Mixture-of-Experts Actor-Critic for Regime-Switching MDPs: Impossibility Results and Performance Guarantees

MING SHI, University at Buffalo (SUNY), USA

Performance-critical computer and communication systems routinely traverse heterogeneous operating regimes, including light/heavy traffic, congestion episodes, workload phase shifts, mobility-induced channel changes, and benign/adversarial operation. This regime heterogeneity induces structured nonstationarity: the Markov decision process (MDP) governing the system can switch in its transition dynamics and/or cost structure. We show that this phenomenon is not merely a modeling nuisance, but a structural obstacle for monolithic stationary actor-critic methods when objectives couple efficiency with systems metrics such as stability, delay, and resource costs. We formulate *regime-switching MDPs* (RS-MDPs) with an unobserved, piecewise-constant regime process and evaluate performance against a regime-aware benchmark that applies the per-regime optimal stationary policy. We then propose a *regime-aware mixture-of-experts actor-critic* (RA-MoE-AC) algorithm that combines expert policies, an online gating mechanism for regime-adaptive selection, and a lightweight safety projection that enforces minimum use of a stabilizing expert. Our contributions are twofold. First, we prove impossibility theorems showing that any stationary policy can suffer a non-vanishing optimality gap against the regime-aware benchmark, and that regime mismatch can destroy queue stability even when each regime is individually stabilizable. Second, for RA-MoE-AC we derive switching-aware performance bounds whose leading terms scale as $\tilde{O}(\sqrt{T \log M} + S_T \log M + S_T t_{\text{mix}})$, plus approximation terms that decrease as the expert class is enriched (with larger M), and establish strong stability in queueing. Here, T is the horizon, S_T the number of regime switches, and t_{mix} the per-regime mixing time.

ACM Reference Format:
Ming Shi. 2026. Mixture-of-Experts Actor-Critic for Regime-Switching MDPs: Impossibility Results and Performance Guarantees. In *Proceedings of (SIGMETRICS'26).* ACM, New York, NY, USA, 45 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Modern networked systems, e.g., wireless access networks, edge/cloud platforms, and cyber-physical infrastructures, are usually nonstationary [12, 14, 27, 37]. They operate under shifting traffic intensities, changing connectivity and interference, evolving workload mixes, time-varying resource prices, and occasional failures or adversarial disruptions. These effects are often *structured*. For certain periods, the system behaves according to a relatively stable *operating regime*, and then switches to a different mode in which the dominant bottlenecks, dynamics, and costs change.

We study the fundamental algorithmic and theoretical consequences of such *regime switching* through a latent mode variable z_t that selects among a finite set of stationary/static Markov decision processes (MDPs). This abstraction captures canonical phenomena in systems, e.g.,

- *Queueing and scheduling.* In wireless scheduling, routing, and cross-layer control, policies that are efficient in light traffic can be persistently misaligned in heavy traffic, where stability margins

Author's Contact Information: Ming Shi, mshi24@buffalo.edu, University at Buffalo (SUNY), Buffalo, New York, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGMETRICS'26, Ann Arbor, MI
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

and backlog growth dominate [17, 25, 30, 38]. Mobility and interference can also fundamentally shift the effective service process, changing which users are bottlenecks.

- *Edge/cloud and data centers.* Workload phases (e.g., diurnal patterns, flash crowds, and workload-mix shifts) can abruptly change resource bottlenecks, (e.g., CPU, network, and memory) and the right provisioning logic [7, 18].
- *Cyber-physical and security-aware systems.* Systems may switch between benign operation, partial failures, and adversarial disturbance modes, changing both transition laws and costs (e.g., penalties for risk exposure or safety violations) [9, 27].

A central implication is immediate: when z_t changes, the best control policy can change with it. Our goal is close to a *best-of-many-worlds* guarantee where the *active* "world" (regime) itself can change over time, rather than the classic best-of-both-worlds paradigm [32, 43].

1.1 Why Regime Switching Breaks the Stationary Actor-Critic Method

Actor-critic methods are attractive in large-scale systems because they are online, scalable, and compatible with function approximation [8, 22, 23, 39]. However, standard analyses typically assume a single stationary MDP. Under regime switching, three core objects move simultaneously: (i) the transition kernel (hence the stationary occupancy measure), (ii) the cost landscape, and (iii) the critic fixed point. As a result, the critic becomes a moving target, advantage surrogates become biased after switches, and actor updates can chase transient artifacts.

A tempting counter-argument is that a sufficiently expressive parameterization (e.g., using deep learning and/or neural networks [6, 42]) should learn a single policy that "works everywhere." Our results show that this intuition is not just practically fragile but can be theoretically false. Even in benign dynamics, a single stationary policy may incur a *non-vanishing* performance gap relative to a regime-aware benchmark. Worse, in queueing systems, sustained regime mismatch creates a persistent service deficit, which leads to linear backlog growth and loss of stability.

To characterize systems-level behavior, e.g., post-switch transients, stability regions, and backlog/delay scaling, we adopt a regime-aware and stability-centric evaluation lens. On the *efficiency side*, we compete with a regime-aware benchmark that applies the per-regime optimal stationary policy on each segment, and we ask how the excess cost scales with the number of switches and the per-regime mixing time. On the *safety side*, for queueing instantiations we require strong stability, and we explicitly enforce a Lyapunov-drift safeguard rather than relying on stability as an emergent byproduct of learning. This emphasis is also dictated by our lower bounds in Section 4. That is, without an explicit regime-adaptive mechanism, both tracking efficiency and stability can fail.

1.2 Regime-Switching Markov Decision Processes

We model the system as a *regime-switching MDP* with a finite family of stationary MDPs $\{\mathcal{M}^{(z)}\}_{z\in\mathcal{Z}}$. A latent piecewise-constant process $\{z_t\}$ selects the active regime at time t. The agent does not observe z_t and the switching times. Our primary performance metric is *tracking regret* against the regime-aware benchmark that applies the per-regime optimal stationary policy on each segment.

There are several new challenges under regime switching. First, regime heterogeneity can create structural (policy-class) mismatch. Different regimes may induce *conflicting* optima, e.g., the optimal action flips on a frequently visited state, so any single stationary policy must compromise and can be persistently suboptimal. Second, the regime is latent, so fast post-switch inference is unavoidable. The agent must identify the active mode from the feedback and reallocate control quickly after switches. Third, critic learning becomes nonstationary. The relevant average-cost Bellman/Poisson fixed point (and hence advantage surrogates) changes across regimes, so critics must track segment-wise targets. Otherwise, temporal-difference (TD) bias propagates into the actor updates. Fourth, in queueing systems, inference and exploration errors are state-amplifying. Sustained mis-selection yields a positive service deficit (negative drift fails), which causes backlog to grow even when each regime is individually stabilizable. Together, these challenges motivate a agent that represents multiple regime-specialized behaviors, performs online mode selection, controls post-switch transients via timescale separation, and enforces a stability floor in queueing.

1.3 Main Contributions and Results

The main contributions and results in this paper are summarized as follows.

- **Impossibility results (Section 4).** We establish two complementary lower bounds that formalize when regime adaptivity is *structurally necessary*. First, we construct an RS-MDP with regime-independent dynamics and conflicting per-regime optimal actions, for which every stationary policy incurs linear tracking regret $\mathrm{Reg}(T) = \Omega(T)$ against the regime-aware benchmark, and this persists even under slow regime-switching with an arbitrary minimum segment length $L_{\min}$ (Section 4.1). Second, we show that in queueing systems regime mismatch can destroy stability: no fixed randomized priority rule stabilizes two regimes that swap the bottleneck queue, and within a long "bad" segment the backlog grows at least linearly in $L_{\min}$ (Section 4.2). Together, these results explain why regime adaptivity is required for both efficiency and safety.
- **Algorithm: RA-MoE-AC (see discussion above and Section 5).** Our new algorithm, *Regime-Aware Mixture-of-Experts Actor-Critic* (RA-MoE-AC), is designed around four coupled mechanisms (see Algorithm 1). First, we adopt an MoE policy class with M expert actors so that different experts can represent regime-specialized behaviors. This avoids the single-policy limitation highlighted by Impossibility I and reduces the challenge to online mode selection. Second, we train a state-dependent gate using a TD-residual-based mismatch signal. Within a fixed regime, an expert whose critic is approximately Bellman-consistent exhibits small centered TD residuals, whereas after a regime switch the previously well-matched expert typically produces systematic residual spikes. The gate interprets the resulting bounded residual losses as online feedback and reallocates probability mass after switches. Third, we maintain per-expert critics and enforce timescale separation (critic fastest, gate intermediate, actor slow), so that value surrogates track segment-wise fixed points after a short burn-in while limiting the propagation of transient critic bias into actor updates. Finally, because Impossibility II shows that inference errors can be unsafe in queueing systems, we enforce an explicit stability floor via a safety projection that guarantees a minimum selection probability $p_{\min}$ for a stabilizing baseline policy embedded as a dedicated expert. This converts stability from an emergent property into an enforced constraint and enables tracking-performance analysis within a guaranteed safe envelope.
- **Achievable performance guarantees (Section 6).** Our analysis chain starts from per-regime geometric mixing, then establishes critic/baseline tracking and switching-aware gate regret, and culminates in the main tracking theorem, yielding an explicit decomposition of the tracking regret. In particular, the bound separates as follows: (i) expert-selection overhead $O(\sqrt{T \log M} + S_T \log M)$, (ii) within-segment actor-critic learning terms (sublinear in T under timescale separation), (iii) post-switch transients $O(S_T t_{\mathrm{mix}})$, and (iv) irreducible approximation errors $\mathrm{Approx}_\pi + \mathrm{Approx}_V$. Here T is the horizon, M the number of experts, S_T the number of regime switches, t_{mix} the per-regime mixing time, and Approx_π / Approx_V denote policy/value approximation errors. Consequently, if $S_T = o(T)$ and approximation errors vanish, then $\mathrm{Reg}(T)/T \to 0$.
- **Stability/backlog guarantees via safety projection (Theorem 6).** Under a standard baseline Lyapunov drift condition for the stabilizing expert, the safety projection yields strong stability and

an explicit backlog bound, i.e., $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|Q_t\|_1 \le B/\epsilon$. This modular result provides a regime-agnostic stability envelope within which the tracking guarantees operate.

