Generalized Principal–Agent Design in Nonstationary Environments: A Dynamic Benchmark and Stability Bounds Against Learning Agents

Liz Lemma Future Detective

January 16, 2026

Abstract

Classic principal–agent models (Stackelberg games, contract design, Bayesian persuasion) compare the principal's performance to a commitment benchmark U^* with a best-responding agent. Lin and Chen (2025) show that in stationary environments without agent private information, a no-swap-regret learning agent effectively restores this benchmark: the principal cannot exploit adaptivity beyond $U^* + O(\mathrm{SReg}(T)/T)$, while a fixed principal strategy guarantees $U^* - O(\sqrt{\mathrm{Reg}(T)/T})$.

This paper modernizes the theory for the 2026 setting where payoffs and constraints drift over time due to model updates, demand shocks, and changing regulations. We study repeated generalized principal–agent problems with time-varying utility primitives and feasibility sets. We introduce an economically interpretable dynamic benchmark—defined as the Stackelberg value of an averaged (or restricted-policy) one-shot problem—and derive sharp additive decompositions of the principal's performance into (i) strategic learning error (agent regret) and (ii) environmental drift (variation/mixing). Our main results show that no-swap-regret learning continues to immunize the agent against principal exploitation, but only up to an explicit drift penalty; conversely, a principal can secure near-benchmark performance with a fixed strategy against a no-regret agent, again up to drift. We provide specializations to dynamic Bayesian persuasion with drifting priors and to Markov environments, yielding closed-form finite-horizon rates and condition-number dependence on interiority of constraints. When exact benchmarks require computation (e.g., large state spaces), we flag the convex programs/LPs needed and provide approximation-aware statements.

Table of Contents

regret term + drift term).

2. 2. Dynamic generalized principal–agent model: primitives, time-varying constraints, information structure, and learning assumptions (contextual no-regret / no-swap-regret under drift).

3. 3. Benchmarks: (a) averaged-game Stackelberg value $\bar{U}^*$; (b) restricted-policy dynamic benchmark U_{Π}^* (stationary Markov or bounded-variation principal policies); discussion of when each is appropriate.

4. 4. Core reduction under drift: extend Lin–Chen's joint-signal argument to time-varying primitives using occupation measures; isolate the exact points where drift enters and prove generic drift-to-average inequalities.

5. 5. Main theorems (general case): upper bound vs no-swap-regret (anti-exploitation) and lower bound vs no-regret (achievability with fixed strategies), with explicit dependence on $\mathrm{diam}(X)$, G, B, L, and $\mathrm{dist}(\bar{C}, \partial X)$.

6. 6. Specialization A — Dynamic Bayesian persuasion with drifting priors: closed-form drift term via $\sum_t \|\mu_{t+1} - \mu_t\|_1$; computational notes (LP size, signal constraints).

7. 7. Specialization B — Markov environments: translate mixing assumptions into explicit drift penalties (e.g., $O(\tau_{\mathrm{mix}}/T)$); discuss when agents can learn without modeling dynamics.

8. 8. Tightness and examples: construct instances showing necessity (up to constants) of the drift penalty and of the $\sqrt{\cdot}$ vs linear regret asymmetry.

9. 9. Discussion and implications: what regulation/robust design can and cannot fix under drift; how to choose benchmark; extensions (principal learning, agent private info, attention constraints).

1 Introduction

Nonstationarity is no longer a technical nuisance that we can relegate to robustness checks; it is the operating condition of most principal–agent and persuasion settings that motivate modern applications. In 2026, platforms iterate product designs and ranking rules weekly, regulators update compliance regimes on short notice, macro conditions move demand and outside options, and learning systems retrain on data whose composition shifts endogenously with the very policies being deployed. Even when the underlying strategic interaction is stable, the mapping from decisions to realized outcomes drifts because the environment (users, suppliers, competitors, measurement pipelines) drifts. In such settings, treating the game as stationary—a single payoff matrix with a fixed feasible set—is often descriptively wrong and, more importantly for theory, can be normatively misleading about what commitment power and information design can achieve.

Classical Stackelberg and Bayesian persuasion analyses obtain sharp benchmarks by solving a one-shot commitment problem: the principal commits to a scheme, the agent best-responds, and equilibrium outcomes are characterized by concavification or linear programming arguments. These tools implicitly rely on primitives that do not change between the commitment stage and the response stage. When payoffs and constraints vary across rounds, the object "commit to a scheme" becomes ambiguous: commit to what, exactly, when the mapping from messages to outcomes evolves? One can attempt to define a dynamic mechanism with history-dependent messages and decisions, but then the benchmark ceases to be a transparent object tied to welfare or implementation and becomes a moving target that conflates adaptation with exploitation. Conversely, if we insist on a fixed benchmark (for interpretability and policy relevance), we must confront the fact that any such benchmark is necessarily approximate in a drifting world.

A second difficulty is behavioral: even if the principal is fully strategic, the agent is often not a textbook best-responder to the principal's hidden randomization. In online marketplaces, workers, advertisers, or consumers adapt using heuristics or learning algorithms with limited feedback. In organizational settings, the "agent" may be a team following a playbook that updates slowly. This means that the principal's effective leverage is mediated by how the agent learns from the realized sequence of interactions. Stationary persuasion predictions—which hinge on exact best responses to posteriors or recommendations—can fail quantitatively when the agent is instead minimizing some form of regret. Crucially, what the principal can extract depends not merely on whether the agent learns, but on *which* no-regret guarantee the agent satisfies: external-regret learning supports convergence to coarse correlated behavior, while swap-regret learning pushes the realized play toward correlated equilibria and sharply limits manipulability.

These observations motivate a simple organizing question: in a repeated

principal–agent interaction with drifting primitives, how should we decompose performance into a *strategic* component (what is enabled or disabled by the agent's learning rule) and a *statistical/drift* component (what is unavoidable because the world changes)? Our results answer this by isolating two additive terms that appear throughout: a strategic-regret term that vanishes when the agent's regret is sublinear (e.g., $\mathrm{CSReg}(T) = o(T)$ or $\mathrm{CReg}(T) = o(T)$), and a drift term that vanishes when primitives are stable (e.g., $\mathcal{V}_u \approx 0$, and analogously for other sources of change). The point is not that nonstationarity is "small" in practice, but that it is conceptually distinct from strategic sophistication, and the theory becomes clearer when we measure them separately.

At a high level, we proceed by defining a transparent benchmark that remains meaningful under drift: the Stackelberg value of a *single* one-shot problem constructed from time averages of the primitives. This benchmark captures what a principal could achieve if it had to commit to a stable policy and faced a stable response problem, with the averaging step reflecting the fact that only long-run performance is being evaluated. The main question then becomes: how far can an adaptive principal push the realized average payoff above (or below) this averaged benchmark, given that the agent learns from experience? The answer depends sharply on whether the agent satisfies a no-swap-regret condition or merely a no-regret condition. When the agent has no-swap-regret, we show an *anti-exploitation* upper bound: regardless of how informed or adaptive the principal is, its average payoff cannot exceed the averaged Stackelberg benchmark by more than an additive term of order $\mathrm{CSReg}(T)/T$ plus a drift penalty proportional to the magnitude of nonstationarity (e.g., $B\mathcal{V}_u$) and a conditioning penalty tied to time-varying feasibility.

This upper bound is economically interpretable. Swap regret controls deviations that depend on the realized recommendation and action pair, which is exactly the leverage an adaptive principal would like to use to "reshuffle" the agent's behavior across contexts in its favor. When the agent prevents such profitable reshufflings, the principal is effectively constrained to outcomes consistent with an approximate correlated equilibrium of an averaged interaction, even though the principal may be changing policies round by round. Environmental drift then enters as a separate term: even if strategic incentives are tightly regulated by the agent's learning, the realized payoffs u_t can differ from their time-average $\bar{u}$ along the realized path, and this mismatch scales with how fast the environment moves. In other words, no amount of learning stability can make a moving payoff landscape behave like a fixed one.

Complementing the upper bound, we provide a matching achievability statement under the weaker assumption of contextual no-regret. Here the conclusion is existential rather than universal: there exists a *fixed* principal strategy (chosen once, in hindsight from the averaged one-shot benchmark)

such that, if the agent has low external regret, the principal's realized average payoff is close to the averaged Stackelberg value up to a term that typically scales like $\sqrt{\mathrm{CReg}(T)/T}$ and the same drift penalties. The economic content is that when the principal restrains itself to a stable policy, the agent's external-regret learning is enough to ensure approximate best responding *in the long run*. Thus, while swap-regret is needed to immunize against a potentially manipulative principal, external regret suffices to guarantee that stable commitment policies achieve the appropriate benchmark. Taken together, the two results identify the "value of adaptivity" in repeated principal–agent problems as being bounded by (i) the agent's swap-regret slack and (ii) the world's nonstationarity.

Time-varying feasibility constraints add a subtle but practically important wrinkle. Many applications impose per-period resource, safety, or compliance constraints: budgets reset, inventories fluctuate, and acceptable action sets shift with policy. A naive averaging argument can fail near the boundary of the decision space, where small perturbations in feasible sets require large "repairs" to restore feasibility. For this reason, our guarantees include a condition-number-like factor that depends on an interiority parameter such as $\mathrm{dist}(\bar{C}, \partial X)$. Economically, this reflects a simple stability principle: organizations that operate with slack (interior feasibility) can smooth shocks cheaply, while organizations that run at the edge of capacity pay a high price for the same volatility. This dependence is not an artifact of proof technique; it captures a real fragility that should inform how one interprets theoretical benchmarks in tightly constrained systems.

We also show how the general drift term specializes to familiar stochastic-process primitives in dynamic persuasion. When payoffs are induced by an ergodic Markov state, nonstationarity over a finite horizon is governed by mixing: empirical averages converge to stationary expectations at a rate controlled by τ_{mix}. In this case, the drift penalty can be instantiated as an explicit finite-horizon term on the order of $B\tau_{\mathrm{mix}}/T$, separating transient dynamics from strategic learning effects. This specialization is important in applications where the "environment" is best modeled as a slowly mixing demand or preference state: the theory predicts that even with extremely sophisticated learning (vanishing regret), finite-horizon outcomes can systematically differ from stationary persuasion benchmarks whenever mixing is slow.

From a policy and practice perspective, the decomposition into a strategic-regret term and a drift term has two immediate implications. First, limiting manipulation is not only about restricting the principal's information or commitment power; it is also about the agent's learning guarantees. Agents who use swap-regret-minimizing procedures (or institutions that approximate them) effectively cap the gains a principal can extract from adaptive, history-dependent schemes. Second, even perfect strategic discipline does not eliminate performance gaps in a drifting world; the drift term quantifies

how much benchmark comparisons must be discounted when the system itself is changing. This is particularly relevant when one evaluates algorithmic policies against static "optimal" baselines: apparent outperformance may reflect favorable drift rather than genuine strategic advantage, and apparent underperformance may be mechanically forced by volatility.

Finally, we emphasize what our results do *not* claim. We do not assume that nonstationarity is small, only that it can be measured in a way that yields interpretable bounds; if the environment changes adversarially and rapidly, any stationary benchmark becomes less meaningful, and our drift terms correctly become large. We also abstract from several dimensions that may matter in applications: private information on the agent side, multiple agents with strategic interactions, and principals who themselves must learn the primitives rather than observe them contemporaneously. These extensions are important, but we view them as complementary: the present analysis isolates a clean tradeoff between strategic learning and environmental drift, and clarifies which aspects of stationary persuasion logic survive when the world is moving.

2 Dynamic generalized principal–agent model

We study a repeated interaction over rounds $t \in \{1, \dots, T\}$ between a principal (leader, sender, or platform) and an agent (follower, receiver, or decision maker). The distinguishing features of the model are (i) a *generalized* decision instrument for the principal—a continuous decision vector $x \in X \subset \mathbb{R}^d$ that can encode prices, allocations, ranking weights, or policy parameters—and (ii) *nonstationary primitives*, allowing payoffs and feasibility constraints to drift across rounds. Our focus is on how the principal's ability to adapt over time interacts with the agent's learning guarantees.

Decision and message spaces. The principal's decision space is a convex compact set $X \subset \mathbb{R}^d$. The agent chooses from a finite action set A (e.g., accept/reject, bid levels, effort choices). Communication is mediated by a finite signal set S with $|S| \geq |A|$. The requirement $|S| \geq |A|$ is without loss in most applications: signals can be interpreted as *recommendations* or *menus* rich enough to label each agent action, while still allowing additional "informational" messages when needed.

