret yet remains significantly misaligned with the target policy, a discrepancy that would be disastrous in safety- or compliance-critical domains.

Viewed in this light, the TRS Principle is less a new algorithm and more a *diagnostic*: a way to ask whether a non-stationary RL system is truly teachable under drift. It suggests several directions for future work. On the theoretical side, one can seek sharper bounds, weaker regularity conditions, and extensions to partially observable or multi-agent settings. On the algorithmic side, it is natural to design learners that are explicitly optimized for teachability—for example, by encouraging Lipschitz policy updates or exposing interfaces that make low-cost teaching strategies easier to implement. Finally, at a broader level, the TRS viewpoint invites the community to revisit a foundational assumption: that "performance" in non-stationary RL can be fully captured by regret. Our results suggest a different answer: in a world where environments drift and stakeholders care about *which* policies are learned, non-stationary RL must be judged not only by how well it tracks, but by how gracefully it can be taught.

7 Conclusion

The dominant story about non-stationary reinforcement learning has been one of tracking: environments drift, algorithms adapt, and success is declared when dynamic regret remains small. In this work we argued that this story is incomplete. It overlooks the presence of a teacher—a designer, regulator, or user—and it treats the content and stability of the learned policy as secondary concerns. As a result, an algorithm can look excellent under standard benchmarks while remaining fundamentally *unteachable*: no reasonable amount of shaping or poisoning can reliably steer it toward a desired, stable behavior.

We proposed an alternative narrative, embodied in the *Teaching–Regret–Stability (TRS) Principle* for *Teachable Non-stationary RL (TNRL)*. In our framework, a teacher with a bounded poisoning budget interacts with a non-stationary MDP and a standard RL algorithm. We quantify three coupled objectives: teaching error with respect to a fixed target policy, dynamic regret in the *true* drifting environment, and the stability of the policy trajectory. Our main theorem shows that, under natural structural assumptions, there exists a teacher strategy and an algorithm for which all three quantities can be controlled simultaneously: the teaching error can be driven to an arbitrarily small target, dynamic regret remains sublinear, and the policy sequence evolves smoothly on average.

References

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

Yuan Cheng, Jing Yang, and Yingbin Liang. Provably efficient algorithm for nonstationary low-rank mdps. *Advances in Neural Information Processing Systems*, 36:6330–6372, 2023.

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*, pp. 1843–1854. PMLR, 2020.

Yuhao Ding, Ming Jin, and Javad Lavaei. Non-stationary risk-sensitive reinforcement learning: Near-optimal dynamic regret, adaptive detection, and separation design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7405–7413, 2023.

Yingjie Fei, Zhuoran Yang, Zhaoran Wang, and Qiaomin Xie. Dynamic regret of policy optimization in non-stationary environments. *Advances in Neural Information Processing Systems*, 33:6743–6754, 2020.

Songtao Feng, Ming Yin, Ruiquan Huang, Yu-Xiang Wang, Jing Yang, and Yingbin Liang. Non-stationary reinforcement learning under general function approximation. In *International Conference on Machine Learning*, pp. 9976–10007. PMLR, 2023.

Pratik Gajane, Ronald Ortner, and Peter Auer. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*, 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2 (3-4):157–325, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In *International conference on machine learning*, pp. 7447–7458. PMLR, 2021.

Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Başar. Model-free nonstationary reinforcement learning: Near-optimal regret and applications in multiagent reinforcement learning and inventory control. *Management Science*, 71(2):1564–1580, 2025.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *International Conference on Machine Learning*, pp. 7974–7984. PMLR, 2020.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Honghao Wei, Arnob Ghosh, Ness Shroff, Lei Ying, and Xingyu Zhou. Provably efficient model-free algorithms for non-stationary cmdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 6527–6570. PMLR, 2023.

Peng Zhao, Long-Fei Li, and Zhi-Hua Zhou. Dynamic regret of online markov decision processes. In *International Conference on Machine Learning*, pp. 26865–26894. PMLR, 2022.