2 Related work

We review adjacent lines of work and clarify what remains unaddressed in our setting.

Learning for networks and systems. Large-scale resource-management problems in networking and computing are routinely implemented atop cluster and control substrates such as Mesos and Borg [19, 41], which has motivated a wave of learning-based controllers that optimize end-to-end performance objectives directly from operational traces (e.g., deep reinforcement learning (RL) for cluster management and adaptive bitrate (ABR) control [28, 29]). Their evaluations are typically trace-driven and their analyses rarely provide stability-centric, switching-aware, and robustness guarantees against regime mismatch. In parallel, rigorous robustness and optimization guarantees have been developed, e.g., via Lyapunov/drift or performance guarantees under uncertainty, for related operational settings (e.g., data-center demand response and workload shifting) [11, 26]. Our work is closest in spirit to this latter lens, but targets the regime-switching RL setting.

Regret-optimal RL in stationary MDPs. A large theory literature studies regret/sample-complexity guarantees for stationary MDPs, mainly in episodic or communicating settings [2, 3, 22, 36]. These results are foundational, but do not directly model piecewise-stationary regime switching, latent regime inference, or stability constraints that amplify transient errors.

Nonstationary and adversarial bandits/MDPs. Nonstationarity has been studied under variation budgets, change-point models, and piecewise-stationary assumptions, predominantly in bandits and partially in MDPs [5, 10, 16, 35]. These formulations often compete with a best-in-hindsight stationary comparator or assume smoothly varying dynamics. In contrast, our benchmark is explicitly *regime-aware* (piecewise stationary), and the analysis must couple learning performance to stability metrics (backlog growth and strong stability), which leads to qualitatively different failure modes (Impossibility II) and motivates explicit stability safeguards.

Latent-regime models and latent-state RL. Regime-switching can be viewed as a latent-variable control problem. Related works include hidden-parameter or latent-task MDPs, where a latent variable indexes system modes and the learner must adapt online [13, 24]. This literature typically emphasizes transfer efficiency, whereas our focus is on *switching-aware tracking* against a regime-aware stationary benchmark together with stability guarantees. Our MoE gate provides a lightweight online mechanism for latent-mode selection that is directly tied to performance certificates (TD-residual-based losses) and to stability enforcement (safe-expert floor).

Quick change detection (QCD) and "detect-then-control." A classical approach to nonstationarity is to detect distributional changes and then restart or switch agents. Quickest change detection offers principled detectors such as CUSUM and Shiryaev-type procedures [4, 33, 40]. These tools provide strong detection-delay/false-alarm tradeoffs, but do not by themselves resolve how to maintain stability during detection uncertainty, or how to integrate detection signals with continuous control updates. Our gating mechanism can be interpreted as an online, control-coupled "soft" alternative, where TD residuals act as mismatch signals that continuously reweight experts, and the safety projection ensures stability even when the mismatch signal is noisy.

Robustness and competitiveness with imperfect predictions. A parallel systems tradition studies robustness via competitive analysis and robustness-consistency tradeoffs when algorithms leverage imperfect forecasts of demand, prices, or workloads [11, 26]. These frameworks typically benchmark against an offline clairvoyant optimum via competitive ratio and aim for graceful degradation as prediction quality deteriorates. Our regime-switching setting differs in two fundamental respects. First, uncertainty is a *latent operating mode* that changes the identity of the optimal policy, rather than a forecast of future inputs. Second, the dominant constraint is *stability*, where mis-control induces state-amplifying backlog growth and cannot be smoothed by time-averaging.

Mixture-of-experts and modular policies. MoE architectures are a standard mechanism for representing heterogeneous behaviors and enabling conditional computation [15, 21, 34]. Relatedly, modular and compositional policies have shown strong empirical effectiveness in RL [1, 20]. However, existing theory does not target tracking a regime-aware stationary benchmark under piecewise-constant switching, while simultaneously enforcing queue stability guarantees. In our work, MoE is a structural requirement dictated by our impossibility results, i.e., when regimes induce conflicting optima on frequently visited states, any single stationary policy suffers a non-vanishing gap.

Queueing control, stability, and drift-based optimization. MaxWeight and related Lyapunov-drift policies are stability-optimal for broad classes of queueing networks under stationary primitives [17, 38]. The drift-plus-penalty framework further unifies stability and long-run cost optimization under stationary randomness [31], and cross-layer control connects these ideas to wireless systems [17]. Our setting departs from this classical regime, since arrivals/service statistics and/or cost tradeoffs switch across *latent* operating regimes.

3 Problem Formulation

This section formalizes the regime-switching Markov decision process (RS-MDP) studied in this paper. Concretely, the system evolves according to one of Z regimes (operating modes), where each regime $z \in \{1, \ldots, Z\}$ is associated with its own stationary MDP $\mathcal{M}^{(z)}$. This regime abstraction captures common systems phenomena, e.g., workload phase changes in data centers, mobility/interference shifts in wireless networks, and time-varying resource prices or policy constraints in provisioning. It is also sufficiently structured to enable switching-aware performance guarantees, and stability guarantees for our queueing instantiations. For the convenience of the reader, Table 2 at the beginning of the appendix summarizes the key notation.

3.1 Regime-Switching MDP (RS-MDP)

We consider an agent interacting with a system over discrete time slots $t = 1, 2, \ldots, T$. At each time t, the system is governed by a *latent regime* z_t, which is not revealed to the agent. The agent observes the current state s_t, selects an action a_t, then incurs an instantaneous cost and the system transitions to a next state s_{t+1}. Specifically,

- $s_t \in \mathcal{S}$ denotes the system state (e.g., queue lengths, channel state, server state, workload type);
- $a_t \in \mathcal{A}$ denotes the agent's action (e.g., which queue/user to serve, how much resource to allocate);
- $z_t \in \mathcal{Z} \triangleq \{1, 2, \ldots, Z\}$ denotes the latent operating mode (regime) of the environment.

The agent does not observe z_t or the switching times of the regime process $\{z_t\}$. Instead, it only observes the realized state-action trajectory (and the instantaneous cost defined below). This modeling choice reflects many systems in which the root cause of a mode change (e.g., interference pattern, workload phase, or adversarial activity) is not explicitly revealed at decision time, which motivates regime-adaptive policies utilizing mixture-of-experts with online gating.

Per-regime stationary MDP. For each regime $z \in \mathcal{Z}$, we define a stationary MDP $\mathcal{M}^{(z)} \triangleq (\mathcal{S}, \mathcal{A}, P^{(z)}, c^{(z)})$, where:

- $P^{(z)} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel under regime z, where $\Delta(\mathcal{S})$ denotes the set of probability distributions over $\mathcal{S}$. In particular, $P^{(z)}(\cdot \mid s, a)$ specifies the distribution of the next state given a state-action pair (s, a).

- $c^{(z)} : \mathcal{S} \times \mathcal{A} \to [0, c_{\max}]$ is the instantaneous cost under regime z, i.e., the cost incurred by taking action a in state s, where $c_{\max} > 0$ is a known uniform upper bound (w.l.o.g., bounded costs can be rescaled to $[0, 1]$).

Intuitively, $\mathcal{M}^{(z)}$ describes how the system behaves if it were to remain in regime z over a time window. At time t, conditioned on $z_t = z$, the agent incurs cost $c_t = c^{(z)}(s_t, a_t)$, and the next state is drawn as $s_{t+1} \sim P^{(z)}(\cdot \mid s_t, a_t)$. Thus, the nonstationarity in our model arises from the switching of the latent regime process $\{z_t\}$, which selects which stationary MDP $\mathcal{M}^{(z)}$ governs the system at each time. The regime variable z_t can affect the system in two systems-relevant ways: (i) *switching dynamics*, where transition kernel $P^{(z)}$ changes across regimes (e.g., different channel/workload/failure statistics); and/or (ii) *switching objectives*, where instantaneous cost $c^{(z)}$ changes across regimes (e.g., energy price/carbon intensity or service-level agreement (SLA) weights).

Regime switching model. We model $\{z_t\}$ as piecewise constant with finitely many switches up to horizon T. Specifically, there exist switch times $1 = \tau_0 < \tau_1 < \cdots < \tau_{S_T} \le T$ and we set $\tau_{S_T+1} \triangleq T+1$, such that z_t is constant on each segment $\mathcal{I}_k \triangleq \{\tau_{k-1}, \ldots, \tau_k - 1\}$, $k = 1, \ldots, S_T + 1$. We denote by S_T the number of switches and by $L_{\min} \triangleq \min_{k \in \{1,\ldots,S_T+1\}} (\tau_k - \tau_{k-1})$ the minimum segment length.

3.2 Performance Metric

This subsection defines the performance metric. We first define finite-horizon (time-T) costs for any policy, and then define infinite-horizon (steady-state) average costs under a *fixed* regime, which serve to define the per-regime optimal stationary benchmark.