Round structure and strategies. Each round proceeds in the following order.

1. The agent selects a response map (policy) $\rho_t : S \to \Delta(A)$, potentially as a function of past observations. Thus, upon receiving a signal $s \in S$, the agent draws an action $a \sim \rho_t(s)$.

2. The principal selects a (possibly randomized) scheme

$$\pi_t = \{(\pi_{t,s}, x_{t,s})\}_{s \in S},$$

where $\pi_{t,s} \geq 0$, $\sum_{s \in S} \pi_{t,s} = 1$, and $x_{t,s} \in X$ is the decision implemented if signal s is realized.

3. A signal $s_t \sim \pi_t$ is realized, the principal implements $x_t := x_{t,s_t}$, and the agent draws $a_t \sim \rho_t(s_t)$.

4. Payoffs $u_t(x_t, a_t)$ and $v_t(x_t, a_t)$ are realized.

We allow the principal to be *adaptive*: π_t may depend on the full history and on contemporaneous observations of the environment. In contrast, the agent does not observe π_t directly; it only experiences the realized sequence (s_t, x_t, a_t) and whatever payoff feedback is available.

Time-varying feasibility constraints. A key modeling ingredient is that the principal faces a per-round convex feasibility constraint on its *average implemented decision* under its own randomization:

$$\sum_{s \in S} \pi_{t,s} x_{t,s} \in C_t \subseteq X, \tag{1}$$

where each C_t is convex and may vary with t. This captures resource or compliance constraints that apply in expectation over randomized deployment. For instance, $\sum_s \pi_{t,s} x_{t,s}$ can represent average spending across user segments in an advertising platform, average risk exposure in a credit setting, or average distortion in a ranking policy; C_t then captures a budget, safety, or regulatory region that can shift with external conditions.

Because the constraint is imposed on the expectation over the principal's own randomization, it naturally accommodates mixed policies (e.g., A/B tests) while preventing the principal from "hiding" infeasibility in low-probability branches. We will later aggregate these time-varying constraints through the Minkowski average set

$$\bar{C} := \left\{ \frac{1}{T} \sum_{t=1}^{T} c_t : c_t \in C_t \right\},$$

which is convex whenever each C_t is convex.

Payoffs, linearity, and boundedness. In each round t, the principal's payoff is $u_t(x, a)$ and the agent's payoff is $v_t(x, a)$. We maintain two standard regularity conditions. First, payoffs are bounded:

$$|u_t(x, a)| \leq B, \qquad |v_t(x, a)| \leq B, \qquad \forall t,\ x \in X,\ a \in A.$$

Second, payoffs are linear in x (equivalently, affine since X is compact), and in particular $u_t(\cdot, a)$ is L-Lipschitz under a chosen norm $\|\cdot\|$ on $\mathbb{R}^d$:

$$|u_t(x,a) - u_t(x',a)| \le L\|x - x'\|, \qquad \forall x, x' \in X,\ a \in A.$$

Linearity is natural when x represents a vector of weights or transfers, and it is technically convenient because it allows us to reason about expected utilities under the principal's randomization using simple averaging arguments.

Nonstationarity and drift measures. The primitives $\{(u_t, v_t, C_t)\}_{t=1}^T$ may drift arbitrarily over time. To quantify the magnitude of this drift in a way that is both interpretable and compatible with worst-case analysis, we measure total variation in payoffs by

$$\mathcal{V}_u := \frac{1}{T}\sum_{t=1}^{T-1} \sup_{x\in X, a\in A} |u_{t+1}(x,a) - u_t(x,a)|, \qquad \mathcal{V}_v := \frac{1}{T}\sum_{t=1}^{T-1} \sup_{x\in X, a\in A} |v_{t+1}(x,a) - v_t(x,a)|.$$

The interpretation is direct: $\mathcal{V}_u$ (resp. $\mathcal{V}_v$) is the average per-round worst-case change in the principal's (resp. agent's) payoff function across the joint action space. When constraints drift, we measure it via the average Hausdorff variation

$$\mathcal{V}_C := \frac{1}{T}\sum_{t=1}^{T-1} d_H(C_{t+1}, C_t),$$

where d_H denotes Hausdorff distance under the norm induced by X. These metrics are agnostic about *why* the world is changing—seasonality, policy shocks, endogenous market response—and are therefore suited to an economic objective that evaluates long-run average performance without committing to a full parametric model of dynamics.

It is also useful to keep in mind an equivalent "state" representation. One can posit an exogenous process ω_t generating (u_t, v_t, C_t); the general model makes no probabilistic assumptions on $\{\omega_t\}$, but later specializations (e.g., ergodic Markov states) provide sharper instantiations of the drift term.

Information structure. We impose an asymmetric information assumption that reflects many platform and organizational settings. The principal is "informed" in the sense that at time t it knows the current primitives u_t, v_t, and C_t (or observes the state ω_t from which they are derived). The agent may not know these objects and, crucially, does not observe the principal's full scheme π_t; it only sees the realized signal s_t (and potentially the realized x_t), and it receives feedback sufficient to run a regret-minimizing procedure.

We deliberately leave the feedback model flexible. In full-information variants, after choosing a_t the agent might observe $v_t(x_t, a)$ for all $a \in A$; in bandit variants it may observe only $v_t(x_t, a_t)$; intermediate feedback (e.g.,

partial monitoring) is also possible. The only requirement we will use is that the agent's learning rule achieves a stated regret bound relative to an appropriate comparator class, defined below.

Agent learning as contextual regret minimization. Because the agent conditions its behavior on the realized signal, the relevant notion of learning is *contextual* regret. We consider two increasingly strong guarantees.

First, the agent satisfies *contextual external no-regret* if there exists a sublinear function $\mathrm{CReg}(T) = o(T)$ such that for every deterministic mapping $d : S \to A$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\big(v_t(x_t, d(s_t)) - v_t(x_t, a_t)\big)\right] \le \mathrm{CReg}(T). \tag{2}$$

Here d is a fixed "policy" mapping signals to actions, evaluated on the realized sequence of contexts (s_t) and principal decisions (x_t). Economically, (2) says that the agent learns a near-best fixed response rule to the principal's induced signal process, even when the environment drifts.

Second, the agent satisfies *contextual no-swap-regret* if there exists $\mathrm{CSReg}(T) = o(T)$ such that for every mapping $d : S \times A \to A$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\big(v_t(x_t, d(s_t, a_t)) - v_t(x_t, a_t)\big)\right] \le \mathrm{CSReg}(T). \tag{3}$$

Swap regret is stronger because the comparator can condition not only on the signal s_t but also on the *action actually taken* a_t, thereby capturing profitable "action relabelings" within each context. This is precisely the deviation structure that characterizes correlated equilibrium, and it is the relevant notion when we ask whether an adaptive principal can manipulate the agent by shaping correlations between recommendations and realized responses.

In both definitions, expectations are taken over the internal randomization of the principal and the agent (and over any exogenous randomness generating u_t, v_t, C_t, when present). We emphasize that the bounds are required to hold *on the realized sequence* induced by the principal's (possibly adaptive) strategy; that is, the agent's guarantee is robust to the principal's endogeneity.

No dominated actions and an inducibility gap. Following Lin–Chen, we assume the agent has no weakly dominated actions. This mild regularity condition rules out degenerate cases in which the principal can create arbitrarily small payoff perturbations that swing the agent between payoff-equivalent actions in a way that is discontinuous and thus exploitable. Under no dominated actions, one can define an *inducibility gap* $G > 0$ that lower

bounds the separation between best and strictly suboptimal actions in terms of the agent's payoff; informally, G is a margin parameter ensuring that if an action is not optimal for the agent under a given decision x, then it is worse by at least G under a suitably constructed "joint-signal" representation. We treat G as an environment-dependent constant that affects the quantitative strength of the principal's ability to induce behavior, but not the qualitative decomposition into regret and drift terms.

Economically, $G > 0$ can be read as a discipline or responsiveness parameter: when the agent's incentives are sharply separated, approximate best responses (as delivered by no-regret learning) translate cleanly into predictable behavior; when incentives are nearly flat, small drift or noise can produce large behavioral variation, and any benchmark comparison becomes correspondingly less informative.

Averaging and interiority for drifting constraints. Since we evaluate the principal by its expected *average* payoff,

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} u_t(x_t, a_t)\right],$$

it is natural to compare dynamic play to a one-shot benchmark constructed from time averages of the primitives. The appropriate feasibility region for such a benchmark is the Minkowski average $\bar{C}$. However, turning per-round feasibility (1) into stable control of averaged feasibility typically requires a quantitative interiority condition. We therefore assume

$$\operatorname{dist}(\bar{C}, \partial X) > 0,$$

so that $\bar{C}$ lies a positive distance away from the boundary of X. This assumption has a clear operational meaning: the principal can implement averaged-feasible policies with some slack in X, making "repairs" for time-varying constraints stable. Without slack, even small changes in C_t can force the principal to make large compensating changes in the implemented decisions, and bounds that scale with $\mathcal{V}_C$ necessarily deteriorate.

Discussion and limitations. Two remarks clarify what this model is and is not designed to capture. First, we do not assume that the agent observes or understands the principal's objective; all strategic discipline is encapsulated in the regret inequalities (2)–(3). This is consistent with the view that many agents are algorithmic learners optimizing their own payoff stream rather than solving a full equilibrium problem. Second, we allow the principal to be fully strategic and informed; thus, any performance limitation we obtain under swap regret reflects a genuine constraint imposed by the agent's learning rule, not by informational frictions on the principal.

At the same time, we abstract from several important extensions. We do not model private information on the agent side, we consider a single agent rather than a population with externalities, and we do not require the principal to learn the primitives. These extensions can change the appropriate benchmark and, in some cases, the relevant notion of regret. Our purpose here is narrower: to isolate, in the cleanest possible setting, how nonstationarity (captured by $\mathcal{V}_u, \mathcal{V}_v, \mathcal{V}_C$) and learning guarantees (captured by $\mathrm{CReg}(T), \mathrm{CSReg}(T)$) jointly determine what a principal can and cannot achieve in repeated generalized principal–agent environments.

3 Benchmarks: what we compare dynamic play to

Our welfare statements require a reference point: a benchmark that is (i) economically interpretable, (ii) well defined under time variation in payoffs and constraints, and (iii) compatible with the agent-facing learning guarantees in (2)–(3). In nonstationary environments there is no single canonical choice. We therefore distinguish two benchmarks that serve different purposes. The first is an *averaged-game* Stackelberg value, which is deliberately "static" and is the right baseline for impossibility (anti-exploitation) results. The second is a *restricted-policy dynamic* benchmark, which is closer to an operations or algorithm-design objective when the principal itself is constrained (by policy, engineering, or regulation) to a structured family of time-varying schemes.

(a) The averaged-game Stackelberg benchmark $\bar{U}^*$. Because our objective evaluates average payoff over time, it is natural to compress the nonstationary primitives into their empirical averages,

$$\bar{u}(x,a) = \frac{1}{T}\sum_{t=1}^{T} u_t(x,a), \qquad \bar{v}(x,a) = \frac{1}{T}\sum_{t=1}^{T} v_t(x,a),$$

and to aggregate feasibility via the Minkowski average set

$$\bar{C} := \left\{ \frac{1}{T}\sum_{t=1}^{T} c_t : \; c_t \in C_t \right\}.$$

We then define $\bar{U}^*$ as the one-shot generalized Stackelberg value of the averaged instance $(\bar{u}, \bar{v}, \bar{C})$:

$$\bar{U}^* := \max_{\pi : \sum_{s\in S} \pi_s x_s \in \bar{C}} \sum_{s\in S} \pi_s \max_{a \in \arg\max_{a'' \in A} \bar{v}(x_s, a'')} \bar{u}(x_s, a). \tag{4}$$

This object should be read as follows. Imagine a counterfactual world in which the principal faces a *single* generalized principal–agent problem with

payoffs equal to time averages and a feasible region equal to the time-averaged constraint set. The principal commits to a randomized signal–decision scheme $\{(\pi_s, x_s)\}_{s\in S}$ whose average decision lies in $\bar{C}$, the agent best-responds to each realized decision x_s according to the averaged agent payoff $\bar{v}$, and the principal receives the corresponding averaged payoff $\bar{u}$.

The economic appeal of (4) is that it captures the *long-run commitment value* of persuasion or policy design when the analyst refuses to privilege any particular time period. In stationary or ergodic settings, $\bar{u}$ and $\bar{v}$ approximate the population primitives, and $\bar{C}$ approximates the long-run constraint region; in that case $\bar{U}^*$ aligns with the textbook persuasion benchmark. In adversarially drifting environments, by contrast, (4) remains well defined and places discipline on what we can guarantee uniformly over all sequences: it depends only on empirical averages and is therefore immune to "cherry-picking" favorable subperiods.