Huozhi Zhou, Jinglin Chen, Lav R Varshney, and Ashish Jagmohan. Nonstationary reinforcement learning with linear function approximation. *Transactions on Machine Learning Research*, 2022, 2022.

A Additional Theoretical Results

A.1 Proof of Lemma 4

We provide a complete proof of the stability bound stated in Lemma 4.

Proof of Lemma 4. For brevity, write

$$d_k := d(\pi_k, \pi_{k-1}), \qquad \Delta_k^P := \sup_{s,a} \left\| \tilde{P}_k(\cdot \mid s, a) - \tilde{P}_{k-1}(\cdot \mid s, a) \right\|_1,$$

$$\Delta_k^r := \sup_{s,a} \left| \tilde{r}_k(s, a) - \tilde{r}_{k-1}(s, a) \right|.$$

By Assumption 5, for all $k \geq 3$ we have

$$d_k \;\leq\; \alpha\, d_{k-1} + L_P \Delta_k^P + L_r \Delta_k^r. \tag{27}$$

Step 1: Unrolling the recursion. We first show by induction that for all $k \geq 3$,

$$d_k \;\leq\; \alpha^{k-2} d_2 + \sum_{j=2}^{k} \alpha^{k-j} \big(L_P \Delta_j^P + L_r \Delta_j^r \big). \tag{28}$$

For $k = 3$, equation 27 gives

$$d_3 \leq \alpha d_2 + L_P \Delta_3^P + L_r \Delta_3^r,$$

which coincides with equation 28 for $k = 3$. Assume equation 28 holds for some $k \geq 3$. Then, using equation 27,

$$\begin{aligned} d_{k+1} &\leq \alpha d_k + L_P \Delta^P_{k+1} + L_r \Delta^r_{k+1} \\ &\leq \alpha\Big(\alpha^{k-2} d_2 + \sum_{j=2}^{k} \alpha^{k-j}(L_P \Delta^P_j + L_r \Delta^r_j)\Big) + L_P \Delta^P_{k+1} + L_r \Delta^r_{k+1} \\ &= \alpha^{k-1} d_2 + \sum_{j=2}^{k} \alpha^{(k+1)-j}(L_P \Delta^P_j + L_r \Delta^r_j) + \alpha^0 (L_P \Delta^P_{k+1} + L_r \Delta^r_{k+1}) \\ &= \alpha^{(k+1)-2} d_2 + \sum_{j=2}^{k+1} \alpha^{(k+1)-j}(L_P \Delta^P_j + L_r \Delta^r_j), \end{aligned}$$

which is exactly equation 28 with k replaced by $k+1$. Thus equation 28 holds for all $k \geq 3$.

Step 2: Bounding the average policy change. By definition,

$$\text{Stab}_K = \frac{1}{K-1} \sum_{k=2}^{K} d_k = \frac{1}{K-1}\Big(d_2 + \sum_{k=3}^{K} d_k\Big).$$

Using equation 28 for $k \geq 3$, we obtain

$$\begin{aligned} \sum_{k=2}^{K} d_k &\leq d_2 + \sum_{k=3}^{K}\Big(\alpha^{k-2} d_2 + \sum_{j=2}^{k} \alpha^{k-j}(L_P \Delta^P_j + L_r \Delta^r_j)\Big) \\ &= d_2 \sum_{k=2}^{K} \alpha^{k-2} + \sum_{k=3}^{K}\sum_{j=2}^{k} \alpha^{k-j}(L_P \Delta^P_j + L_r \Delta^r_j). \end{aligned}$$

We bound the two terms separately. For the first term, since $\alpha \in [0,1)$, the geometric series is bounded:

$$\sum_{k=2}^{K} \alpha^{k-2} = \sum_{m=0}^{K-2} \alpha^m \leq \frac{1}{1-\alpha},$$

so

$$\frac{1}{K-1} d_2 \sum_{k=2}^{K} \alpha^{k-2} \leq \frac{d_2}{(1-\alpha)(K-1)} = \mathcal{O}\Big(\frac{1}{K}\Big).$$

For the second term, define

$$S := \sum_{k=3}^{K}\sum_{j=2}^{k} \alpha^{k-j}(L_P \Delta^P_j + L_r \Delta^r_j).$$