Finite-horizon cumulative cost and average cost. Given a (possibly history-dependent) policy π over horizon T (denoted $\pi \triangleq \pi_{1:T}$), let $\{(s_t^\pi, a_t^\pi)\}_{t=1}^T$ denote the state-action trajectory generated by π interacting with the regime-switching environment. Define the finite-horizon cumulative cost

$$C_T(\pi) \triangleq \mathbb{E}_\pi \left[\sum_{t=1}^{T} c^{(z_t)} \left(s_t^\pi, a_t^\pi \right) \right], \tag{1}$$

and the corresponding finite-horizon average cost $V_T(\pi) \triangleq \frac{1}{T} C_T(\pi)$. The expectation $\mathbb{E}_\pi[\cdot]$ is taken with respect to the randomness of the regime sequence $\{z_t\}$, the controlled state transitions under $\{P^{(z_t)}\}$, and the policy's (possibly randomized) action selection.

Infinite-horizon average cost under a fixed regime. For any stationary randomized policy $\pi \in \Pi_{\text{stat}}$ and fixed regime $z \in \mathcal{Z}$, define its steady-state infinite-horizon average cost

$$J^{(z)}(\pi) \triangleq \limsup_{N \to \infty} \frac{1}{N} \mathbb{E}_{P^{(z)},\pi} \left[\sum_{t=1}^{N} c^{(z)}(s_t, a_t) \right], \tag{2}$$

where the regime is held fixed at z for all time (and the expectation is with respect to the trajectory induced by π under $P^{(z)}$). The lim sup is used for generality, since the limit need not exist without additional ergodicity or unichain assumptions. Then, for each regime $z \in \mathcal{Z}$, define a per-regime optimal stationary policy (breaking ties arbitrarily) by

$$\pi^{*,(z)} \in \arg\min_{\pi \in \Pi_{\text{stat}}} J^{(z)}(\pi). \tag{3}$$

Regime-aware tracking regret. Since the environment can switch regimes over time, a natural comparator is the *regime-aware* (nonstationary) benchmark policy $\pi_t^*(\cdot \mid s) \triangleq \pi^{*,(z_t)}(\cdot \mid s)$, which applies the regime-optimal stationary policy for the currently active regime. Let $\{(s_t^{\pi^*}, a_t^{\pi^*})\}_{t=1}^T$ be the trajectory induced by $\{\pi_t^*\}$. Define the benchmark cumulative cost $C_T^* \triangleq \mathbb{E}_{\pi^*} \left[\sum_{t=1}^T c^{(z_t)} \left(s_t^{\pi^*}, a_t^{\pi^*} \right) \right]$, and the cumulative tracking regret

$$\mathrm{Reg}(T) \triangleq C_T(\pi) - C_T^*. \tag{4}$$

We also report the average (per-step) regret $\mathrm{Reg}(T)/T$, which is directly comparable to the finite-horizon average cost $V_T(\pi)$. The metric (4) evaluates how well an online agent tracks the best stationary behavior for each regime and is the standard benchmark for obtaining switching-aware bounds that scale with the number of regime changes.

3.3 Policy Class: Mixture-of-Experts Actor-Critic

We now specify the parametric policy class considered in this paper. The key idea is to represent regime-dependent behavior via a mixture-of-experts (MoE), i.e., different experts specialize to different operating modes, while a gating network adaptively selects experts online based on the observed state.

Actor and mixture policy. We consider a mixture policy with M experts. Let $\pi^{(m)}_{\phi_m}(\cdot \mid s)$ be the m-th expert policy parameterized by ϕ_m and denote $\phi \triangleq (\phi_1, \dots, \phi_M)$. Then, the actor with mixture policy is

$$\pi_{\theta,\phi}(a \mid s) \triangleq \sum_{m=1}^{M} g_\theta(m \mid s)\, \pi^{(m)}_{\phi_m}(a \mid s), \tag{5}$$

where $g_\theta(\cdot \mid s) \in \Delta_M$ is a gating distribution over experts, parameterized by θ. A standard choice is a softmax gate $g_\theta(m \mid s) = \frac{\exp(u_\theta^{(m)}(s))}{\sum_{j=1}^{M} \exp(u_\theta^{(j)}(s))}$, where $u_\theta^{(m)}(s)$ is a score function (e.g., linear or a shallow network). Our analysis assumes the score functions are regular enough so that $\nabla_\theta \log g_\theta(m \mid s)$ is uniformly bounded. Although $\pi_{\theta,\phi}$ is stationary as a mapping from s to a distribution over actions, it can effectively adapt to regime changes through state-dependent gating and expert specialization.

Critic and value function approximation. We maintain per-expert critics $\{V^{(m)}_{w_m}\}_{m=1}^{M}$, which estimate the (differential) value of states under expert m. For theoretical analysis, we focus on linear critics:

$$V^{(m)}_{w_m}(s) = \psi(s)^\top w_m, \tag{6}$$

where $\psi : \mathcal{S} \to \mathbb{R}^d$ is a bounded feature map with $\|\psi(s)\| \le 1$, and $w_m \in \mathbb{R}^d$ is the critic parameter for expert m.

Average-cost TD error and advantage estimate. Our performance metric is average cost (Section 3.2). Accordingly, we use the standard average-cost (relative-value) actor-critic surrogate based on a centered TD error. For each expert m, we maintain an estimate $\eta^{(m)}$ of the average cost under that expert, and define the per-expert TD error

$$\delta_t^{(m)} \triangleq c_t - \bar{c}^{(m)} + V^{(m)}_{w_m}(s_{t+1}) - V^{(m)}_{w_m}(s_t). \tag{7}$$

We use $\widehat{A}_t^{(m)}$ as an advantage estimate and a common single-step choice is $\widehat{A}_t^{(m)} \approx \delta_t^{(m)}$.

3.4 System Instantiations

This subsection provides two instantiations of our problem setting.

3.4.1 *Instantiation A: Single-Queue System with Regime-Dependent Arrivals and Energy Prices.*

We instantiate the RS-MDP with a canonical single-server queueing system whose operating conditions (workload intensity and energy price) vary across regimes. The system state is $s_t = (Q_t, x_t)$, where $Q_t \in \mathbb{R}_+$ is the queue backlog and $x_t \in \mathcal{X}$ is an exogenous mode variable (e.g., workload phase, channel condition, or server state).

Specifically, at each slot t, the agent chooses a service action $a_t \in \mathcal{A} \subseteq [0, \mu_{\max}]$. Conditioned on the latent regime $z_t = z$, arrivals $A_t^{(z)}$ are drawn and the queue evolves as

$$Q_{t+1} = \left[Q_t + A_t^{(z_t)} - a_t\right]^+. \tag{8}$$

We model the exogenous process as regime-dependent Markov dynamics $x_{t+1} \sim \mathcal{F}^{(z_t)}(\cdot \mid x_t)$, and arrivals as a regime- and state-dependent distribution,

$$A_t^{(z_t)} \sim \mathcal{D}^{(z_t)}(\cdot \mid x_t), \tag{9}$$

e.g., Poisson arrivals with mean $\lambda^{(z)}(x_t)$. This captures workload phase shifts where both the transition law of x_t and the arrival intensity depend on the current regime. We consider a per-slot cost that penalizes backlog (delay proxy) and energy expenditure with a regime-dependent price:

$$c^{(z)}(Q, x, a) = wQ + \kappa^{(z)} a^2, \tag{10}$$

where $w > 0$ weights delay/backlog and $\kappa^{(z)} > 0$ is a regime-dependent energy/price coefficient (e.g., reflecting electricity price or carbon intensity). This model captures the canonical tradeoff: in high-price regimes, aggressive service is costly, whereas in low-price regimes, aggressive service is attractive for backlog reduction. Let $s = (Q, x)$ and $s' = (Q', x')$. Using (8)-(9), the regime-z transition kernel admits the factorization

$$P^{(z)}(s' \mid s, a) = \sum\nolimits_{A \geq 0} \mathcal{D}^{(z)}(A \mid x) \cdot \mathcal{F}^{(z)}(x' \mid x) \cdot \mathbb{1}\left\{Q' = [Q + A - a]^+\right\}. \tag{11}$$

In this instantiation, regimes can alter (i) the arrival law $\mathcal{D}^{(z)}(\cdot \mid x)$, i.e., traffic intensity, (ii) the evolution of the exogenous mode $\mathcal{F}^{(z)}(\cdot \mid x)$ i.e., phase persistence, and (iii) the energy/price coefficient $\kappa^{(z)}$ in (10).

3.4.2 Instantiation B: Downlink Wireless Scheduling with Regime-Dependent Channel Law. We next instantiate the RS-MDP with a canonical downlink scheduling problem in a wireless base station serving d users. The system maintains per-user queues $Q_{t,i} \in \mathbb{R}_+$ for $i \in [d]$, and the wireless channel state at time t is denoted by $H_t \in \mathcal{H}$ (capturing fading and interference).