A second, more technical, reason for using $\bar{U}^*$ is that it matches the informational structure of contextual regret. The inequalities (2)–(3) compare the agent's realized play to deviations that are themselves *time-invariant mappings* (from S to A, or from $S \times A$ to A). Thus, when we translate these inequalities into restrictions on the long-run joint distribution of (s_t, x_t, a_t), the natural induced object is an *occupation measure* and its average payoffs, precisely the ingredients that define $\bar{u}, \bar{v}$, and $\bar{C}$. In this sense, $\bar{U}^*$ is the benchmark that is "dual" to the agent's learning guarantee: it is the value that would obtain if the agent were exactly best-responding to the *average* incentives that its own regret bound forces it to respect.

That said, $\bar{U}^*$ is not meant to represent what a fully informed, fully rational principal could necessarily achieve in a genuinely dynamic world. When payoffs drift, an adaptive principal can sometimes do better than any time-averaged one-shot policy by exploiting periods in which the mapping from x to payoffs is temporarily favorable. Our results will make this precise: the gap between realized performance and $\bar{U}^*$ is controlled by explicit drift terms (e.g., $B\mathcal{V}_u$ and conditioning penalties from $\mathcal{V}_C$), so $\bar{U}^*$ is best interpreted as a *stable baseline* rather than an upper envelope on dynamic possibilities.

(b) A restricted-policy dynamic benchmark U^*_Π. In many applications, the principal cannot (or should not) implement arbitrary history-dependent schemes. Product teams limit "policy churn" to preserve user experience; regulators may require stability or explainability; and operational constraints often induce a small set of admissible parameter updates. In such cases, the relevant comparator is not the full-information, fully adaptive principal, but rather the best policy inside a structured class Π.

We formalize this by fixing a nonempty family Π of principal policies,

where each policy specifies a feasible scheme each round.[1] Given such a class, we define the corresponding dynamic benchmark as

$$U^*_\Pi := \sup_{\{\pi_t\}\in\Pi} \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S}\pi_{t,s} \max_{a\in\arg\max_{a''\in A} v_t(x_{t,s},a'')} u_t(x_{t,s},a), \tag{5}$$

subject to per-round feasibility $\sum_s \pi_{t,s}x_{t,s} \in C_t$ for each t. The interpretation of (5) is a *clairvoyant* or *planning* benchmark: within class Π, what is the best average payoff the principal could obtain if, in each round, the agent played a myopic best response to the decision actually implemented in that round? This benchmark is common in mechanism and policy design because it isolates the principal's design capacity from the agent's learning dynamics.

Two examples illustrate the kinds of restrictions that lead to useful U^*_Π benchmarks.

Stationary Markov policies. In dynamic persuasion problems with an exogenous state ω_t generating (u_t, v_t, C_t), a natural restriction is that the principal may condition only on the current state and must use a stationary rule. Writing $\pi(\omega)$ for a state-contingent scheme, the class

$$\Pi_{\mathrm{SM}} := \{\ \pi_t = \pi(\omega_t) \text{ for some fixed map } \pi(\cdot)\ \}$$

yields a benchmark $U^*_{\Pi_{\mathrm{SM}}}$ that corresponds to an implementable operating policy: "given the current market/regulatory state, deploy the corresponding scheme." In an ergodic environment, $U^*_{\Pi_{\mathrm{SM}}}$ is closely related to the stationary-prior value $U^*(\mu_\infty)$, while still allowing the principal to react to state realizations rather than to time averages.

Bounded-variation (low-churn) policies. Another salient restriction is that the principal may update its scheme, but only gradually. One way to capture this is to endow the space of schemes with a metric (e.g., total variation on $\pi_{t,\cdot}$ plus a norm on the implemented decisions) and to restrict cumulative movement:

$$\Pi_{\mathrm{BV}}(V) := \left\{\ \{\pi_t\}_{t=1}^{T} :\ \sum_{t=1}^{T-1}\mathrm{dist}(\pi_{t+1},\pi_t) \le V\ \right\}.$$

This benchmark reflects organizational practice: policies can adapt, but frequent or abrupt changes are disallowed. When V is small relative to T, $U^*_{\Pi_{\mathrm{BV}}(V)}$ is a dynamic benchmark that still rules out aggressive intertemporal manipulation.

[1] Formally, an element of Π can be taken to be a measurable map from the principal's information at time t (history and possibly ω_t) to a scheme $\pi_t = \{(\pi_{t,s}, x_{t,s})\}_{s\in S}$ satisfying $\sum_s \pi_{t,s}x_{t,s} \in C_t$. We keep the definition abstract because different applications restrict different objects: the support of π_t, the allowable $x_{t,s}$, or the dynamics of how π_t may change.

When is each benchmark appropriate? The choice between $\bar{U}^*$ and U_Π^* depends on what question we ask.

If we are interested in *limits on exploitation* of a learning agent—for example, whether a platform can systematically extract more value by correlating recommendations with actions in a way the agent does not anticipate—then $\bar{U}^*$ is the right reference point. It is deliberately conservative with respect to temporal structure, and it is the benchmark that emerges from the occupation-measure perspective implied by contextual (swap) regret. Put differently, $\bar{U}^*$ is the quantity we can hope to upper bound the principal by *uniformly over all adaptive principal strategies* once the agent satisfies a strong enough equilibrium-selection property (swap regret) and the environment does not drift too violently.

If instead we are evaluating *performance of a constrained principal design or learning pipeline*, then U_Π^* is often the more meaningful comparator. Here we want to know whether a particular implementable policy class can track a moving environment and how much value is lost due to restrictions such as stationarity, Markovian dependence, or bounded policy churn. This is especially relevant for policy optimization: even if an unconstrained principal could in principle exploit nonstationarity, such exploitation may be infeasible, undesirable, or illegal; a restricted benchmark makes explicit what is being optimized.

Finally, it is important to acknowledge a limitation common to both benchmarks. Neither $\bar{U}^*$ nor U_Π^* is an "equilibrium value" of the full repeated game with forward-looking strategic behavior by both sides. We use them because our maintained behavioral assumption on the agent is *learning-theoretic* rather than equilibrium-theoretic: the agent is disciplined by regret bounds on realized play, not by common knowledge of rationality. The role of the next section is to show that, under drift, these benchmarks can still be connected to realized outcomes through a reduction that (i) converts regret inequalities into approximate optimality in an averaged one-shot problem and (ii) isolates exactly where time variation produces additive error terms.

4 Core reduction under drift: from repeated play to an averaged one-shot instance

Our main theorems rest on a single organizing idea: even though the interaction is genuinely dynamic and the principal may be adaptive, the agent-side guarantees in (2)–(3) only compare realized play to *time-invariant* deviations. This restriction is not a weakness; it is precisely what lets us "compress" the repeated game into an averaged one-shot object. The technical challenge under drift is that the compression must (i) keep track of feasibility when C_t changes over time and (ii) quantify, in a modular way, how much we lose when we replace u_t, v_t by their averages $\bar{u}, \bar{v}$. We now spell out the

reduction and isolate the two places where drift enters.

Step 1: an occupation-measure view of the repeated interaction. Fix an arbitrary (possibly adaptive and informed) principal strategy sequence $\{\pi_t\}_{t=1}^T$ and an arbitrary agent response sequence $\{\rho_t\}_{t=1}^T$. Taking expectations over the internal randomization of both players, the expected average principal payoff can be written without reference to histories as

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} u_t(x_t, a_t)\right] = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S}\sum_{a\in A}\pi_{t,s}\,\rho_t(a \mid s)\,u_t(x_{t,s}, a). \tag{6}$$

The right-hand side makes it natural to define the round-t *occupation weights*

$$m_t(s,a) := \pi_{t,s}\rho_t(a \mid s) \in [0,1], \qquad \sum_{s,a} m_t(s,a) = 1,$$

and their time average $m(s,a) := \frac{1}{T}\sum_{t=1}^{T} m_t(s,a)$. Intuitively, $m(s,a)$ is the empirical (expected) frequency with which the interaction visits signal–action pair (s,a). The virtue of this view is that all regret inequalities are linear in these same weights, so we can translate learning guarantees directly into linear constraints on an averaged distribution.

Feasibility also becomes transparent. Let $c_t := \sum_{s\in S}\pi_{t,s}x_{t,s} \in C_t$ denote the principal's round-t average decision. Then

$$\bar{c} := \frac{1}{T}\sum_{t=1}^{T} c_t \in \bar{C} \qquad \text{and} \qquad \bar{c} = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S}\pi_{t,s}x_{t,s}. \tag{7}$$

Thus, no matter how violently C_t moves within X, averaging across rounds produces an exactly feasible point for the Minkowski average constraint set $\bar{C}$. The conditioning issues associated with $\mathrm{dist}(\bar{C}, \partial X)$ do not arise here; they arise later, when we ask whether a *fixed* averaged feasible point can be implemented *round-by-round* under the varying sets $\{C_t\}$.

Step 2: the joint-signal lift (Lin–Chen) and why it still works with drift. In a static persuasion/Stackelberg problem, we typically require that the agent best-responds *signal-by-signal*. In a repeated setting with contextual *swap* regret, however, the agent is disciplined against deviations that can depend on both the observed signal and the realized action, i.e., maps $d : S \times A \to A$ as in (3). Lin–Chen's key trick is to build an auxiliary one-shot instance whose "signals" are precisely these joint realizations.

Formally, define an augmented signal set $\widetilde{S} := S \times A$, with elements $\tilde{s} = (s,a)$. We now construct a *time-averaged joint-signal scheme* $\widetilde{\pi} = \{(\widetilde{\pi}_{\tilde{s}}, \widetilde{x}_{\tilde{s}})\}_{\tilde{s}\in\widetilde{S}}$ as follows:

$$\widetilde{\pi}_{s,a} := m(s,a) = \frac{1}{T}\sum_{t=1}^{T}\pi_{t,s}\rho_t(a \mid s),$$

and, whenever $\widetilde{\pi}_{s,a} > 0$,

$$\widetilde{x}_{s,a} := \frac{1}{T\,\widetilde{\pi}_{s,a}} \sum_{t=1}^{T} \pi_{t,s}\rho_t(a \mid s)\, x_{t,s} \ \in X, \tag{8}$$

with an arbitrary $\widetilde{x}_{s,a} \in X$ if $\widetilde{\pi}_{s,a} = 0$. The inclusion $\widetilde{x}_{s,a} \in X$ uses only convexity of X.

This construction preserves feasibility *exactly* at the averaged level:

$$\sum_{s\in S}\sum_{a\in A} \widetilde{\pi}_{s,a}\,\widetilde{x}_{s,a} = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S} \pi_{t,s}x_{t,s} = \bar{c} \in \bar{C}. \tag{9}$$

It also preserves expected payoffs whenever we evaluate them under *time-averaged* primitives. Because $\bar{u}(\cdot, a)$ and $\bar{v}(\cdot, a)$ are linear in x, we have the exact identities

$$\sum_{s,a} \widetilde{\pi}_{s,a}\,\bar{u}(\widetilde{x}_{s,a}, a) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s,a} \pi_{t,s}\rho_t(a \mid s)\,\bar{u}(x_{t,s}, a), \tag{10}$$

and the analogous equality with $\bar{v}$ in place of $\bar{u}$. In words: once we replace u_t, v_t by their empirical averages, the joint-signal lift converts the dynamic, time-varying scheme into a one-shot randomized signal–decision plan over $\widetilde{S}$ with no loss.

Step 3: translating swap regret into (approximate) obedience constraints for $\bar{v}$. The contextual swap-regret guarantee (3) can be expanded, taking expectation and conditioning on (t, s, a), as

$$\forall d : S\times A \to A, \qquad \sum_{t=1}^{T}\sum_{s,a} \pi_{t,s}\rho_t(a \mid s)\Big(v_t(x_{t,s}, d(s,a)) - v_t(x_{t,s}, a)\Big) \le \mathrm{CSReg}(T). \tag{11}$$

At this point the only obstacle to a clean one-shot statement is that the payoff inside the sum is v_t, not $\bar{v}$. The next lemma shows that the difference is controlled by the drift measure $\mathcal{V}_v$ and, importantly, does *not* depend on how the principal correlates $(x_{t,s})$ with time.