We exchange the order of summation:

$$S = \sum_{j=2}^{K} (L_P \Delta^P_j + L_r \Delta^r_j) \sum_{k=j}^{K} \alpha^{k-j}.$$

Again using $\alpha \in [0,1)$ and the geometric series,

$$\sum_{k=j}^{K} \alpha^{k-j} = \sum_{m=0}^{K-j} \alpha^m \leq \frac{1}{1-\alpha},$$

hence

$$S \leq \frac{1}{1-\alpha} \sum_{j=2}^{K} (L_P \Delta^P_j + L_r \Delta^r_j).$$

Therefore

$$\mathrm{Stab}_K \le \frac{1}{K-1} \cdot \frac{1}{1-\alpha} \sum_{j=2}^{K} (L_P \Delta_j^P + L_r \Delta_j^r) + \mathcal{O}\Big(\frac{1}{K}\Big).$$

Using $L_P \Delta_j^P + L_r \Delta_j^r \le (L_P + L_r)(\Delta_j^P + \Delta_j^r)$ and the fact that, up to universal constants,

$$\sum_{j=2}^{K} (\Delta_j^P + \Delta_j^r) \;\lesssim\; V_{\mathrm{env}} + C$$

(Assumptions 1 and 2), we obtain

$$\mathrm{Stab}_K \;\le\; \frac{L_P + L_r}{1-\alpha} \cdot \frac{V_{\mathrm{env}} + C}{K-1} + \mathcal{O}\Big(\frac{1}{K}\Big),$$

which is equivalent to equation 18 after replacing $K-1$ by K in the denominator. □

Remark 2. *The parameter $\alpha \in [0,1)$ in Assumption 5 plays the role of a contraction factor for the policy update: when the environment is fixed ($\Delta_k^P = \Delta_k^r = 0$), the recursion $d_k \le \alpha d_{k-1}$ implies that successive policy changes decay at rate α^k. The bound $\sum_{k=j}^{K} \alpha^{k-j} \le (1-\alpha)^{-1}$ used above is the standard geometric-series estimate associated with this contraction.*

A.2 Proof of Lemma 5

Proof. Fix t and (x,a). Then

$$\begin{aligned} \big|\bar{r}_{t+1}(x,a) - \bar{r}_t(x,a)\big| &= \big|\sigma(\langle \theta_{t+1}, \phi(x,a)\rangle) - \sigma(\langle \theta_t, \phi(x,a)\rangle)\big| \\ &\le L_\sigma \big|\langle \theta_{t+1} - \theta_t, \phi(x,a)\rangle\big| \quad \text{(by Lipschitzness of } \sigma) \\ &\le L_\sigma \|\theta_{t+1} - \theta_t\|_2 \|\phi(x,a)\|_2 \quad \text{(Cauchy–Schwarz)} \\ &\le L_\sigma L_\phi \|\theta_{t+1} - \theta_t\|_2, \end{aligned}$$

where we used the uniform bound on $\|\phi(x,a)\|_2$ in the last step. Taking the supremum over (x,a) yields

$$\big\|\bar{r}_{t+1} - \bar{r}_t\big\|_\infty \;\le\; L_\sigma L_\phi \|\theta_{t+1} - \theta_t\|_2.$$

Summing over $t = 1, \dots, T-1$ gives

$$\sum_{t=1}^{T-1} \big\|\bar{r}_{t+1} - \bar{r}_t\big\|_\infty \;\le\; L_\sigma L_\phi \sum_{t=1}^{T-1} \|\theta_{t+1} - \theta_t\|_2 \;=\; L_\sigma L_\phi B_\theta,$$

which is exactly equation 26. □

A.3 Expected path length of the synthetic generator

Lemma 6 (Expected path length of the synthetic generator)**.** *In Algorithm 1, let B_θ be defined as in equation 25. Then there exists a constant $c_d > 0$ depending only on the dimension d such that $\mathbb{E}[B_\theta] \le c_d M \eta$.*

A.4 Evaluation Metrics

We report three metrics that correspond directly to the quantities in Theorem 1.