Specifically, the system state is $s_t = (Q_t, H_t)$, where $Q_t = (Q_{t,1}, \ldots, Q_{t,d})$. The latent regime $z_t \in \mathcal{Z}$ captures operating conditions such as mobility/interference patterns. Conditioned on $z_t = z$, we model the channel as a regime-dependent Markov process $H_{t+1} \sim \mathcal{H}^{(z)}(\cdot \mid H_t)$, where $\mathcal{H}^{(z)}$ can specialize to the i.i.d. case by dropping the conditioning. At each time, the scheduler selects one user $a_t \in \mathcal{A} \triangleq \{1, 2, \ldots, d\}$ to serve. (Extensions to selecting rate vectors or multiple users per slot are standard and omitted for clarity.) User i receives exogenous arrivals $A_{t,i}^{(z_t)}$, potentially regime-dependent (e.g., demand phases), $A_{t,i}^{(z)} \sim \mathcal{D}_i^{(z)}(\cdot \mid H_t)$. If user i is scheduled, it receives service at an achievable rate $r_i(H_t; z_t)$, which can also depend on the regime (e.g., reflecting interference levels or mobility). The per-user queue update is

$$Q_{t+1,i} = \left[Q_{t,i} + A_{t,i}^{(z_t)} - \mathbb{1}\{a_t = i\}\, r_i(H_t; z_t)\right]^+, i \in [d]. \tag{12}$$

We consider a standard delay-power objective $c^{(z)}(Q, H, a) = \sum_{i=1}^{d} w_i Q_i + \lambda^{(z)} \text{Power}(a, H)$, where $w_i > 0$ are queue weights, $\text{Power}(a, H)$ is the transmit power incurred by serving user a under channel state H (e.g., due to adaptive modulation/coding or target rate constraints), and $\lambda^{(z)} > 0$ is a regime-dependent price coefficient that can encode time-varying energy prices or tighter power constraints in certain operating modes.

In this instantiation, regimes can correspond to (i) mobility/interference patterns that change the channel law $\mathcal{H}^{(z)}$ and the achievable rate functions $r_i(\cdot; z)$, and/or (ii) traffic demand phases that change the arrival laws $\{\mathcal{D}_i^{(z)}\}$ and the power price $\lambda^{(z)}$. Because both channel statistics and traffic intensities can shift across regimes, the scheduling policy that is optimal in one regime can be persistently misaligned in another.

Definition 1 (Strong stability [31]). The queue process $\{Q_t\}$ is strongly stable if

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|Q_t\|_1] < \infty. \tag{13}$$

4 Impossibility Results: Regret Lower Bounds and Queue Instability

This section explains why regime switching fundamentally changes what can be achieved by *stationary* control rules. Even when each regime is individually stationary and admits a well-behaved optimal stationary policy, a single stationary policy can be *persistently misaligned* with at least one regime. The consequence is twofold: (i) a stationary policy can incur a *non-vanishing* average-cost gap relative to a regime-aware benchmark, and (ii) in queueing instantiations, regime mismatch can create sustained overload within long segments, leading to instability.

4.1 Impossibility I: A Stationary Policy Can Have a Non-Vanishing Cost Gap

We construct a simple RS-MDP where two regimes require *opposite* actions on a frequently visited state. Let $\mathcal{S} = \{s_0, s_1\}$ and $\mathcal{A} = \{a^+, a^-\}$. The state evolution is deterministic and regime-independent:

$$s_{t+1} = s_1 \text{ if } s_t = s_0; \text{ and } s_{t+1} = s_0 \text{ if } s_t = s_1.$$

Hence s_0 is visited exactly every other step, independent of the policy. The two regimes differ only in the cost at state s_0:

- **Regime 1:** $c^{(1)}(s_0, a^+) = 0$, $c^{(1)}(s_0, a^-) = 1$, and $c^{(1)}(s_1, a) = 0$ for both actions.
- **Regime 2:** $c^{(2)}(s_0, a^-) = 0$, $c^{(2)}(s_0, a^+) = 1$, and $c^{(2)}(s_1, a) = 0$ for both actions.

Therefore, the optimal regime-aware benchmark policy π_t^* will simply choose a^+ at s_0 in regime 1 and a^- at s_0 in regime 2, and achieve zero cost at all times.

Theorem 1 (Linear regret for stationary policies under conflicting regimes). *For the above two-regime construction, for any stationary randomized policy π, there exists a piecewise-constant regime sequence $\{z_t\}$ such that*

$$\mathrm{Reg}(T) \geq \left\lfloor \frac{T}{2} \right\rfloor \cdot \max\left\{\pi(a^+ \mid s_0), \pi(a^- \mid s_0)\right\} \geq \frac{T}{4} - \frac{1}{2}. \tag{14}$$

and hence $\mathrm{Reg}(T) = \Omega(T)$ *and* $\mathrm{Reg}(T)/T = \Omega(1)$.

Theorem 1 isolates a fundamental obstruction: when two regimes require *conflicting* actions on a state that is visited with nontrivial frequency, any *single* stationary policy must randomize and therefore be persistently suboptimal for at least one regime. Notably, this lower bound holds even though the dynamics are completely benign (deterministic and regime-independent), so the failure is not caused by slow mixing or hard exploration. Instead, it is caused by *latent regime dependence of the optimal action.* Consequently, achieving sublinear tracking regret in RS-MDPs requires a mechanism that can represent and select among multiple specialized behaviors (e.g., via mixtures) and an online adaptation/inference component (e.g., gating) that identifies which behavior is appropriate from the observation. We provide the proof sketch for Theorem 1 below, and please see Appendix C for the complete proof.

Proof sketch. Let $p \triangleq \pi(a^+ \mid s_0) \in [0,1]$ for the stationary randomized policy π. Since the transition is deterministic and alternates between s_0 and s_1 regardless of the action, the state s_0 is visited exactly $\lfloor T/2 \rfloor$ times over horizon T. In regime 1, the expected cost incurred at each visit to s_0 equals $1-p$ (because only action a^- is costly), while in regime 2 it equals p (because only action a^+ is costly). Costs at s_1 are always 0. Hence, if the entire horizon is spent in regime $z \in \{1,2\}$, the expected cumulative cost of π is

$$\mathbb{E}\left[\sum_{t=1}^{T} c^{(z)}(s_t,a_t)\right] = \left\lfloor \frac{T}{2} \right\rfloor \times \left[(1-p)\mathbb{1}\{z=1\} + p\mathbb{1}\{z=2\}\right].$$

The regime-aware benchmark π_t^* chooses the zero-cost action at s_0 for the active regime, so it achieves zero cumulative cost under either regime. Therefore, for the regime z that is worse for π ($z=1$ if $1-p \geq p$ and $z=2$ otherwise),

$$\mathrm{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} c^{(z)}(s_t,a_t)\right] \geq \left\lfloor \frac{T}{2} \right\rfloor \cdot \max\{p, 1-p\} \geq \frac{T}{4} - \frac{1}{2}, \tag{15}$$

which implies $\mathrm{Reg}(T) = \Omega(T)$ and $\mathrm{Reg}(T)/T = \Omega(1)$. □

The lower bound of regret in Theorem 1 is not an artifact of rapid switching. In fact, linear regret persists even when regimes are required to be piecewise constant with a prescribed minimum segment length $L_{\min}$.

Theorem 2 (Linear regret under slow switching (minimum segment length))**.** *Fix any stationary randomized policy π and any $L_{\min} \geq 1$. For any horizon $T \geq L_{\min}$, there exists a piecewise-constant regime sequence $\{z_t\}_{t=1}^{T}$ satisfying the segment-length constraint $L_{\min}$ such that* $\mathrm{Reg}(T) = \Omega(T)$*. In particular, one may choose either:*

(1) ***No switching:*** *$S_T = 0$ (a single regime for all t), in which case $L_{\min} = T$ and* $\mathrm{Reg}(T) \geq T/4 - 1/2$*; or*

(2) ***With switching:*** *an alternating regime sequence with segment length exactly $L_{\min}$, which yields*

$$\mathrm{Reg}(T) \;\geq\; \left(\frac{1}{4} - \frac{1}{2L_{\min}}\right)T, \;\; \textit{for all } T \textit{ that are multiples of } 2L_{\min}. \tag{16}$$

The construction in the proof highlights a key insight that *regime switching creates competing optima*. Even if the transition dynamics are benign and perfectly predictable, the optimal action can depend on an unobserved regime. Thus, achieving sublinear tracking regret requires either (i) explicit regime inference, or (ii) a policy class capable of representing multi-modal behavior (e.g., mixtures) together with an online selection mechanism. We provide the proof sketch for Theorem 2 below, and please see Appendix D for the complete proof.

Proof sketch. For (i), select the regime (1 or 2) that maximizes the stationary policy's per-step loss (as in Theorem 1). This produces $\mathrm{Reg}(T) \geq T/4$ and trivially satisfies any minimum-segment constraint since there is only one segment.

For (ii), partition time into segments of length $L_{\min}$ and alternate regimes $1,2,1,2,\ldots$. Within any segment, the state s_0 is visited at least $\lfloor L_{\min}/2 \rfloor$ times (because the chain alternates deterministically), and each visit to s_0 incurs expected cost $1-p$ in regime 1 and p in regime 2, where $p = \pi(a^+ \mid s_0)$. If there are $K = T/L_{\min}$ segments with K even, then exactly $K/2$ segments are in each regime, and the total expected cost is at least

$$\frac{K}{2}\left\lfloor \frac{L_{\min}}{2} \right\rfloor (1-p) + \frac{K}{2}\left\lfloor \frac{L_{\min}}{2} \right\rfloor p = \frac{K}{2}\left\lfloor \frac{L_{\min}}{2} \right\rfloor \geq \left(\frac{1}{4} - \frac{1}{2L_{\min}}\right)T.$$

The regime-aware benchmark achieves zero cost, hence (16) follows. □

4.2 Impossibility II: Regime Mismatch Can Destroy Queue Stability

We now show that, in queueing systems, regime mismatch can have a qualitatively stronger effect than a constant cost suboptimality, i.e., persistent mismatch within a long regime segment creates a *service deficit* that accumulates over time, leading to sustained backlog growth and loss of stability. This motivates regime-adaptive scheduling and, in our instantiations, the inclusion of a stabilizing baseline expert to protect against inference errors.