Lemma 4.1 (Drift-to-average for time-varying payoffs)**.** *Let $\{f_t\}_{t=1}^{T}$ be bounded real-valued functions on a common domain $\mathcal{Z}$, and let $\bar{f} := \frac{1}{T}\sum_{t=1}^{T} f_t$. For any (possibly adaptive) sequence $z_1, \ldots, z_T \in \mathcal{Z}$,*

$$\frac{1}{T}\sum_{t=1}^{T} \big|f_t(z_t) - \bar{f}(z_t)\big| \le \frac{2}{T}\sum_{t=1}^{T-1} \sup_{z\in\mathcal{Z}} \big|f_{t+1}(z) - f_t(z)\big|.$$

Applying Lemma 4.1 with $\mathcal{Z} = X \times A$ and $f_t(\cdot,\cdot) = v_t(\cdot,\cdot)$ (and similarly for the deviation payoff $v_t(\cdot, d(\cdot))$) yields, after dividing by T, the averaged incentive constraints

$$\forall d : S\times A \to A, \qquad \sum_{s,a} \widetilde{\pi}_{s,a}\Big(\bar{v}(\widetilde{x}_{s,a}, d(s,a)) - \bar{v}(\widetilde{x}_{s,a}, a)\Big) \;\le\; O\bigg(\frac{\mathrm{CSReg}(T)}{T}\bigg) + O(\mathcal{V}_v). \tag{12}$$

We read (12) as an *approximate obedience* condition in the auxiliary one-shot problem on signals $\widetilde{S}$: if the principal were to publicly reveal (s,a) and implement decision $\widetilde{x}_{s,a}$, then the agent would have negligible incentive (under $\bar{v}$) to replace action a by any alternative rule $d(s,a)$.

This is exactly where the assumption of no weakly dominated actions enters via the inducibility gap $G > 0$. In the static Lin–Chen argument, G is used to upgrade an approximate obedience inequality into a statement that, on most of the probability mass, the realized action must be close to a true best response (and hence the principal cannot extract unbounded gains by inducing "nearly dominated" behavior). Under drift, the same logic goes through with an additive slack given by the right-hand side of (12).

Step 4: translating principal payoffs back and forth between u_t and $\bar{u}$. Once we have reduced the agent-side behavior to approximate obedience for the averaged utility $\bar{v}$, we would like to compare the principal's realized payoff (6) to the payoff of the induced averaged one-shot scheme, namely $\sum_{s,a} \widetilde{\pi}_{s,a} \bar{u}(\widetilde{x}_{s,a}, a)$. The only difference is again drift. Applying Lemma 4.1 to u_t yields

$$\left| \frac{1}{T}\sum_{t=1}^{T} \sum_{s,a} \pi_{t,s}\rho_t(a \mid s)\, u_t(x_{t,s}, a) - \frac{1}{T}\sum_{t=1}^{T} \sum_{s,a} \pi_{t,s}\rho_t(a \mid s)\, \bar{u}(x_{t,s}, a) \right| \le O(\mathcal{V}_u). \tag{13}$$

Combining (13) with the exact identity (10) shows that, up to an additive $O(\mathcal{V}_u)$ term, the principal's dynamic payoff is the payoff of the averaged joint-signal one-shot scheme. This is the second and final place where drift in payoffs matters: it enters only through the generic replacement of u_t by $\bar{u}$ and v_t by $\bar{v}$.

Step 5: where drifting constraints enter (and why conditioning appears). The occupation-measure and joint-signal steps treat feasibility at the *averaged* level, producing (9) at no cost. However, our lower-bound (achievability) statement will require a converse maneuver: starting from a *fixed* scheme that is feasible for $\bar{C}$, we must implement a feasible scheme in every round t under the moving set C_t. This is where the geometry of X and the interiority parameter $\mathrm{dist}(\bar{C}, \partial X)$ become essential.

At a high level, we use a "repair" argument: given a target averaged decision $\bar{c} \in \bar{C}$, we select per-round decisions $c_t \in C_t$ whose average is $\bar{c}$ and

whose deviations $\|c_t - \bar{c}\|$ are controlled by the path variation of $\{C_t\}$ measured in Hausdorff distance. When $\bar{C}$ lies strictly inside X, such a selection is stable; near the boundary, small set movements can force large corrections, and the resulting loss is amplified by a condition-number-like factor proportional to $\mathrm{diam}(X)/\mathrm{dist}(\bar{C}, \partial X)$. Because $u_t(\cdot, a)$ is L-Lipschitz, any such correction translates linearly into payoff loss, explaining the form of the constraint-drift term that appears in the main bounds.

Summary of the reduction. Putting the pieces together, the reduction produces the following conceptual pipeline.

1. We represent the repeated interaction by an averaged occupation measure over (s, a) and an averaged feasible decision $\bar{c} \in \bar{C}$.

2. Via the joint-signal lift, we convert this averaged object into a bona fide one-shot scheme over $\widetilde{S} = S \times A$ with decisions $\{\widetilde{x}_{s,a}\}$.

3. Contextual swap regret yields approximate obedience constraints for the averaged agent payoff $\bar{v}$, with an additive slack $O(\mathrm{CSReg}(T)/T) + O(\mathcal{V}_v)$.

4. Drift-to-average inequalities control the error incurred when we replace u_t, v_t by $\bar{u}, \bar{v}$, contributing $O(\mathcal{V}_u)$ and $O(\mathcal{V}_v)$ terms.

5. Constraint drift matters only when we move from averaged feasibility in $\bar{C}$ back to per-round feasibility in C_t, introducing the conditioning dependence on $\mathrm{dist}(\bar{C}, \partial X)$ and the variation term $\mathcal{V}_C$.

This modularity is valuable beyond our specific application. In practice, platform policies and regulatory constraints often evolve slowly, while payoffs may be subject to seasonalities or market shocks. The reduction cleanly separates (i) what learning rules prevent a principal from exploiting (the incentive constraints driven by regret) from (ii) what nonstationarity inevitably obscures (the drift terms) and (iii) what geometry makes fragile (constraint repair near the boundary). The next section instantiates this pipeline into explicit upper and lower bounds, tracking the dependence on $B, L, G, \mathrm{diam}(X)$, and $\mathrm{dist}(\bar{C}, \partial X)$.

5 Main theorems (general case): anti-exploitation and achievability with explicit constants

We now instantiate the reduction in Section 4 into two quantitative statements that together pin down the value of adaptivity in the repeated interaction. The first is an *anti-exploitation* upper bound: if the agent attains contextual *swap* regret, then no principal—even one who is fully informed

about (u_t, v_t, C_t) and who adapts arbitrarily over time—can exceed the averaged one-shot Stackelberg benchmark $\bar{U}^*$ by more than terms that are proportional to (i) the agent's swap-regret rate and (ii) the amount of drift. The second is an *achievability* lower bound: if the agent attains contextual (external) no-regret, then the principal can guarantee nearly $\bar{U}^*$ using a *fixed* strategy, again up to regret and drift penalties.

Throughout we work with the norm used to define Lipschitzness and Hausdorff distance. We summarize the geometry of the feasible region by the diameter

$$\operatorname{diam}(X) := \sup_{x,x' \in X} \|x - x'\|$$

and the (dimensionless) conditioning parameter

$$\kappa(X, \bar{C}) := \frac{\operatorname{diam}(X)}{\operatorname{dist}(\bar{C}, \partial X)}. \tag{14}$$

Interiority of $\bar{C}$ ensures $\kappa(X, \bar{C}) < \infty$. Economically, $\kappa(X, \bar{C})$ measures how sensitive feasibility is to perturbations: when $\bar{C}$ approaches the boundary of X, small changes in constraints can force large changes in implementable decisions, amplifying welfare losses by Lipschitzness of payoffs.

5.1 From approximate obedience to a quantitative cap (the role of G)

The reduction produces an auxiliary one-shot scheme on joint signals $\widetilde{S} = S \times A$ satisfying approximate obedience inequalities for the averaged agent utility $\bar{v}$ (cf. (12)). To turn such inequalities into a bound on the principal's value, we need a stability property of best responses. This is exactly what the inducibility gap $G > 0$ (implied by the absence of weakly dominated actions, as in Lin–Chen) provides: it prevents the principal from extracting large gains by inducing behavior that is only *nearly* optimal for the agent.

We record the implication we use as a black box; its proof is standard in this literature and follows Lin–Chen's argument, combined with a simple averaging step.

Lemma 5.1 (Gap-based stability of obedience)**.** *Fix the averaged primitives $(\bar{u}, \bar{v}, \bar{C})$ and suppose the inducibility gap is $G > 0$. Consider any one-shot scheme $\{(\eta_\sigma, y_\sigma)\}_{\sigma \in \Sigma}$ with $y_\sigma \in X$ and average feasibility $\sum_\sigma \eta_\sigma y_\sigma \in \bar{C}$. Suppose that for some $\varepsilon \geq 0$ the agent satisfies ε-approximate obedience:*

$$\forall d : \Sigma \to A, \qquad \sum_{\sigma \in \Sigma} \eta_\sigma \big(\bar{v}(y_\sigma, d(\sigma)) - \bar{v}(y_\sigma, a_\sigma)\big) \leq \varepsilon,$$

where a_σ denotes the action taken under signal σ (or, equivalently, the recommended action in a direct scheme). Then the principal payoff under $\bar{u}$ is

bounded by

$$\sum_{\sigma\in\Sigma} \eta_\sigma \bar{u}(y_\sigma, a_\sigma) \le \bar{U}^* + \frac{2B}{G}\,\varepsilon. \tag{15}$$

Two comments clarify what Lemma 5.1 is doing. First, the dependence on B/G is inevitable: if the agent can be made indifferent among actions up to ε, then the principal might be able to shift probability mass to actions that differ in principal payoff by as much as $2B$, and the gap G controls how much mass can be moved without violating obedience. Second, the lemma is compatible with our signal-size assumptions: while our reduction naturally produces joint signals $\widetilde{S} = S \times A$, the one-shot benchmark $\bar{U}^*$ is computed over the original class of schemes with signal set S (and $|S| \ge |A|$). Under linearity in x, standard "revelation" and signal-compression arguments imply that allowing additional dummy signals does not increase the Stackelberg value relative to $\bar{U}^*$, so (15) indeed compares to the intended benchmark.

5.2 Upper bound: anti-exploitation under contextual swap regret

We can now state the anti-exploitation theorem. The proof is modular: (i) apply the reduction to obtain an ε-obedient one-shot scheme for $\bar{v}$, where ε is controlled by $\mathrm{CSReg}(T)/T$ and the drift $\mathcal{V}_v$ (Lemma 4.1); (ii) invoke Lemma 5.1 to compare the induced principal payoff under $\bar{u}$ to $\bar{U}^*$; (iii) translate $\bar{u}$ back to realized u_t using the drift control for $\mathcal{V}_u$; and (iv) incorporate constraint drift through a feasibility-repair bound (stated below) when we insist on comparing to a benchmark defined with the averaged feasible set $\bar{C}$ while the principal must satisfy the moving constraints C_t.

Theorem 5.2 (Anti-exploitation cap under drift)**.** *Suppose $|u_t(x,a)|, |v_t(x,a)| \le B$ for all t, x, a, and the agent satisfies contextual swap regret with bound $\mathrm{CSReg}(T)$. Assume no weakly dominated actions and inducibility gap $G > 0$. Then for any (possibly adaptive and informed) principal strategy sequence $\{\pi_t\}_{t=1}^T$ satisfying $\sum_s \pi_{t,s} x_{t,s} \in C_t$ in each round,*

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} u_t(x_t, a_t)\right] \le \bar{U}^* + \frac{4B}{G}\left(\frac{\mathrm{CSReg}(T)}{T} + 2\mathcal{V}_v\right) + 2B\,\mathcal{V}_u + 4L\,\kappa(X, \bar{C})\,\mathcal{V}_C. \tag{16}$$

In particular, if payoffs are stationary ($\mathcal{V}_u = \mathcal{V}_v = 0$) and constraints do not drift ($\mathcal{V}_C = 0$), then

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} u_t(x_t, a_t)\right] \le \bar{U}^* + \frac{4B}{G}\cdot\frac{\mathrm{CSReg}(T)}{T}.$$

Theorem 5.2 formalizes the sense in which swap regret prevents dynamic exploitation. If the agent's learning dynamics eliminate profitable *signal–action contingent* deviations, then the principal's additional degrees of freedom—observing the environment each round, randomizing over signals, correlating decisions with time—cannot create value beyond what could have been attained by committing to the best scheme in the averaged environment, except to the extent that (a) the agent's swap regret is non-negligible and (b) the environment itself is moving. The $B\,\mathcal{V}_u$ and $B\,\mathcal{V}_v$ terms are unavoidable: even a benevolent principal cannot predict the realized payoff of a fixed averaged scheme when primitives drift, and a strategic principal can at most take advantage of such drift linearly in its magnitude. The constraint term highlights a different limitation: even if the principal's objective and the agent's incentives were perfectly stable, time-varying feasibility may prevent the principal from implementing, round-by-round, what would be optimal in the averaged feasible set; the blow-up factor $\kappa(X, \bar{C})$ is precisely the fragility near the boundary.