Teaching error (policy mismatch). We measure how well the teacher succeeds at steering the learner towards the fixed target policy $\pi^\dagger$ by the average mismatch

$$\mathrm{Mismatch}_K = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\big[d(\pi_k, \pi^\dagger)\big],$$

where $d(\cdot,\cdot)$ is the ℓ_1 distance between action distributions, taken uniformly over contexts.

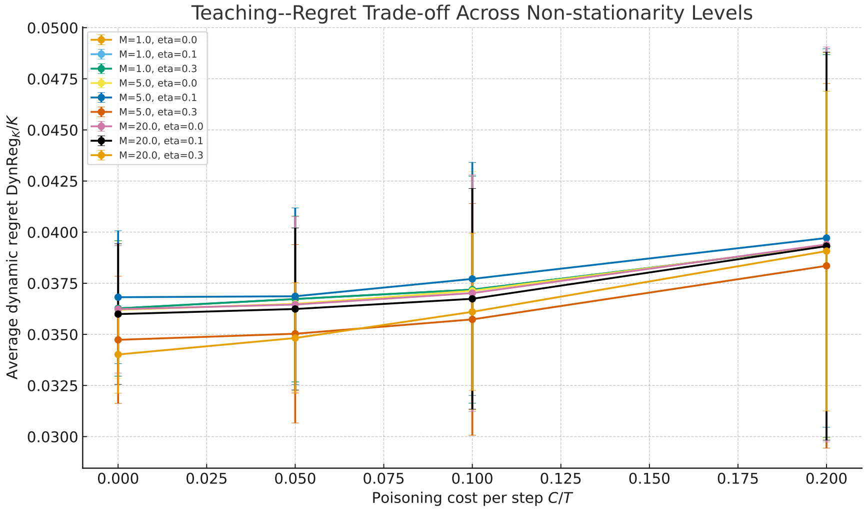

Figure 1: Average dynamic regret DynReg_K/K versus normalized poisoning cost per step C/T. Each curve corresponds to a different non-stationarity configuration (M, η); error bars denote standard deviation over seeds.

Dynamic regret in the true environment. Although all updates happen in the (possibly poisoned) environments, performance is evaluated in the true non-stationary environment. We therefore track the cumulative dynamic regret

$$\mathrm{DynReg}_K = \sum_{k=1}^{K} \big(V_k(\pi_k^\star) - V_k(\pi_k) \big),$$

where $V_k(\pi)$ denotes the expected return of policy π in the true environment at episode k, and $\pi_k^\star$ is the per-episode optimal policy. We report both the cumulative regret and the average regret DynReg_K/K.

Policy stability. Finally, we quantify the smoothness of the policy trajectory via the average step-to-step change

$$\mathrm{Stab}_K = \frac{1}{K-1} \sum_{k=2}^{K} \mathbb{E}\big[d(\pi_k, \pi_{k-1}) \big].$$

Small values of Stab_K indicate that the learner updates its policy gradually over time, whereas large values point to unstable behavior with frequent drastic shifts.