Consider two queues $Q_{t,1}, Q_{t,2}$. In each slot, the agent chooses $a_t \in \{1,2\}$ and serves one packet from queue a_t (if nonempty). Let arrivals be regime-dependent with means

- **Regime 1:** $\mathbb{E}[A_{t,1} \mid z_t = 1] = \lambda_H$, $\mathbb{E}[A_{t,2} \mid z_t = 1] = \lambda_L$,
- **Regime 2:** $\mathbb{E}[A_{t,1} \mid z_t = 2] = \lambda_L$, $\mathbb{E}[A_{t,2} \mid z_t = 2] = \lambda_H$,

where $\lambda_H \in (1/2, 1)$ and $\lambda_L > 0$ is small. We focus on stationary randomized *priority* policies that serve queue 1 with a fixed probability $p \in [0,1]$ (and queue 2 otherwise), independent of state.

Theorem 3 (No fixed randomized priority stabilizes both regimes)**.** *There exist $\lambda_H \in (1/2, 1)$ and sufficiently small $\lambda_L > 0$ such that:*

*(1) (*Per-regime stabilizability*) For each fixed regime $z \in \{1,2\}$, there exists a stationary policy that stabilizes the two-queue system when the regime is held fixed at z.*

*(2) (*Global impossibility for fixed priorities*) No stationary randomized priority policy that serves queue 1 with a fixed parameter $p \in [0,1]$ stabilizes the system under regime switching when regimes persist for sufficiently long contiguous periods.*

Theorem 3 isolates a *structural incompatibility*. Specifically, under regime 1, stability requires devoting a large service fraction to queue 1, whereas under regime 2 it requires devoting a large fraction to queue 2. When $\lambda_H > 1/2$, these requirements cannot be satisfied simultaneously by any fixed service split $(p, 1-p)$. Unlike regret gaps in cost-only objectives, a queueing mismatch is *state-amplifying*: once a queue becomes overloaded in a regime, the backlog accumulates and cannot be instantaneously eliminated after the regime changes. This is precisely why regime adaptivity and stability-aware safeguards (e.g., a safe expert or drift-based guardrails) matter in systems. We provide the proof sketch for Theorem 3 below, and please see Appendix E for the complete proof.

Proof sketch. A fixed-p policy allocates long-run service fractions $(p, 1-p)$. In regime 1, queue 1 is the heavy queue and stability requires $p > \lambda_H$ (up to an arbitrarily small slack). In regime 2, queue 2 is the heavy queue and stability requires $1 - p > \lambda_H$, i.e., $p < 1 - \lambda_H$. When $\lambda_H > 1/2$, the inequalities $p > \lambda_H$ and $p < 1 - \lambda_H$ cannot both hold. Therefore, for any fixed p, there exists a regime in which the heavy queue has a strict service deficit; if that regime persists for long intervals, backlog grows without bound, violating any strong stability notion. □

The above incompatibility implies not only instability but also an explicit *linear growth* of backlog within a long segment of the "bad" regime. This result connects directly to our piecewise-constant switching model and clarifies why slow switching does not rescue fixed stationary priorities.

Theorem 4 (Backlog grows linearly with $L_{\min}$ under slow switching)**.** *Fix any stationary randomized priority policy with parameter $p \in [0,1]$. Choose $\lambda_H \in (1/2, 1)$ and sufficiently small $\lambda_L > 0$ as in Theorem 3. Then there exists a piecewise-constant regime sequence satisfying the minimum segment length constraint $L_{\min}$ such that, for infinitely many regime-segment endpoints t,*

$$\mathbb{E}\left[Q_{t,1} + Q_{t,2}\right] \geq \Omega(L_{\min}). \tag{17}$$

More concretely, letting $\delta(p) \triangleq \lambda_H - \max\{p, 1-p\} > 0$, there exists a regime segment of length $L_{\min}$ starting at some t_0 such that

$$\mathbb{E}\left[Q_{t_0+L_{\min},i^*} - Q_{t_0,i^*}\right] \geq \delta(p)\, L_{\min} - O(1), \tag{18}$$

where $i^* \in \{1, 2\}$ is the heavy queue in that segment.

Theorem 4 quantifies the compounding effect of mismatch. Specifically, in a bad regime segment of length $L_{\min}$, the overloaded queue accumulates at least $\Theta(L_{\min})$ additional backlog in expectation. Hence, even if regime switches are infrequent (large $L_{\min}$), fixed stationary priorities are not merely suboptimal. They can be *unsafe* in the sense of creating large transient backlogs that break stability objectives. This directly motivates two design principles used later: (i) *fast adaptation* after a detected switch (via gating/expert selection), and (ii) *stability protection* when the regime inference is uncertain (via a stabilizing baseline expert with enforced minimum usage). We provide the proof sketch for Theorem 4 below, and please see Appendix F for the complete proof.

Proof sketch. Fix $p \in [0, 1]$ and define $\delta(p) \triangleq \lambda_H - \min\{p, 1 - p\} > 0$, which is positive since $\lambda_H > 1/2$ implies $\min\{p, 1 - p\} \leq 1/2$. Choose the "bad" regime so that the *heavy* queue is the one that the fixed-p policy serves *less often*: if $p \leq 1/2$, use regime 1 (queue 1 is heavy); otherwise use regime 2 (queue 2 is heavy). Let $i^* \in \{1, 2\}$ denote the heavy queue in this bad regime, and consider a regime segment $[t_0, t_0 + L_{\min} - 1]$ during which the bad regime persists. In each slot of this segment, queue i^* has expected arrival rate λ_H. Under the fixed-p priority rule, the probability of selecting the heavy queue i^* is exactly $\min\{p, 1 - p\}$ (by construction of the bad regime). Hence, whenever $Q_{t,i^*} > 0$, the expected service (departure) from queue i^* in that slot is $\min\{p, 1 - p\}$, and the one-step conditional drift satisfies

$$\mathbb{E}\left[\, Q_{t+1,i^*} - Q_{t,i^*} \mid Q_{t,i^*} > 0 \right] \geq \lambda_H - \min\{p, 1 - p\} = \delta(p). \tag{19}$$

Summing these positive drifts over the $L_{\min}$ slots yields an expected backlog increase on the order of $\delta(p)L_{\min}$, up to an $O(1)$ boundary term accounting for possible emptiness at the very beginning of the segment and the $[\cdot]^+$ truncation. Therefore, at the end of such a bad segment, $\mathbb{E}[Q_{t_0+L_{\min},i^*} - Q_{t_0,i^*}] \geq \delta(p)L_{\min} - O(1)$, proving the stated linear-in-$L_{\min}$ growth. Repeating such bad segments infinitely often (consistent with the piecewise-constant switching model) forces arbitrarily large expected backlog, precluding strong stability. □

Together, Theorems 3-4 show that in queueing systems, regime switching can turn a seemingly mild modeling change into a sharp stability challenge. Fixed stationary priorities are incompatible with regimes that swap the identity of the bottleneck queue. This justifies regime-adaptive control architectures and stability-aware safeguards in RL.

5 Algorithm Design: Regime-Aware Mixture-of-Experts Actor-Critic with Safety Projection

Our goal is to achieve sublinear tracking regret in RS-MDPs while maintaining stability in queueing instantiations. Impossibility I shows that a *single* stationary behavior can have an $\Omega(1)$ per-step gap under conflicting regimes, hence we must represent multiple specialized behaviors and *select* among them online. Impossibility II further shows that, in queueing systems, persistent regime mismatch can create a sustained service deficit and destroy stability, hence the selection mechanism must be robust to inference errors and exploration.

We propose *Regime-Aware Mixture-of-Experts Actor-Critic* (RA-MoE-AC) with Safety Projection (see Algorithm 1), which combines four tightly coupled components, including (i) M expert actors $\{\pi^{(m)}_{\phi_m}\}_{m=1}^{M}$ to represent regime-dependent behaviors, (ii) a state-dependent gating rule $g_\theta(\cdot \mid s)$ to perform online expert selection for regime inference, (iii) per-expert critics $\{V^{(m)}_{w_m}\}$ (and average-cost estimates $\{\bar{c}^{(m)}\}$) to generate low-variance advantage surrogates, and (iv) a *safety projection* that enforces a minimum selection probability on a stabilizing expert in queueing instantiations.

Algorithm 1 Regime-Aware MoE Actor-Critic with Safety Projection (RA-MoE-AC)

Require: Experts $\{\pi^{(m)}_{\phi_m}\}_{m=1}^{M}$, gating $g_\theta(\cdot \mid s)$, critics $\{V^{(m)}_{w_m}\}_{m=1}^{M}$, stepsizes $\{\beta_t, \alpha_t, \eta_t, \rho_t\}$, safe expert m_{safe}, minimum probability $p_{\min} \in (0,1)$, clipping constant C.