5.3 A repair bound for drifting constraints (why $\mathrm{dist}(\bar{C}, \partial X)$ matters)

The remaining ingredient needed for a clean lower bound is a constructive statement: starting from a *fixed* scheme whose average decision lies in $\bar{C}$, we must implement a per-round feasible scheme under C_t while staying close in X. This is where $\mathrm{dist}(\bar{C}, \partial X)$ and $\mathrm{diam}(X)$ enter.

Lemma 5.3 (Constraint repair with controlled displacement)**.** *Fix $\bar{c} \in \bar{C}$ and suppose $\mathrm{dist}(\bar{C}, \partial X) > 0$. Then there exists a sequence $\{c_t\}_{t=1}^T$ with $c_t \in C_t$ for all t and $\frac{1}{T}\sum_{t=1}^T c_t = \bar{c}$ such that*

$$\frac{1}{T}\sum_{t=1}^{T} \|c_t - \bar{c}\| \le 2\,\kappa(X, \bar{C})\,\mathcal{V}_C. \tag{17}$$

Consequently, for any fixed weights $\{\pi_s\}_{s\in S}$ and decisions $\{x_s\}_{s\in S} \subseteq X$ with $\sum_s \pi_s x_s = \bar{c}$, the translated decisions

$$x_{t,s} := x_s + (c_t - \bar{c})$$

remain in X and satisfy $\sum_s \pi_s x_{t,s} = c_t \in C_t$, while inducing an average Lipschitz payoff loss at most

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S} \pi_s \big|u_t(x_{t,s}, a) - u_t(x_s, a)\big| \le 2L\,\kappa(X, \bar{C})\,\mathcal{V}_C \qquad (\forall a \in A). \tag{18}$$

Lemma 5.3 is the precise form of the intuition in Step 5 of the reduction: when $\bar{C}$ is well inside X, we can "absorb" moderate movements of C_t by

small translations that preserve feasibility; near the boundary, the same movement forces larger corrections, and the resulting loss is amplified linearly by $\kappa(X, \bar{C})$.

5.4 Lower bound: achievability via a fixed strategy under contextual no-regret

We turn to the robust achievability guarantee. The natural candidate strategy is to pick an optimizer of the averaged one-shot problem, and then hold it fixed across time. With fixed play by the principal, the agent's contextual no-regret (external regret) ensures that the agent cannot consistently gain by switching, signal-by-signal, to a better fixed action. Combining this with the same gap-based stability and the drift-to-average bounds yields a guarantee relative to $\bar{U}^*$.

Theorem 5.4 (Achievability with a fixed strategy)**.** *Suppose $|u_t(x,a)|, |v_t(x,a)| \le B$ and $u_t(\cdot, a)$ is L-Lipschitz for each a. Assume no weakly dominated actions with inducibility gap $G > 0$. If the agent satisfies contextual no-regret with bound $\mathrm{CReg}(T)$, then there exists a* fixed *principal strategy $\pi = \{(\pi_s, x_s)\}_{s \in S}$ (chosen as a function of $(\bar{u}, \bar{v}, \bar{C})$) and a per-round feasible implementation $\{(\pi_s, x_{t,s})\}_{t=1}^T$ with $\sum_s \pi_s x_{t,s} \in C_t$ for all t such that*

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} u_t(x_t, a_t)\right] \ge \bar{U}^* - \frac{4B}{G}\sqrt{\frac{\mathrm{CReg}(T)}{T}} - 2B\,\mathcal{V}_u - 4L\,\kappa(X, \bar{C})\,\mathcal{V}_C. \tag{19}$$

In particular, if (u_t, v_t, C_t) are stationary, the principal can guarantee $\bar{U}^* - O\Big(\frac{B}{G}\sqrt{\mathrm{CReg}(T)/T}\Big)$ with a fixed strategy.

Theorem 5.4 has a clear practical interpretation. If the agent's learning rule is only disciplined against *external* deviations, then the principal does not need adaptivity to achieve the long-run benchmark: a simple commitment to a fixed policy computed from the time-averaged environment is enough. The bound also makes transparent when this prescription is fragile. First, if incentives drift ($\mathcal{V}_u > 0$), then an averaged policy is necessarily a misspecified predictor of current payoffs; this is a statistical limitation rather than a strategic one. Second, if constraints drift ($\mathcal{V}_C > 0$), then feasibility requires round-by-round adjustments; the associated welfare loss is governed by the geometry of X through $\kappa(X, \bar{C})$ and by the Lipschitz constant L. Third, the inducibility gap G again governs how quickly approximate incentive alignment translates into realized behavior that is close to a best response favorable to the principal.

Putting the two theorems together. Taken jointly, Theorems 5.2 and 5.4 identify $\bar{U}^*$ as the correct limiting benchmark for repeated principal–agent

interaction under standard learning dynamics, with a sharp separation of roles:

- swap regret governs what an informed, adaptive principal *cannot* do (anti-exploitation);
- external regret governs what a non-adaptive principal *can* do (achievability);
- drift in payoffs and constraints governs what neither side can eliminate, and the conditioning parameter $\kappa(X, \bar{C})$ tells us when constraint drift is economically innocuous versus when it is amplified by tight feasibility.

The next section specializes these general statements to dynamic Bayesian persuasion with drifting priors, where the drift terms admit closed-form expressions in total variation and where the benchmark $\bar{U}^*$ can be computed by a linear program whose size is explicit in $|A|, |S|$, and the state dimension.

6 Specialization A: dynamic Bayesian persuasion with drifting priors

We now specialize the abstract principal–agent model to a canonical *dynamic Bayesian persuasion* environment in which the only source of nonstationarity is a slowly drifting public prior. This specialization is useful for two reasons. First, it turns the abstract drift quantities into an explicit, interpretable expression—a total-variation path length of the prior. Second, it clarifies what our theorems imply for practice: when the prior changes gradually, a principal (sender/platform) can compute a *single* persuasion policy from the time-averaged prior and use it throughout, incurring a loss that is linear in the amount of prior movement.

Model. Let the state space be finite, $\Omega = \{1, \dots, n\}$. In round t, a public prior $\mu_t \in \Delta(\Omega)$ is realized. The principal observes μ_t and can commit (within the round) to an information structure that induces a distribution over posteriors. The agent observes the realized signal and chooses an action $a \in A$ to maximize expected utility given the induced posterior.

We map this into our notation by taking

$$X = \Delta(\Omega) \subset \mathbb{R}^n, \qquad x \in X \text{ is a posterior belief.}$$

State-dependent utilities are time-invariant functions $u(\omega, a)$ and $v(\omega, a)$, with $|u(\omega, a)|, |v(\omega, a)| \le B$ for all (ω, a). The induced (belief-based) payoffs are linear in the posterior:

$$u_t(x, a) := \sum_{\omega \in \Omega} x(\omega)\, u(\omega, a), \qquad v_t(x, a) := \sum_{\omega \in \Omega} x(\omega)\, v(\omega, a).$$

Thus all time dependence enters through feasibility, not through payoffs: we have $u_t \equiv u$ and $v_t \equiv v$ as functions of (x, a).

The persuasion feasibility constraint is Bayes plausibility: the average posterior must equal the prior. In our language,

$$C_t = \{\mu_t\} \subseteq X, \qquad \sum_{s\in S} \pi_{t,s} x_{t,s} \in C_t \quad \Longleftrightarrow \quad \sum_{s\in S} \pi_{t,s} x_{t,s} = \mu_t.$$

This is exactly the standard reduced form in Bayesian persuasion: a scheme is a distribution over posteriors $\{(\pi_{t,s}, x_{t,s})\}_{s\in S}$ that averages to μ_t.

Averaging and the benchmark. Because each C_t is a singleton, the Minkowski average constraint set collapses to a singleton as well:

$$\bar{C} = \left\{ \frac{1}{T} \sum_{t=1}^{T} \mu_t \right\} = \{\bar{\mu}\}, \qquad \bar{\mu} := \frac{1}{T} \sum_{t=1}^{T} \mu_t.$$

Consequently, the benchmark $\bar{U}^*$ becomes the ordinary one-shot persuasion value under the averaged prior $\bar{\mu}$ and the fixed utilities (u, v). In particular, in this specialization there is no ambiguity about what we are comparing to: $\bar{U}^*$ is the sender-optimal commitment payoff for the *time-averaged* persuasion instance.

Closed-form drift term from the prior path length. We next make the drift penalties in Theorems 5.2–5.4 explicit. Since $u_t \equiv u$ and $v_t \equiv v$, we have

$$\mathcal{V}_u = 0, \qquad \mathcal{V}_v = 0.$$

All non-stationarity is in the constraint sets $\{C_t\}$, and because C_t is a singleton, Hausdorff distance is simply the ambient norm distance:

$$d_H(C_{t+1}, C_t) = \|\mu_{t+1} - \mu_t\|.$$

If we adopt the ℓ_1 norm on $X = \Delta(\Omega)$ (natural for beliefs and total variation), then

$$\mathcal{V}_C = \frac{1}{T} \sum_{t=1}^{T-1} \|\mu_{t+1} - \mu_t\|_1. \tag{20}$$

This is the promised closed-form term: the average total-variation *path length* of the prior process.

The geometric conditioning term $\kappa(X, \bar{C})$ is also interpretable on the simplex. Under ℓ_1, the diameter of the simplex is $\mathrm{diam}(X) = 2$, and the distance from $\bar{\mu}$ to the boundary is the smallest coordinate:

$$\mathrm{dist}(\bar{C}, \partial X) = \mathrm{dist}(\bar{\mu}, \partial\Delta(\Omega)) = \min_{\omega\in\Omega} \bar{\mu}(\omega).$$

Hence

$$\kappa(X, \bar{C}) = \frac{2}{\min_\omega \bar{\mu}(\omega)}. \tag{21}$$

This clarifies when constraint drift is benign versus amplified: if the averaged prior assigns very small probability to some state, then $\bar{\mu}$ lies near the boundary and feasibility becomes ill-conditioned. Economically, rare states create fragility because Bayes plausibility forces the distribution of posteriors to "balance" these small masses; small changes in μ_t can then require large shifts in posteriors.

Finally, the Lipschitz constant L for $u(\cdot, a)$ as a function of x is immediate under ℓ_1:

$$|u(x, a) - u(x', a)| = \left| \sum_\omega (x(\omega) - x'(\omega)) u(\omega, a) \right| \leq B \, \|x - x'\|_1,$$

so we may take $L = B$.

Implications of the general bounds. Plugging (20)–(21) and $L = B$ into the main theorems yields particularly transparent corollaries.

First, under contextual swap regret, an informed and fully adaptive sender cannot beat the averaged-prior persuasion value by more than swap-regret plus a prior-drift penalty:

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} u(x_t, a_t)\right] \leq \bar{U}^* + \frac{4B}{G} \cdot \frac{\mathrm{CSReg}(T)}{T} + \frac{16B}{\min_\omega \bar{\mu}(\omega)} \cdot \left(\frac{1}{T} \sum_{t=1}^{T-1} \|\mu_{t+1} - \mu_t\|_1 \right). \tag{22}$$

The first additive term is the familiar "approximate obedience" slack scaled by B/G. The second is the cost of implementing, round by round, a scheme that is only guaranteed Bayes-plausible with respect to the *average* prior: the larger the movement in priors, the more translation is needed to restore feasibility, and the more welfare can be lost (or gained) through Lipschitzness.

Second, under contextual external regret, the sender can achieve (up to the same prior-drift penalty) the averaged-prior value with a *fixed* scheme computed from $\bar{\mu}$:

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} u(x_t, a_t)\right] \geq \bar{U}^* - \frac{4B}{G} \sqrt{\frac{\mathrm{CReg}(T)}{T}} - \frac{16B}{\min_\omega \bar{\mu}(\omega)} \cdot \left(\frac{1}{T} \sum_{t=1}^{T-1} \|\mu_{t+1} - \mu_t\|_1 \right). \tag{23}$$

The operational message is simple: when the receiver is disciplined only against signal-by-signal fixed deviations (external regret), the sender does not need to track the drifting prior in order to secure near-optimal long-run performance. Tracking may still help, but any improvement over $\bar{U}^*$ is limited by the total variation budget of the prior path and the conditioning of the averaged prior.