B Additional Experimental Figures

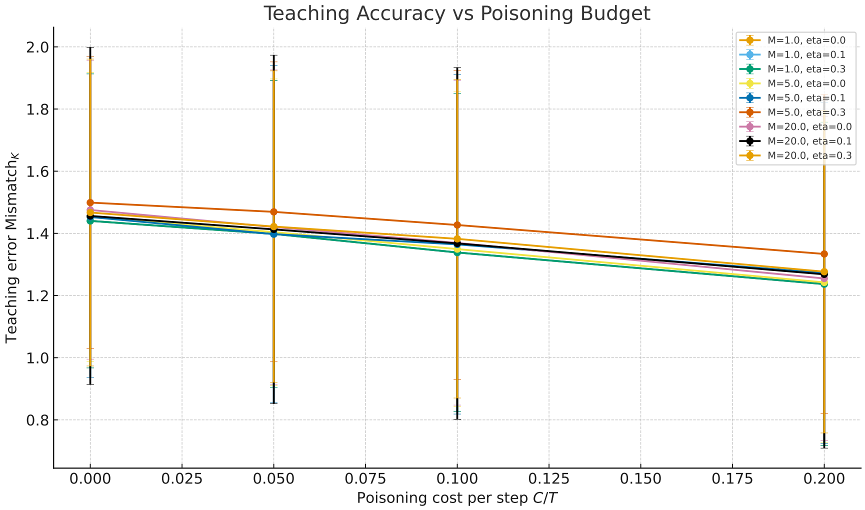

Figure 2: Teaching error Mismatch$_K$ versus normalized poisoning cost per step C/T. Larger budgets consistently reduce policy mismatch across all non-stationarity levels.

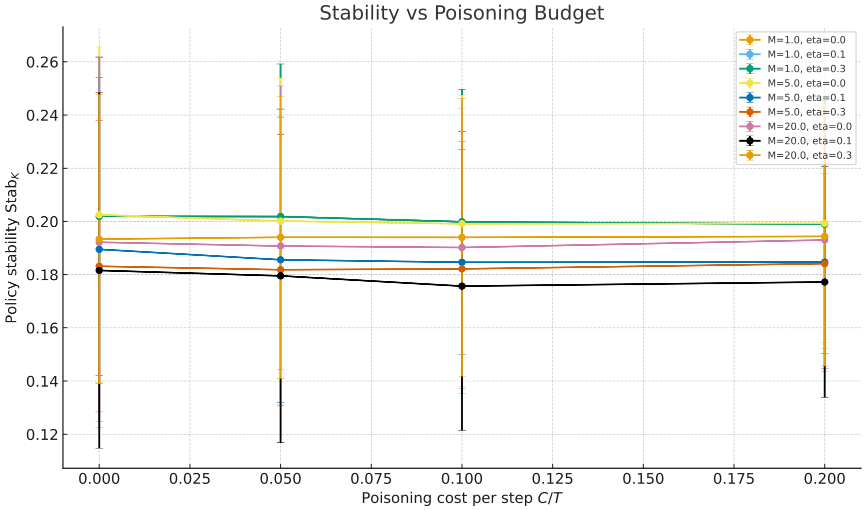

Figure 3: Policy stability Stab$_K$ versus normalized poisoning cost per step C/T. Stability remains essentially flat, indicating that teaching does not destabilize the policy updates.

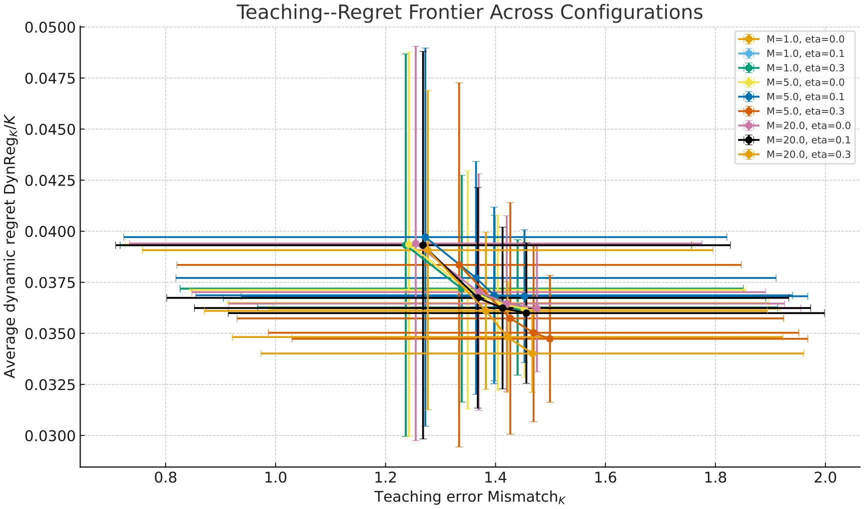

Figure 4: Empirical teaching–regret frontier: average dynamic regret DynReg_K/K versus teaching error $\mathrm{Mismatch}_K$, across budgets and non-stationarity levels. The curves illustrate that substantial reductions in mismatch can be obtained at a modest regret cost.