1: Initialize $\phi_m^{(1)}, w_m^{(1)}, \bar{c}^{(m,1)}$ for all m, and $\theta^{(1)}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Observe s_t
4: Compute gate $g_t(m) \leftarrow g_{\theta^{(t)}}(m \mid s_t)$ for all m
5: **Safety projection:** $\tilde{g}_t(\cdot) \leftarrow \Pi\big(g_t(\cdot)\,;\ \tilde{g}_t(m_{\text{safe}}) \ge p_{\min}\big)$
6: Sample expert $m_t \sim \tilde{g}_t(\cdot)$
7: Sample action $a_t \sim \pi^{(m_t)}_{\phi^{(t)}_{m_t}}(\cdot \mid s_t)$
8: Execute action a_t, and observe cost c_t and next state s_{t+1}
9: **(Selected expert) average-cost update:** $\bar{c}^{(m_t,t+1)} \leftarrow (1-\rho_t)\bar{c}^{(m_t,t)} + \rho_t\, c_t$
10: **(Selected expert) TD residual:** $\delta_t^{(m_t)} \leftarrow c_t - \bar{c}^{(m_t,t)} + V^{(m_t)}_{w^{(t)}_{m_t}}(s_{t+1}) - V^{(m_t)}_{w^{(t)}_{m_t}}(s_t)$
11: **Critic (fast timescale):** $w_{m_t}^{(t+1)} \leftarrow w_{m_t}^{(t)} - \beta_t\, \delta_t^{(m_t)}\, \nabla_w V_w^{(m_t)}(s_t)\big|_{w=w^{(t)}_{m_t}}$
12: **Actor (slow timescale):** $\phi_{m_t}^{(t+1)} \leftarrow \phi_{m_t}^{(t)} - \alpha_t\, \delta_t^{(m_t)}\, \nabla_\phi \log \pi_\phi^{(m_t)}(a_t \mid s_t)\big|_{\phi=\phi^{(t)}_{m_t}}$
13: **Gating loss:**
14: **if** full information **then**
15: Compute $\delta_t^{(m)}$ for all m via (20) and set $\widehat{\ell}_t(m) \leftarrow \ell_t(m)$ via (21)
16: **else**
17: Set $\widehat{\ell}_t(\cdot)$ via the bandit estimator (22)
18: **end if**
19: **Gating update:** update $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_t \nabla_\theta \Big(\sum_{m=1}^{M} g_\theta(m \mid s_t)\, \widehat{\ell}_t(m) \Big)\Big|_{\theta=\theta^{(t)}}$
20: **end for**
21: **return** mixture policy $\pi_{\theta,\phi}$ in (5)

5.1 Technical Difficulties and Design Ideas

Before formally introducing Algorithm 1, we highlight the core technical difficulties created by regime switching and explain the corresponding design ideas implemented in Algorithm 1.

D1 (critic drift under switching): Bellman targets change within the horizon. When z_t switches, both the stationary distribution and the (average-cost) Bellman equations change. As a result, a critic trained on one segment can be systematically biased on the next, which then corrupts actor gradients. *To address this,* we maintain per-expert critics and update the selected expert critic on the fastest timescale (Algorithm 1, TD residual and critic update; Lines 9–11). Under slow switching and per-regime mixing, these updates track the segment-local value surrogate.

D2 (latent regime inference): the agent must select the right expert without observing z_t. Sublinear tracking regret requires a mechanism that quickly concentrates probability on the expert most compatible with the current regime, even though z_t is hidden. *To address this,* we treat gating as online learning driven by a bounded mismatch loss based on TD residuals (Algorithm 1, loss computation and gate update; Lines 13–18). This directly operationalizes "regime inference" from observable trajectories.

D3 (coupled learning): gate quality depends on experts and expert learning depends on the gate. If the gate collapses early, non-selected experts stop receiving samples and cannot improve. In addition, if experts are inaccurate, the gate receives noisy mismatch signals and may oscillate after switches. *To address this,* we develop (i) timescale separation $\beta_t \gg \eta_t \gg \alpha_t$ (critic fast, gate intermediate, actor slow), (ii) clipped gating losses to control variance (Line 13), and (iii) an explicit sampling floor via safety projection (Line 5), which prevents complete starvation of the stabilizing expert.

D4 (stability under inference errors): wrong expert selection can be catastrophic in queues. Impossibility II shows mismatch can create a sustained service deficit and backlog growth within a long segment, so stability cannot be left to "emerge" from regret minimization alone. *To address this,* we embed a known stabilizing baseline π_{safe} as expert m_{safe} and enforce $\tilde{g}_t(m_{\text{safe}} \mid s_t) \geq p_{\min}$ at every time (Algorithm 1, safety projection; Line 5). This provides a direct handle for Lyapunov-drift arguments in the stability analysis.

D5 (switching overhead): the gate must react fast after a regime change. Even under slow switching, the post-switch transient dominates performance if the gate re-concentrates too slowly. *To address this,* gate stepsize η_t is chosen to balance noise and responsiveness (Line 18), and the TD-residual-based loss naturally spikes right after a switch, accelerating reweighting.

5.2 Main algorithm: RA-MoE-AC with safety projection

We present a regime-aware MoE actor-critic algorithm (with safety projection for queue stability). For each expert m, we maintain actor parameters ϕ_m, critic parameters w_m for a differential value approximation $V_{w_m}^{(m)}(s)$, and an average-cost estimate $\bar{c}^{(m)}$. At each time t, the gate produces a distribution over experts, and then we enforce safety and sample an expert to act. Given transition (s_t, a_t, c_t, s_{t+1}), we define the average-cost TD residual for expert m by

$$\delta_t^{(m)} \triangleq c_t - \bar{c}^{(m)} + V_{w_m}^{(m)}(s_{t+1}) - V_{w_m}^{(m)}(s_t). \tag{20}$$

We use the clipped squared residual as a bounded mismatch loss, i.e.,

$$\ell_t(m) \triangleq \text{clip}\big((\delta_t^{(m)})^2, 0, C\big), \tag{21}$$

so that experts whose critics are more Bellman-consistent on the current trajectory receive lower loss and therefore higher gate weight. When M is small, we compute $\delta_t^{(m)}$ (and thus $\ell_t(m)$) for all m and update the gate with full-information losses (Lines 13–14). However, when M is large, we can update using only the selected expert m_t via an unbiased importance-weighted estimator (Lines 15–17)

$$\widehat{\ell}_t(m) \triangleq \frac{\mathbb{1}\{m = m_t\}}{\tilde{g}_t(m_t \mid s_t)} \,\text{clip}\big((\delta_t^{(m_t)})^2, 0, C\big), \tag{22}$$

Our main theorems later analyze the full-information gate. However, the bandit variant follows standard EXP3-style arguments with the usual $\sqrt{M}$ dependence. Moreover, for queueing-system instantiations, we assume there exists a known stabilizing baseline policy π_{safe} (e.g., a MaxWeight-type rule) that ensures a Lyapunov drift condition for the queue component Q_t, implying positive recurrence (strong stability) of the induced queueing process. This baseline can be embedded as a dedicated "safe expert" in the mixture, and our algorithm enforces a minimum selection probability for it to guarantee stability. Specifically, our algorithm includes four components addressing the difficulties discussed in Section 5.1.

C1 (Lines 3–8): gating for safe sampling. The gate $g_{\theta^{(t)}}(\cdot \mid s_t)$ converts the latent-regime problem into online expert selection. The safety projection (Line 5) enforces $\tilde{g}_t(m_{\text{safe}}) \geq p_{\min}$, guaranteeing that stabilizing actions remain available even when regime inference is wrong. Sampling $m_t \sim \tilde{g}_t$ (Line 6) implements the mode-selection decision required to circumvent Impossibility I. Then, the algorithm acts using the selected expert policy (Line 7) and observes (c_t, s_{t+1}) (Line 8). Under regime switching, this single transition is the only information available.

C2 (Lines 9–11): fast critic tracking produces a usable advantage signal. Line 9 updates $\bar{c}^{(m_t)}$ to maintain a running estimate of the expert's average cost. Line 10 forms the centered TD residual $\delta_t^{(m_t)}$, which is an advantage surrogate for the actor and a mismatch certificate for the gate. Line 11 updates w_{m_t} on the fast timescale, reducing critic drift within a regime segment (D1).

C3 (Line 12): slow actor update avoids chasing switching transients. Line 12 performs the policy-gradient step using $\delta_t^{(m_t)}$ as the advantage estimate. Keeping α_t smaller than the gate step-sizes prevents the actor from overreacting to the immediate post-switch transient, which is critical when $L_{\min}$ is only moderately larger than the mixing time (D3–D5).

C4 (Lines 13–18): gate update reweights experts using bounded mismatch losses. Lines 13–17 define $\widehat{\ell}_t(\cdot)$ either from full-information losses (small M) or from the bandit estimator (large M). Line 18 then updates θ to reduce the expected mismatch under the gate at state s_t. After a regime change, TD residuals for mismatched experts spike, so the gate update naturally re-concentrates on the best expert for the new segment (D2, D5), while clipping controls variance and supports finite-time analysis.

6 Theoretical Analysis

This section states finite-time guarantees for Algorithm 1 that are consistent with: (i) the RS-MDP model and tracking-regret metric in Section 3, (ii) the necessity of regime adaptivity highlighted by the impossibility results in Section 4, and (iii) the MoE actor-critic with TD-residual-driven online gating and safety projection in Section 5. We emphasize *switching-aware* bounds that scale explicitly with the number of regime switches S_T, the minimum segment length $L_{\min}$, and the per-regime mixing time t_{mix}.