Interpretation and limitations. Equations (22)–(23) connect three economically meaningful objects: (i) how quickly the receiver learns to respond to recommendations (regret rates), (ii) how much the belief environment changes (total variation path length), and (iii) how close the average prior is to having full support.

Two caveats are worth emphasizing. First, the factor $1/\min_\omega \bar{\mu}(\omega)$ is not an artifact of the analysis; it reflects a real instability near the boundary of the belief simplex. If $\bar{\mu}$ assigns vanishing mass to some state, then Bayes plausibility severely restricts which posteriors can be mixed, and small movements in μ_t can force large changes in feasible posterior distributions. Second, our specialization has held (u, v) fixed. If payoffs also drift (for example, the receiver's objective changes due to policy, or the sender's payoff from actions changes due to market conditions), then $\mathcal{V}_u$ and $\mathcal{V}_v$ re-enter exactly as in the general theorems; the prior-path term (20) is then only one component of drift.

Computational notes: LP formulation and signal size. In this persuasion setting, the averaged benchmark $\bar{U}^*$ is computationally tractable. Because $\bar{C} = \{\bar{\mu}\}$, computing $\bar{U}^*$ is equivalent to solving the one-shot persuasion problem under prior $\bar{\mu}$.

A standard direct-recommendation (or "action advice") formulation is convenient and aligns with our standing assumption $|S| \geq |A|$. We may identify signals with recommended actions, i.e., take $S = A$ without loss for the value. Introduce variables $\phi(\omega, a) \geq 0$ representing the joint distribution of state and recommended action:

$$\phi(\omega, a) = \bar{\mu}(\omega)\, \sigma(a \mid \omega),$$

where σ is the signaling rule. The Bayes plausibility constraint becomes linear:

$$\sum_{a \in A} \phi(\omega, a) = \bar{\mu}(\omega) \qquad (\forall \omega \in \Omega).$$

Obedience (receiver incentive) constraints are also linear: for each recommended action a and deviation $a' \in A$,

$$\sum_{\omega \in \Omega} \phi(\omega, a)\, v(\omega, a) \geq \sum_{\omega \in \Omega} \phi(\omega, a)\, v(\omega, a').$$

The objective is linear in ϕ:

$$\max_{\phi \geq 0} \sum_{\omega \in \Omega} \sum_{a \in A} \phi(\omega, a)\, u(\omega, a).$$

This is a linear program with $n|A|$ variables, n Bayes constraints, and $|A|(|A|-1)$ obedience constraints (plus nonnegativity). Its size depends polynomially

on $|\Omega|$ and $|A|$, and it can be solved once to obtain the fixed strategy used in (23). If instead the sender insisted on re-optimizing every round using μ_t, it would solve the same LP with $\bar{\mu}$ replaced by μ_t each time; our bounds quantify when such repeated re-optimization can only marginally improve long-run performance.

Finally, note how the signal-size issue fits into this computation. Although a persuasion scheme can, in principle, use many signals (and our general model allows $|S| \geq |A|$), the direct formulation above shows that for value comparisons it suffices to use at most $|A|$ signals: one per recommended action. When $|S| > |A|$, the extra signals can be treated as redundant labels that do not change the feasible set of implementable joint distributions over (ω, a). Thus, in this specialization, both the benchmark $\bar{U}^*$ and the fixed strategy achieving (23) admit an explicit, efficiently computable representation.

7 Specialization B: Markov environments and mixing-based drift bounds

The prior-drift persuasion specialization above illustrates a particularly transparent (and deterministic) source of non-stationarity: Bayes-plausibility constraints that move along a bounded-variation path. In many economic and computational settings, however, the environment is neither adversarial nor smoothly drifting; instead it evolves according to a *stochastic* law with temporal dependence. A canonical example is an ergodic Markov process capturing demand regimes, political states, or macro conditions. In this section we show how standard mixing assumptions translate directly into an explicit "drift penalty" of order $O(\tau_{\text{mix}}/T)$, and we discuss why the agent can still be modeled as a regret-minimizer *without* explicitly learning the transition kernel.

Model: a Markov state process generating primitives. Let ω_t be a time-homogeneous Markov chain on a finite state space Ω with transition matrix P and unique stationary distribution μ_∞. In round t, the realized state ω_t generates the principal's and agent's stage primitives:

$$u_t(x, a) = u(\omega_t, x, a), \qquad v_t(x, a) = v(\omega_t, x, a), \qquad C_t = C(\omega_t),$$

where $u(\omega, \cdot, a)$ and $v(\omega, \cdot, a)$ are linear in $x \in X$ and uniformly bounded by B in absolute value, and $C(\omega) \subseteq X$ is convex for each ω. We allow the principal to observe ω_t (or equivalently to know (u_t, v_t, C_t) at time t), while the agent need not. The within-round timing is as in the global model: the agent chooses a response map ρ_t, the principal chooses a feasible signaling/decision scheme π_t, then (s_t, x_t, a_t) are realized and feedback is observed by the agent.

Two clarifications are useful. First, the Markov assumption is *not* an informational restriction on the principal: the principal may still use an arbitrary history-dependent strategy. Second, the agent's learning rule is still summarized solely by regret bounds (external or swap), which are defined pathwise and therefore remain valid under any stochastic dependence.

Mixing time and empirical distributions. To connect Markov dependence to our drift-to-average logic, it is convenient to focus on the empirical occupation measure of states

$$\hat{\mu}_T := \frac{1}{T}\sum_{t=1}^{T} \delta_{\omega_t} \in \Delta(\Omega).$$

When the chain is ergodic and mixes rapidly, $\hat{\mu}_T$ concentrates around μ_∞. We quantify "rapidly" through a standard total-variation mixing time. Define

$$\tau_{\mathrm{mix}} := \min\Big\{\tau \geq 1 : \sup_{\omega\in\Omega} \big\|P^\tau(\omega,\cdot) - \mu_\infty\big\|_1 \leq \tfrac{1}{4}\Big\},$$

where $\|\cdot\|_1$ is total variation distance on $\Delta(\Omega)$ (up to the usual factor $1/2$). Many equivalent definitions are available; any choice yields the same qualitative conclusion: dependence across rounds reduces the effective sample size by a factor proportional to τ_{mix}.

A standard consequence (see, e.g., concentration bounds for Markov chains) is that for bounded test functions $f : \Omega \to [-1, 1]$,

$$\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T} f(\omega_t) - \mathbb{E}_{\omega\sim\mu_\infty}[f(\omega)]\right| \;\leq\; O\Big(\frac{\tau_{\mathrm{mix}}}{T}\Big), \tag{24}$$

and, more generally, that $\mathbb{E}\|\hat{\mu}_T - \mu_\infty\|_1 \leq O(\tau_{\mathrm{mix}}/T)$ (up to constants depending on the particular mixing-time convention). The economic interpretation is that the chain reaches stationarity quickly enough that the early transient phase has vanishing weight in the T-round average.

From mixing to "drift penalties": replacing variation by stationarity gaps. Our main theorems are stated for arbitrary time-varying primitive sequences (u_t, v_t, C_t) and measure non-stationarity by path-length quantities such as $\mathcal{V}_u$ and $\mathcal{V}_C$. A Markov chain does not generally have small path length: successive states may jump, so $\sup_{x,a} |u_{t+1}(x,a) - u_t(x,a)|$ can be large even when the chain mixes fast. The right object in Markov environments is therefore not variation of consecutive primitives, but rather *distance between the empirical environment and its stationary limit.*

To formalize this, consider the stationary-prior (or stationary-environment) one-shot benchmark defined by the stationary distribution:

$U^*(\mu_\infty)$:= Stackelberg value of the one-shot problem with primitives averaged under μ_∞,

i.e., with payoff functions $\tilde{u}(x,a) := \mathbb{E}_{\omega\sim\mu_\infty}[u(\omega,x,a)]$, $\tilde{v}(x,a) := \mathbb{E}_{\omega\sim\mu_\infty}[v(\omega,x,a)]$, and an appropriate stationary analogue of the feasibility set (for example, the Minkowski average of $\{C(\omega)\}$ under μ_∞). In parallel, conditional on a realized state path, the *time-averaged* primitives correspond to the empirical distribution $\hat{\mu}_T$:

$$\bar{u}_T(x,a) = \mathbb{E}_{\omega\sim\hat{\mu}_T}[u(\omega,x,a)], \qquad \bar{v}_T(x,a) = \mathbb{E}_{\omega\sim\hat{\mu}_T}[v(\omega,x,a)],$$

and an empirical averaged feasibility set $\bar{C}_T$ obtained by Minkowski averaging $C(\omega_t)$ across t.

Because payoffs are bounded and linear in x, differences between stationary and empirical averages are Lipschitz in $\|\hat{\mu}_T - \mu_\infty\|_1$. Concretely, for any fixed (x,a),

$$\left|\bar{u}_T(x,a) - \tilde{u}(x,a)\right| = \left|\sum_{\omega\in\Omega}\left(\hat{\mu}_T(\omega) - \mu_\infty(\omega)\right)u(\omega,x,a)\right| \le B\,\|\hat{\mu}_T - \mu_\infty\|_1,$$

and similarly for v. Thus, whatever bound our general analysis delivers in terms of mismatch between realized play and the *time average* of primitives, mixing allows us to replace that mismatch (in expectation) by $O(B\tau_{\mathrm{mix}}/T)$ when comparing to a *stationary* benchmark.

If constraints also depend on ω, an analogous statement holds provided the map $\omega \mapsto C(\omega)$ is well-behaved in the sense needed to control the distance between $\bar{C}_T$ and its stationary analogue (e.g., via a Lipschitz bound in Hausdorff distance). In that case, the usual conditioning factor $\mathrm{diam}(X)/\mathrm{dist}(\cdot,\partial X)$ reappears, but the stochastic component entering the bound is still $\mathbb{E}\|\hat{\mu}_T - \mu_\infty\|_1$ (or a comparable mixing-controlled discrepancy).

A representative corollary (upper bound under swap regret). To highlight the message, suppose first that feasibility is state-invariant, $C(\omega) \equiv C$, so that only payoffs vary with ω. Then the stationary benchmark is particularly clean. Combining (a) the general anti-exploitation logic under contextual no-swap-regret with (b) the mixing control (24), we obtain the following implication: for any (possibly fully adaptive and state-informed) principal strategy and any agent satisfying contextual no-swap-regret $\mathrm{CSReg}(T)$,

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} u(\omega_t, x_t, a_t)\right] \;\le\; U^*(\mu_\infty) + O\!\left(\frac{\mathrm{CSReg}(T)}{T}\right) + O\!\left(\frac{B\,\tau_{\mathrm{mix}}}{T}\right). \quad (25)$$

Relative to the fully adversarial formulation, the role of "drift" is now played by *finite-horizon non-stationarity*: the average state distribution differs from stationarity by at most $O(\tau_{\mathrm{mix}}/T)$, and boundedness converts this into an $O(B\tau_{\mathrm{mix}}/T)$ payoff gap. Economically, even a sender who perfectly observes the current regime cannot, on average, obtain much more than the

stationary-optimal value once the horizon greatly exceeds the mixing time, unless the receiver's behavior departs from approximate obedience (captured here by swap regret).

A representative corollary (lower bound under external regret). A parallel statement holds for achievability with a fixed policy under contextual external regret. Again under state-invariant feasibility for simplicity, there exists a principal strategy computed from the stationary one-shot problem (equivalently, from primitives averaged under μ_∞) such that, for any agent satisfying contextual no-regret $\mathrm{CReg}(T)$,

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} u(\omega_t, x_t, a_t)\right] \;\geq\; U^*(\mu_\infty) \;-\; \tilde{O}\left(\sqrt{\frac{\mathrm{CReg}(T)}{T}}\right) \;-\; O\left(\frac{B\,\tau_{\mathrm{mix}}}{T}\right). \tag{26}$$

The interpretation mirrors the persuasion specialization: when the receiver is only disciplined against signal-by-signal fixed deviations, the sender does not need to model or track the Markov dynamics to secure near-stationary-optimal long-run performance. The only Markov-specific loss is the vanishing transient term $O(B\tau_{\mathrm{mix}}/T)$.