6.1 Preliminaries: Two-Level Benchmarks, Regularity, and Mixing

Recall the regime process is piecewise constant, i.e., there exist switch times $1 = \tau_0 < \tau_1 < \cdots < \tau_{S_T} \le T$ such that the regime is constant on each segment $\mathcal{I}_k \triangleq \{\tau_{k-1}, \ldots, \tau_k - 1\}$, $k = 1, \ldots, S_T + 1$. Let z_k denote the regime on $\mathcal{I}_k$, and let $L_k \triangleq |\mathcal{I}_k|$, so that $L_{\min} = \min_k L_k$. Our tracking regret $\mathrm{Reg}(T)$ in (4) benchmarks against the regime-optimal stationary policy $\pi^{*,(z_t)}$, which may lie outside the MoE expert families. To separate *learnability within the class* from *modeling mismatch*, we introduce an intermediate, in-class comparator. For each expert $m \in [M]$, let $\Pi^{(m)} \triangleq \{\pi_\phi^{(m)} : \phi \in \Phi_m\}$ denote its policy family. Define the *in-class per-regime oracle* (breaking ties arbitrarily) by

$$(m^{\text{ic}}(z), \phi^{\text{ic}}(z)) \in \arg\min_{m\in[M],\phi\in\Phi_m} J^{(z)}\left(\pi_\phi^{(m)}\right) \text{ and } \pi^{\text{ic},(z)} \triangleq \pi_{\phi^{\text{ic}}(z)}^{(m^{\text{ic}}(z))}. \tag{23}$$

We quantify the unavoidable policy-class mismatch by the per-regime approximation gap

$$\mathrm{Approx}_\pi \triangleq \max_{z\in\mathcal{Z}} \left(J^{(z)}\left(\pi^{\text{ic},(z)}\right) - J^{(z)}\left(\pi^{*,(z)}\right)\right) \ge 0. \tag{24}$$

Later, our regret bound naturally decomposes as $\mathrm{Reg}(T) \lesssim$ learning regret for $\pi^{\text{ic},(z_t)} + T\mathrm{Approx}_\pi$, so Approx_π captures the portion that no algorithm can remove without enlarging the expert class.

Regularity assumptions. We use two standard mild technical conditions to control stochastic-gradient magnitudes and to ensure the critics are well-posed.

Assumption 1 (Bounded score functions)**.** For the policy class under consideration, we assume bounded score functions, i.e., there exist constants $G_\pi, G_g < \infty$ such that for all $m \in [M]$,

$$\|\nabla_{\phi_m} \log \pi_{\phi_m}^{(m)}(a \mid s)\| \leq G_\pi \text{ and } \|\nabla_\theta \log g_\theta(m \mid s)\| \leq G_g. \tag{25}$$

Assumption 2 (Critic realizability)**.** Fix $z \in \mathcal{Z}$ and an expert policy $\pi_{\phi_m}^{(m)}$. Let $\mu^{(z,m)}$ be the stationary distribution of the induced Markov chain under $(P^{(z)}, \pi_{\phi_m}^{(m)})$, and define $\Sigma^{(z,m)} \triangleq \mathbb{E}_{s\sim\mu^{(z,m)}}[\psi(s)\psi(s)^\top]$. Assume $\Sigma^{(z,m)} \succeq \lambda_{\min} I$ uniformly over z, m, ϕ_m for some $\lambda_{\min} > 0$. Moreover, the average-cost projected Bellman equation admits a unique fixed point $w^{*,(z,m)}(\phi_m)$ in the linear class, up to an additive constant in the differential value.

Remark 1. Assumption 1 holds for standard softmax/gated-softmax parameterizations with bounded features and parameters, e.g., enforced via projection onto a compact set, ensuring uniformly bounded stochastic gradients. Assumption 2 is a standard identifiability condition. It makes the linear TD normal equations well-conditioned under each (z, m), so the critic target (projected Bellman fixed point) is well-defined and trackable.

Mixing as a regime-wise systems property. The property below formalizes that, within any regime and under any stationary policy, the induced Markov chain forgets its initial condition quickly.

Definition 2 (Uniform geometric mixing within regimes)**.** For every regime $z \in \mathcal{Z}$ and every stationary randomized policy $\pi \in \Pi_{\text{stat}}$, let $P_\pi^{(z)}$ be the policy-induced Markov kernel on $\mathcal{S}$,

$$P_\pi^{(z)}(s' \mid s) \triangleq \sum\nolimits_{a\in\mathcal{A}} \pi(a \mid s)\, P^{(z)}(s' \mid s, a). \tag{26}$$

We say the RS-MDP satisfies uniform geometric mixing within regimes if, for each (z, π), the Markov chain with kernel $P_\pi^{(z)}$ admits a unique stationary distribution $\mu_\pi^{(z)}$, and there exist constants $C_{\text{mix}} \geq 1$ and $\rho \in (0, 1)$, independent of (z, π), such that for all $s \in \mathcal{S}$ and all $t \geq 0$,

$$\mathrm{TV}\left((P_\pi^{(z)})^t(s, \cdot),\ \mu_\pi^{(z)}\right) \leq C_{\text{mix}}\, \rho^t, \tag{27}$$

where $(P_\pi^{(z)})^t$ is the t-step kernel, i.e., the t-fold composition of $P_\pi^{(z)}$.

Remark 2. Geometric mixing is a standard consequence of mild "randomization" and "connectivity" conditions, which are often natural in systems models. A concrete sufficient condition is a uniform Doeblin minorization, i.e., if there exist $\alpha \in (0, 1)$ and $\nu \in \Delta(\mathcal{S})$ such that $P^{(z)}(\cdot \mid s, a) \geq \alpha\nu(\cdot)$ for all z, s, a, then $P_\pi^{(z)}(\cdot \mid s) \geq \alpha\nu(\cdot)$ for all z, π, s. Writing $P_\pi^{(z)} = \alpha\mathbf{1}\nu^\top + (1-\alpha)\widetilde{P}$, one obtains the TV contraction $\mathrm{TV}(\mu P_\pi^{(z)}, \mu' P_\pi^{(z)}) \leq (1-\alpha)\mathrm{TV}(\mu, \mu')$, which yields (27) with $C_{\text{mix}} = 1$ and $\rho = 1-\alpha$. In queueing/wireless models, analogous geometric mixing follows from a standard "drift-to-a-small-set + minorization" argument, i.e., exogenous randomness of arrivals and/or channels provides noise, while a stabilizing and exploratory action floor ensures recurrent visits to a small set, together implying geometric ergodicity of the controlled chain under each frozen (z, π).

The property in Definition 2 indicates that burn-in bias decays geometrically (see Lemma 1).

Lemma 1 (Per-regime mixing and burn-in bias)**.** *Fix a regime z and any stationary policy $\pi \in \Pi_{\text{stat}}$, and let $\mu_\pi^{(z)}$ be the stationary distribution of $P_\pi^{(z)}$. Then, for any measurable $f : \mathcal{S} \to [-1, 1]$, any initial state $s \in \mathcal{S}$, and any $t \geq 1$, we have*

$$\left|\mathbb{E}[f(s_t) \mid s_1 = s] - \mathbb{E}_{s\sim\mu_\pi^{(z)}}[f(s)]\right| \leq 2C_{\text{mix}}\rho^{t-1}. \tag{28}$$

In particular, after $t \geq 1 + \left\lceil \frac{\log(\epsilon/C_{\text{mix}})}{\log \rho} \right\rceil$, the bias is at most 2ϵ.

Lemma 1 shows that within any regime z and under any stationary policy π, the distribution of the state s_t converges geometrically fast to the stationary distribution $\mu_\pi^{(z)}$. Consequently, time samples collected after $t_{\rm mix}$ steps are approximately stationary. This lemma is the technical bridge that allows us to treat each regime segment as "nearly stationary" after a short transient. See Appendix G for the proof.

6.2 Main Results: Regime-Aware Tracking and Stability

This subsection formalizes the main theoretical guarantees for RA-MoE-AC. As motivated by the impossibility results (Section 4), sublinear tracking regret requires both a multi-modal policy class (experts) and an online expert-selection mechanism (gate). Our analysis follows a modular pipeline from mixing, critic, to gating regret, regret decomposition, until the final main tracking bound, plus a separate stability guarantee under safety projection.

Lemma 2 (Tracking of $\bar{c}^{(m)}$ and w_m within a fixed regime)**.** *Consider a segment $\mathcal{I}_k$ of length L_k with fixed regime z_k. Under Assumption 1 and Assumption 2, there exists a burn-in $b = \Theta(t_{\rm mix})$, such that for all $t \in \mathcal{I}_k$ with $t \geq \tau_{k-1} + b$ and all experts $m \in [M]$, the full-information iterates satisfy*

$$\mathbb{E}\left[|\bar{c}^{(m,t)} - J^{(z_k)}(\pi^{(m)}_{\phi_m^{(t)}})|\right] \leq O(\rho_{\max}) + O(\beta_{\max}) + O\left(\sup_{u\in\mathcal{I}_k}\frac{\alpha_u}{\beta_u}\right) + O\left(C_{\rm mix}\rho^b\right), \quad (29)$$

$$\mathbb{E}\left[\|w_m^{(t)} - w^{*,(z_k,m)}(\phi_m^{(t)})\|\right] \leq O(\beta_{\max}) + O\left(\sup_{u\in\mathcal{I}_k}\frac{\alpha_u}{\beta_u}\right) + O\left(C_{\rm mix}\rho^b\right), \quad (30)$$

where $\beta_{\max} \triangleq \sup_{u\in\mathcal{I}_k}\beta_u$, $\rho_{\max} \triangleq \sup_{u\in\mathcal{I}_k}\rho_u$, and constants depend only on $(c_{\max}, \lambda_{\min}, C_{\rm mix}, \rho)$.

Lemma 2 shows that within any regime segment, after a short burn-in, each expert's critic behaves as if it were trained on approximately stationary samples from that regime. The dominant penalty is the standard timescale-separation term $\sup(\alpha_t/\beta_t)$. If the actor moves slowly relative to the critic, value surrogates can track quickly and do not corrupt the gate/actor updates. We provide the proof sketch for Lemma 2 below, and please see Appendix H for the complete proof.