When can the agent learn without modeling the dynamics? A subtle point in Markov environments is whether a bounded-regret agent is *behaviorally plausible* when payoffs are temporally correlated. Our view is that it often is, and the reason is conceptual: the agent's learning task in our model is not "estimate P and solve a dynamic program," but rather "choose actions that perform well against the realized stream of recommendation contexts." Standard online-learning algorithms guarantee regret bounds without any stochastic assumptions, hence they remain valid (and are in fact conservative) under Markov dependence.

This matters for applications in which the agent is a human, a firm, or a downstream algorithm that responds to a platform's signals. In such settings, it is often unrealistic to assume the agent knows the state space Ω, the transition matrix P, or even that the environment is Markov. What is more realistic is that the agent can evaluate realized payoffs (or gradients, or bandit feedback) and adjust its behavior so that, in hindsight, it does not systematically prefer a simple deviation mapping from signals to actions. External regret corresponds to "I should not have consistently taken a different fixed action upon seeing a given signal," while swap regret corresponds to "I should not have consistently applied a remapping from recommended actions to alternative actions." Neither notion requires an internal model of P.

From the principal's perspective, this is exactly the discipline that limits exploitation. If the principal attempts to encode predictive information

about future states into signals, a sophisticated agent *could* in principle use such signals to form intertemporal plans. But our regret-based formulation is intentionally myopic: the agent reacts to the signal in the current round. Under this behavioral restriction, the only way Markov dependence matters for long-run outcomes is through how quickly the distribution of states converges to stationarity, which is precisely what τ_{mix} captures. Put differently, the principal may know the dynamics, but if the agent is primarily "learning how to respond" rather than "learning the world model," then the relevant comparison is to the stationary one-shot benchmark, with an explicit and vanishing finite-horizon correction.

Limitations and scope. Two limitations are worth keeping in view. First, when τ_{mix} is large relative to T, the stationary benchmark is not the right yardstick: the chain may spend most of the horizon in a transient region, and the $O(B\tau_{\mathrm{mix}}/T)$ term need not be small. In that case our more general, pathwise drift measures (or a benchmark based on the empirical distribution $\hat{\mu}_T$) may be more informative.

Second, if the agent is forward-looking and can condition its actions on inferred states (for example, by using the principal's signals as a state estimator), then static regret notions may understate the agent's strategic capabilities. Addressing that case would require a genuinely dynamic model of the agent's objective and information (e.g., policy regret or reinforcement-learning guarantees). Our contribution here is complementary: it isolates what can be said when the agent's adaptation is powerful enough to ensure no-regret behavior, but not premised on correct modeling of the Markov dynamics.

In summary, Markov mixing provides a clean route from stochastic dependence to explicit finite-horizon corrections. It therefore yields a particularly interpretable instance of our general message: as long as the agent learns to be approximately obedient in an online sense, the principal's long-run advantage from adaptivity is sharply limited, and in rapidly mixing environments the limiting object is the stationary one-shot Stackelberg value up to a vanishing $O(\tau_{\mathrm{mix}}/T)$ term.

8 Tightness and examples: why the drift and regret terms are not artifacts

Our bounds isolate two distinct "sources of slack": (i) *environmental non-stationarity* (captured by variation-type terms such as $\mathcal{V}_u$ and $\mathcal{V}_C$, or by a mixing-controlled stationarity gap), and (ii) *behavioral slack* on the agent side (captured by $\mathrm{CSReg}(T)$ or $\mathrm{CReg}(T)$). Because both sources enter additively, it is natural to ask whether one can sharpen the dependence—for example, replacing the drift penalty by something smaller, or upgrading the

$\sqrt{\cdot}$-type dependence that appears in the achievability guarantee. In this section we give stylized examples showing that, up to constants (and up to the conditioning factor for drifting constraints), these terms are unavoidable. The examples are deliberately simple: their role is not realism, but to make clear which *mechanism* forces each term to appear.

8.1 A drift penalty is necessary: adaptivity can buy exactly $\Theta(\mathcal{V}_u)$

We begin with a construction where the principal can outperform the averaged one-shot benchmark by an amount proportional to the path-length of payoffs, even when the agent is perfectly disciplined (zero regret) and the principal faces no feasibility complications. For simplicity take $B = 1$ and let $X = [0, 1]$, $A = \{1, 2\}$, and $S = \{1, 2\}$, with the interpretation that each signal $s \in S$ recommends an action. Let the agent's payoff be time-invariant and linear in x:

$$v(x, 1) = x, \qquad v(x, 2) = 1 - x.$$

Thus the agent strictly prefers action 1 when $x > 1/2$ and strictly prefers action 2 when $x < 1/2$ (ties at $x = 1/2$), and each action is uniquely optimal for some x, so there is a positive inducibility gap away from the knife-edge point.

Now let the principal's payoff *alternate across rounds* but not depend on x:

$$u_t(x, 1) = \mathbf{1}\{t \text{ odd}\}, \qquad u_t(x, 2) = \mathbf{1}\{t \text{ even}\}.$$

In odd rounds the principal values action 1 and in even rounds values action 2. Consider two benchmarks:

The averaged one-shot benchmark. The time-averaged principal payoff satisfies $\bar{u}(x, 1) = \bar{u}(x, 2) = 1/2$ (up to an $O(1/T)$ edge effect). Hence the Stackelberg value of the averaged one-shot problem is

$$\bar{U}^* = \frac{1}{2},$$

because no matter what scheme is used in the averaged problem, the principal cannot obtain more than $1/2$ when both actions yield identical averaged payoff.

An adaptive, state-informed principal. In the repeated game, the principal observes u_t and can choose $(\pi_t, x_{t,s})$ accordingly. A simple deterministic strategy is:

if t is odd, send $s = 1$ with $x_{t,1} = 1$; if t is even, send $s = 2$ with $x_{t,2} = 0$.

Given v, the agent's best response is $a_t = 1$ when it sees $x = 1$ and $a_t = 2$ when it sees $x = 0$. Thus the principal earns payoff 1 every round, so the realized average payoff is 1.

Gap and variation. The gain over the averaged benchmark is $1 - \bar{U}^* = 1/2$. Meanwhile the variation measure is large:

$$\mathcal{V}_u = \frac{1}{T}\sum_{t=1}^{T-1} \sup_{x\in X, a\in A} |u_{t+1}(x,a) - u_t(x,a)| = \frac{T-1}{T} \cdot 1 \approx 1.$$

Hence the improvement $1 - \bar{U}^*$ is $\Theta(\mathcal{V}_u)$ (with $B = 1$). By scaling the amplitude of the alternation, one can make this proportionality exact: for $\Delta \in (0, 1]$, define $u_t(x, 1) = \frac{1}{2} + \frac{\Delta}{2}$ and $u_t(x, 2) = \frac{1}{2} - \frac{\Delta}{2}$ on odd t and swap on even t. Then $\bar{U}^* = \frac{1}{2}$ while the adaptive principal achieves $\frac{1}{2} + \frac{\Delta}{2}$, so the gap is $\Delta/2$, whereas $\mathcal{V}_u \approx \Delta$. This shows that a drift penalty of order at least $\Omega(\mathcal{V}_u)$ (and hence of order $\Omega(B\mathcal{V}_u)$ under the original scaling conventions) cannot be removed in general: it is precisely the amount by which an informed principal can "track" the changing payoff landscape.

8.2 Why drifting constraints must be accompanied by conditioning

The role of $\mathrm{dist}(\bar{C}, \partial X)$ is more geometric: it governs how stably one can convert time-varying feasibility into a single averaged feasibility condition. The need for such a conditioning factor can already be seen in one dimension. Let $X = [0, 1]$ and suppose the principal payoff is increasing in the decision (so the principal wants to push x upward), but the feasible sets impose an upper bound that drifts:

$$C_t = \{x \in [0, 1] : x \le b_t\}, \qquad b_t \in (0, 1).$$

Then the Minkowski average constraint is $\bar{C} = \{x : x \le \bar{b}\}$ with $\bar{b} = \frac{1}{T}\sum_t b_t$, and

$$\mathrm{dist}(\bar{C}, \partial X) = 1 - \bar{b}.$$

When $\bar{b}$ is close to 1, the averaged constraint set lies close to the boundary of X and small perturbations of the b_t become hard to "absorb": a change of size δ in some b_t forces a change of comparable size in the feasible x_t, but relative to the tiny remaining slack $1-\bar{b}$ this is a *large* fractional perturbation. In the multi-dimensional setting (or when feasibility is imposed on averages $\sum_s \pi_{t,s} x_{t,s}$ rather than on a scalar), the same phenomenon manifests as follows: to repair a sequence of average-feasible points so that it is feasible for a single averaged set, one applies an affine correction whose operator norm scales like $\mathrm{diam}(X)/\mathrm{dist}(\bar{C}, \partial X)$. Thus, as $\mathrm{dist}(\bar{C}, \partial X) \downarrow 0$, any bound that is uniform over drifting C_t must deteriorate. This is not a proof of the exact

constant in our conditioning term, but it explains why *some* dependence of this form is information-theoretically unavoidable: near the boundary, feasibility "amplifies" drift.

8.3 Linear dependence on swap regret is tight

We next show that the linear dependence on $\mathrm{CSReg}(T)/T$ in the anti-exploitation (upper) bound cannot, in general, be improved. The idea is that swap regret is exactly the resource that limits how often the agent can systematically "mis-implement" a recommended action; if the principal's payoff is concentrated on those mis-implementations, the principal can gain an amount proportional to the swap-regret budget.

Consider a stationary environment with $X = \{x^\star\}$ a singleton (so feasibility and x are irrelevant), $S = \{1\}$ a single signal, and $A = \{1, 2\}$. Let the agent's payoffs satisfy

$$v(1) = 1, \qquad v(2) = 0,$$

and let the principal's payoff be the reverse:

$$u(1) = 0, \qquad u(2) = 1.$$

The unique best response for the agent is action 1; hence the one-shot Stackelberg benchmark is $\bar{U}^* = 0$.

Now consider any realization of play in which the agent plays action 2 on m rounds. Define a swap mapping $d : A \to A$ by $d(2) = 1$ and $d(1) = 1$. On each of the m rounds where the agent played 2, this deviation would have improved the agent's payoff by 1. Therefore the agent's swap regret is at least m. If the agent satisfies $\mathrm{CSReg}(T) \le R$, we must have $m \le R$, and thus the principal's realized average payoff is at most $m/T \le R/T$. Conversely, for any $m \le R$ there exist action sequences with exactly m plays of action 2 and swap regret m. Hence the maximal advantage the principal can extract above $\bar{U}^*$ is *exactly* $\Theta(R/T)$ in this example. This pins down the linear scaling: no bound of order $o(\mathrm{CSReg}(T)/T)$ can hold uniformly over all principal strategies and all swap-regret-bounded agents, because the principal can always "monetize" the agent's allowed remapping mistakes.

8.4 Why the $\sqrt{\mathrm{CReg}(T)/T}$ dependence is unavoidable

Finally, we give an example showing that the square-root dependence appearing in the achievability guarantee under external regret cannot be improved without additional regularity (e.g., strong concavity, stochastic assumptions, or an explicit margin condition at the optimum). The mechanism is a familiar knife-edge from Stackelberg problems: the principal's averaged optimal policy may rely on the agent being (almost) indifferent, but an external-regret bound does not force consistent tie-breaking. To secure obedience,

the principal must introduce an incentive margin, and the cost of doing so trades off against the fraction of "mistakes" allowed by regret, producing a $\sqrt{\cdot}$ rate.

Let $X = [0, 1]$, $S = \{1\}$, and $A = \{1, 2\}$. The agent's payoff is as in the drift example:

$$v(x, 1) = x, \qquad v(x, 2) = 1 - x,$$

so action 1 is optimal when $x > 1/2$ and action 2 is optimal when $x < 1/2$. Let the principal's payoff be

$$u(x, 1) = 1 - x, \qquad u(x, 2) = 0.$$

The principal would like to induce action 1 while keeping x small. In the one-shot Stackelberg problem with favorable tie-breaking, the principal chooses $x^\star = 1/2$, the agent is indifferent, and the principal obtains

$$\bar{U}^* = u(1/2, 1) = \frac{1}{2}.$$

Now fix a horizon T and suppose the agent is only constrained by external regret: for the single signal, the deviation class reduces to choosing a fixed action in hindsight. Let $\varepsilon := \mathrm{CReg}(T)/T$. Consider any *fixed* principal policy that plays $x = 1/2 + \gamma$ for some $\gamma \in [0, 1/2]$ (we can focus on $x \geq 1/2$ because otherwise the agent prefers action 2 and the principal's payoff is 0). Then action 1 exceeds action 2 in agent payoff by 2γ. Construct an agent behavior that plays action 2 on a fraction

$$p = \min\left\{1, \frac{\varepsilon}{2\gamma}\right\}$$

of rounds and action 1 otherwise. The agent's regret with respect to always playing action 1 is exactly $(2\gamma) \cdot pT \leq \varepsilon T$, so this behavior is consistent with the external-regret constraint.

Against this agent, the principal's expected average payoff is at most

$$(1 - p)\, u(1/2 + \gamma, 1) = (1 - p)\left(\frac{1}{2} - \gamma\right).$$

When $\gamma \geq \varepsilon$ (so that $p = \varepsilon/(2\gamma) \leq 1$), the gap to $\bar{U}^* = 1/2$ is at least

$$\frac{1}{2} - \left(1 - \frac{\varepsilon}{2\gamma}\right)\left(\frac{1}{2} - \gamma\right) \;\geq\; \gamma + \frac{\varepsilon}{4\gamma},$$

where the last inequality uses $\frac{1}{2} - \gamma \geq \frac{1}{2}$ for $\gamma \leq 1/2$ up to constants. Minimizing $\gamma + \varepsilon/(4\gamma)$ over $\gamma > 0$ yields an unavoidable loss of order $\sqrt{\varepsilon} = \sqrt{\mathrm{CReg}(T)/T}$. Intuitively, if we choose γ small to keep $u(1/2 + \gamma, 1)$ close to $1/2$, then the agent can afford to play the wrong action on a relatively large fraction of rounds while maintaining low regret (because the incentive

gap 2γ is small). If instead we choose γ large to create a robust incentive margin, we pay linearly in γ in the principal objective. The optimal balance is at $\gamma \asymp \sqrt{\varepsilon}$, producing a $\sqrt{\varepsilon}$ shortfall.

This example also clarifies what additional structure would be needed to beat the square-root rate: one must rule out (or control) these knife-edge optima by imposing either a strict separation between best and second-best agent actions at the optimum (a margin condition at the *chosen* x), or sufficient curvature/regularization that makes the principal's loss from introducing a margin smaller than linear.

Takeaway. Taken together, these constructions justify the qualitative shape of our guarantees. Drift-type terms are necessary because an informed principal can track changing primitives by exactly the amount the environment moves. Conditioning terms are necessary because feasibility repair becomes unstable near the boundary. And the regret terms exhibit a genuine asymmetry: the principal can convert a swap-regret budget into an additive $\Theta(\mathrm{CSReg}(T)/T)$ advantage, while external regret is too weak to prevent knife-edge indifference from degrading robust performance at a $\Theta\big(\sqrt{\mathrm{CReg}(T)/T}\big)$ rate.

9 Discussion and implications: robustness, regulation, and modeling choices under drift

Our results and tightness examples jointly highlight a basic tension that is easy to miss when one starts from a stationary persuasion benchmark. On the one hand, discipline on the agent side (no-regret, and especially no-swap-regret) sharply limits what an adaptive principal can extract *relative to an averaged one-shot commitment problem*. On the other hand, when primitives drift, that averaged benchmark is itself only an approximation to what is feasible in a finite horizon; the approximation error is not an analysis artifact but a real wedge created by time variation. In this concluding discussion we step back from the formal bounds and ask what these observations mean for (i) robust design and regulation, (ii) benchmark selection, and (iii) extensions that relax the informational assumptions or incorporate additional frictions.

9.1 What robust design and regulation can and cannot "fix" under drift

A natural policy instinct is that if principals (platforms, senders, regulators, recruiters) can adapt their information policies to fine-grained fluctuations, then restricting adaptivity or imposing transparency should eliminate exploitation. Our bounds suggest a more nuanced message: behavioral discipline on the agent side can indeed cap exploitation, but the cap must

move with the environment. Even if the agent is perfectly "rational" in the learning-theoretic sense (vanishing swap regret), a principal who observes the current payoff structure can still outperform the stationary or averaged benchmark by an amount proportional to drift. Put differently, regulation that targets *strategic manipulation* may not eliminate *intertemporal selection*: when the world changes, a principal can lawfully re-optimize.

This distinction is useful because it separates two families of interventions.

1. *Interventions that reduce behavioral slack.* If we interpret the agent as a boundedly rational decision-maker (or an algorithm) whose guarantee is of the form $\mathrm{CSReg}(T) = o(T)$, then any intervention that strengthens the agent's response map—better tooling, audits, debiasing, or improved feedback that supports lower regret—directly tightens the additive term involving $\mathrm{CSReg}(T)/T$. In environments where drift is small, this can be close to a complete solution: exploitation opportunities vanish at essentially the same rate as the agent's swap-regret budget.

2. *Interventions that reduce environmental non-stationarity.* When drift is large, the binding term is instead $B\mathcal{V}_u$ (and the constraint-variation term when C_t moves). Here, limiting principal adaptivity does not remove the underlying wedge between any fixed benchmark and realized payoffs; rather it changes *who bears the cost* of non-stationarity. For example, a rule that requires a principal to commit to a single policy for long windows makes the principal absorb drift risk (reducing the principal's ability to track), while a rule permitting frequent re-optimization passes drift risk to the agent (who now faces a moving target). Neither choice makes drift disappear; it allocates it.

A practical implication is that "robust persuasion design" should be evaluated together with a *drift model*. If drift is primarily exogenous (seasonality, macro conditions, shifting user composition), then one should not expect commitment-style regulation to recover a stationary benchmark. If drift is primarily endogenous (the principal's own interventions change u_t or v_t over time, e.g., through habituation or congestion), then the relevant object is not simply $\mathcal{V}_u$ but a joint dynamic system where the principal can *create* drift; bounding or taxing such induced variation may be a more direct lever.

Finally, our conditioning discussion for drifting constraints emphasizes an often-overlooked point: feasibility requirements that are close to binding are inherently fragile. In policy terms, if constraints (budget caps, fairness constraints, safety envelopes) leave little slack, then even modest fluctuations in feasible sets can force large reallocations. This is not an argument against stringent constraints, but it is an argument for acknowledging that enforcement will be sensitive to measurement error and short-run shocks precisely

when slack is smallest.

9.2 How to choose a benchmark: averaged, stationary, or dynamic?

A central modeling decision is what the principal should be compared to. We used $\bar{U}^*$, the Stackelberg value of the *averaged one-shot* problem with primitives $(\bar{u}, \bar{v}, \bar{C})$, because it is (i) well-defined even when the principal is fully informed and adaptive, and (ii) tightly linked to the occupation-measure arguments that connect repeated play to a static feasible scheme. However, benchmark choice is ultimately normative and application-dependent, and the tightness examples show that different benchmarks answer different questions.

Averaged one-shot benchmark $\bar{U}^*$. This benchmark treats drift as "noise" around a stable underlying environment. It is appropriate when a regulator or analyst believes that, absent adaptivity, the relevant target is a time-average notion of performance (e.g., average welfare, average conversion, average compliance). Under this view, our upper bound with swap regret can be read as an *anti-exploitation* guarantee: no adaptive strategy can systematically beat the averaged commitment frontier except through behavioral slack or drift. The cost of this benchmark is that it may understate what is achievable if drift is predictable and societally acceptable to track.

Stationary benchmark under Markov structure. In dynamic persuasion with an ergodic Markov state, the stationary-prior value $U^*(\mu_\infty)$ is often the object of interest because it corresponds to a long-run equilibrium with stable beliefs. Our specialization indicates that, for finite T, the relevant comparison includes a mixing-controlled bias term, essentially quantifying how much time the process spends away from stationarity. This is a useful diagnostic: if τ_{mix} is large relative to T, then the stationary benchmark is simply not a good approximation, and disagreements about "manipulation" may actually be disagreements about whether the horizon is long enough for stationarity to be meaningful.

Dynamic oracle benchmarks. In some applications, one might instead compare to a *non-anticipating oracle* that can choose π_t as a function of current primitives but is restricted in complexity or switching. For instance, one can define a class Π of admissible principal policies (e.g., Lipschitz in time, or with at most K switches) and benchmark against

$$U^*(\Pi) := \sup_{\pi_{1:T} \in \Pi} \frac{1}{T} \mathbb{E}\Big[\sum_{t=1}^{T} u_t(x_t, a_t)\Big].$$

Such benchmarks better match environments where adaptivity is not itself suspicious (e.g., pricing under volatile costs), and the goal is instead to separate learning limitations from strategic effects. The drawback is interpretability: once the benchmark is dynamic, it is harder to say what "exploitation" means, because outperformance is no longer the relevant metric.

In our view, the right benchmark should be chosen by asking a concrete counterfactual: *what policy could the principal have committed to ex ante that would be viewed as legitimate, and what information would have been available?* The averaged benchmark corresponds to commitment without time indexing; the stationary benchmark corresponds to commitment under a stable prior; dynamic oracle benchmarks correspond to legitimate re-optimization subject to explicit frictions. Our framework can be adapted to each, but the drift terms will reappear in some form because they express a genuine mismatch between realized and reference primitives.

9.3 Extensions and limitations

We close by outlining three extensions where we expect the same decomposition—behavioral slack plus environmental drift—to remain informative, but where the technical objects would change.

Principal learning (two-sided uncertainty). We assumed the principal knows u_t, v_t, C_t (or observes a state generating them). In many settings the principal is also learning: a platform experiments with ranking rules, a regulator learns compliance responses, a firm learns demand. If the principal chooses π_t based on noisy feedback, then the relevant performance notion becomes a principal regret relative to a benchmark class, and the agent's regret interacts with the principal's exploration. Conceptually, one expects an additional term reflecting the principal's learning error (e.g., $\mathrm{PReg}(T)/T$) on top of the $\mathrm{CSReg}(T)/T$ and drift penalties, but the more delicate point is that exploration itself may *increase* apparent drift by inducing payoff fluctuations. A clean separation would require modeling the state process and feedback channel explicitly, and distinguishing exogenous variation from endogenous experimentation.

Agent private information and dynamic persuasion. Our agent had no private information, so signals serve only as recommendations that shape behavior via incentives in v_t. Many persuasion problems instead feature agent types or private signals. Introducing private information changes both the feasible set (Bayes plausibility replaces the simple average constraint) and the meaning of regret constraints (the agent's deviations may be type-contingent). Nonetheless, the same high-level question persists: can an informed sender exploit an adaptive receiver beyond a natural static bench-

mark? We conjecture that swap-regret-type discipline will again be the right behavioral notion when the receiver can condition on both messages and realized actions (or reports), while drift will again govern the gap between finite-horizon performance and any static benchmark based on averaged primitives. The main additional complication is that, with types, drift may occur not only in payoffs but also in the distribution of types, raising the question of which distribution should be averaged and how quickly beliefs can track it.

Attention, message complexity, and informational frictions. We took S large enough to encode recommendations, but in practice message spaces are limited (few disclosure categories, coarse scores) and agents have attention constraints (bounded memory, limited processing). These constraints can be modeled either as restrictions on $|S|$ or as costs that penalize complex ρ_t. The immediate implication is that inducibility may fail even in the averaged problem: the principal might be unable to separate actions cleanly, effectively shrinking the attainable $\bar{U}^*$. At the same time, limited attention can *increase* behavioral slack (larger regret bounds for feasible agent algorithms), making the $\mathrm{CSReg}(T)/T$ term economically salient. This suggests a design principle with a policy flavor: when drift is unavoidable, one can mitigate manipulation concerns either by improving agent-side tooling (reducing regret) or by reducing complexity demands (smaller S, simpler mappings), but these levers trade off against efficiency because simpler policies may also reduce the frontier itself.

A final limitation. Our variation measures $\mathcal{V}_u, \mathcal{V}_v, \mathcal{V}_C$ are worst-case (supremum over x and a), which is appropriate for uniform guarantees but may be conservative in applications where drift is localized to regions of X that are never reached. Refining the analysis to "path-dependent" variation along realized play could tighten constants and improve empirical relevance, but it would also complicate benchmark interpretation because the benchmark would then depend on the induced path.

9.4 Takeaway

The main conceptual contribution of the framework is to make explicit that repeated principal–agent interactions with learning agents are governed by two independent scarcity constraints: the agent's ability to implement a stable best-response mapping (captured by regret), and the analyst's ability to summarize a changing environment by a single static object (captured by drift and conditioning). Regulation and robust design can meaningfully reduce the first, and sometimes can reallocate the burden of the second, but neither can eliminate the second without additional assumptions on the environment process. In that sense, the model illuminates a tradeoff: disci-

plining behavior curbs manipulation, while understanding and modeling drift determines which benchmark is economically coherent in the first place.