Proof sketch. Fix (z_k, m) and ϕ_m over a short window. Under Definition 2 and Lemma 1, the Markov noise becomes nearly stationary after $b = \Theta(t_{\rm mix})$, so the TD recursion is close to its mean ODE (projected Bellman equation). Assumption 2 gives well-posedness and uniform conditioning, yielding contraction of the mean dynamics. Finite-time stochastic approximation bounds for linear TD with Markovian noise then give an $O(\beta_{\max})$ term, while the slow drift of ϕ_m contributes $O(\sup\alpha/\beta)$. The baseline estimator $\bar{c}^{(m)}$ is a standard Robbins-Monro average-cost estimator, producing an analogous $O(\rho_{\max})$ term. The $O(C_{\rm mix}\rho^b)$ term is exactly the burn-in bias from Lemma 1. □

Lemma 3 (Gate regret against a piecewise-constant in-class selector)**.** *Let $\tilde{g}_t(\cdot)$ be the post-projection distribution. For the comparator sequence $m_t^{\rm ic} \equiv m_k^{\rm ic}$ for $t \in \mathcal{I}_k$, we have*

$$\sum_{t=1}^T \sum_{m=1}^M \tilde{g}_t(m)\ell_t(m) - \sum_{t=1}^T \ell_t(m_t^{\rm ic}) \leq O\left(C\sqrt{T\log M} + CS_T\log M\right). \quad (31)$$

Lemma 3 shows that the gate pays the standard $\sqrt{T\log M}$ learning term, plus a switching overhead proportional to $S_T \log M$ that accounts for re-concentrating after regime changes. This is the precise formalization of the "switching overhead" discussed in design idea D2/D5. We provide the proof sketch for Lemma 3 below, and please see Appendix I for the complete proof.

Proof sketch. Fixed-share Hedge is a standard reduction from switching comparators to a mixture of restarted Hedge instances. One obtains a dynamic-regret bound against piecewise-constant expert sequences with S_T switches, i.e., a $\sqrt{T\log M}$ term from within-segment learning plus an additive $S_T \log M$ term from the share/restart mechanism. Scaling by the loss range C yields (31). □

Lemma 4 (Regret Decomposition). *Let* $\mathrm{Reg}(T)$ *be the tracking regret in* (4). *Under bounded costs, we have*

$$\mathrm{Reg}(T) \leq \kappa_1 \mathrm{Reg}_{\mathrm{gate}}(T) + \mathrm{Reg}_{\mathrm{AC}}(T) + \mathrm{Reg}_{\mathrm{switch}}(T) + T \cdot \mathrm{Approx}_{\pi} + T \cdot \mathrm{Approx}_{V}, \tag{32}$$

where

- $\mathrm{Reg}_{\mathrm{gate}}(T) \triangleq \sum_{t=1}^{T} \sum_m \tilde{g}_t(m)\ell_t(m) - \sum_{t=1}^{T} \ell_t(m_t^{\mathrm{ic}})$ *is controlled by Lemma 3;*
- $\mathrm{Reg}_{\mathrm{AC}}(T)$ *is the within-expert learning error caused by imperfect advantage surrogates and slow actor updates, controlled by Lemma 2 and standard average-cost policy-gradient arguments;*
- $\mathrm{Reg}_{\mathrm{switch}}(T)$ *is the transient cost incurred in the first* $O(t_{\mathrm{mix}})$ *steps after each switch, before the state distribution re-mixes and critic/baseline estimates re-center;*
- Approx_{π} *is the policy-class gap defined in* (24)*;*
- Approx_{V} *is the critic function-approximation bias.*

Lemma 4 pins each term in the final bound to a concrete design component, including gating controls $\mathrm{Reg}_{\mathrm{gate}}$, per-expert critics control $\mathrm{Reg}_{\mathrm{AC}}$, and mixing controls $\mathrm{Reg}_{\mathrm{switch}}$. The impossibility results imply that without a mechanism reducing $\mathrm{Reg}_{\mathrm{gate}}$, i.e., the multi-modal representation and selection, one cannot generally achieve $\mathrm{Reg}(T) = o(T)$. We provide the proof sketch for Lemma 4 below, and please see Appendix J for the complete proof.

Proof sketch. Add and subtract the in-class selector and the regime-optimal benchmark:

$$\mathrm{Reg}(T) = \underbrace{\big(C_T(\pi) - C_T(\text{in-class selector})\big)}_{\text{learn+gate+transients}} + \underbrace{\big(C_T(\text{in-class selector}) - C_T^*\big)}_{\leq\, T\mathrm{Approx}_{\pi}}. \tag{33}$$

The first bracket is controlled by calibrating instantaneous excess cost by TD-losses, bounding gate dynamic regret, bounding critic/baseline tracking error within segments, and charging at most $c_{\max} t_{\mathrm{mix}}$ per switch for burn-in/mixing transients. The critic approximation bias contributes $T\mathrm{Approx}_{V}$. □

Theorem 5 (Tracking regret for RA-MoE-AC). *Under Assumption 1 and Assumption 2, the full-information RA-MoE-AC variant satisfies*

$$\mathrm{Reg}(T) \leq O\left(\kappa_1 C\sqrt{T \log M} + \kappa_1 C S_T \log M\right) + \tilde{O}\left(\sqrt{T}\right) + O\left(c_{\max} S_T t_{\mathrm{mix}}\right) + T(\mathrm{Approx}_{\pi} + \mathrm{Approx}_{V}). \tag{34}$$

In particular, if $S_T = o(T)$ *and* $\mathrm{Approx}_{\pi} + \mathrm{Approx}_{V} = o(1)$*, then* $\mathrm{Reg}(T)/T \to 0$.

The bound in Theorem 5 decomposes into four important parts: (i) regime inference cost $\sqrt{T \log M} + S_T \log M$; (ii) within-regime learning cost that is sublinear in T under timescale separation; (iii) post-switch mixing/transient cost t_{mix} per switch; (iv) irreducible approximation bias from policy classes. This matches the qualitative lessons of Impossibility I that without multiple experts and a gate, the selection term cannot be sublinear in general. We provide the proof sketch for Theorem 5 below, and please see Appendix K for the complete proof.

Proof sketch. Start from Lemma 4. Bound $\mathrm{Reg}_{\mathrm{gate}}(T)$ by Lemma 3. Bound $\mathrm{Reg}_{\mathrm{AC}}(T)$ using Lemma 2 to control baseline/critic tracking error and standard average-cost actor-critic analysis to convert advantage estimation error into cumulative performance loss, yielding a $\tilde{O}(\sqrt{T})$ term under the two-timescale stepsizes. Bound $\mathrm{Reg}_{\mathrm{switch}}(T)$ by charging at most $O(c_{\max} t_{\mathrm{mix}})$ per switch (burn-in until near-stationarity). Add approximation terms $T\mathrm{Approx}_{\pi}$ and $T\mathrm{Approx}_{V}$. □

Remark 3 (selected-expert-only variants). If one updates critics/actors only for the sampled expert and/or uses bandit losses (e.g., EXP3-style importance weighting), then (34) holds with the standard additional factors, either an explicit exploration-floor assumption ensuring each expert is sampled often enough within each segment, or an importance-weighting variance term (typically yielding a $\sqrt{M}$ factor in the gating bound). All other components remain unchanged.

Theorem 6 (Stability under safety projection). *Let $\tilde{g}_t(m_{\text{safe}}) \geq p_{\min} > 0$ for all t. Then, the queue component Q_t is strongly stable. In particular, there exist constants $B < \infty$ and $\epsilon > 0$, such that for the Lyapunov function $L(Q) = \frac{1}{2}\|Q\|_2^2$, we have*

$$\mathbb{E}[L(Q_{t+1}) - L(Q_t) \mid Q_t] \leq B - \epsilon\|Q_t\|_1, \tag{35}$$

and hence

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|Q_t\|_1] \leq \frac{B}{\epsilon}. \tag{36}$$

Theorem 6 operationalizes the key message of Impossibility II. That is, queueing mismatch can be catastrophic, so stability should be enforced by design. Safety projection yields a regime-agnostic stability floor, while Theorem 5 quantifies efficiency loss relative to the regime-aware benchmark within that stable envelope. Please see Appendix L for the complete proof of Theorem 6.

7 Numerical Results

We evaluate *RA-MoE-AC* (Algorithm 1) on the queueing system instantiation in Section 3.4.1 and focus on two questions aligned with our theory, i.e., switch tracking relative to the regime-aware benchmark and stability in queueing instantiations under latent regime mismatch. We compare three methods, RA-MoE-AC, single-expert actor-critic (Single-AC, i.e., the same actor-critic update but with one expert and no regime selection), and Safe-only (always deploy the stabilizing baseline π_{safe}). All methods share the same policy and critic parameterizations when applicable. The only difference is whether the agent can represent and select multiple regime-specialized behaviors.

We report time-average cost $V_T(\pi) = \frac{1}{T}\sum_{t=1}^{T} c_t$, average tracking regret $\text{Reg}(T)/T$ in (4), estimated by simulating the regime-aware benchmark with oracle regime labels, and queueing stability statistics $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|Q_t\|_1$ and $\max_{t\leq T} \|Q_t\|_1$. Due to page limits, all concrete simulation values (horizon, switch schedule, arrivals/channels, feature maps, stepsizes, and $p_{\min}$) are provided in Appendix A. Additional experiments (ablations and scaling studies) are deferred to Appendix B.

7.1 Switch Tracking and Regime Inference

Figure 1 visualizes regime tracking on the queueing instantiation. The left figure plots the smoothed instantaneous cost, and the right figure plots the gate probabilities $\tilde{g}_t(m)$ over experts (vertical dashed lines indicate true switches). RA-MoE-AC exhibits two consistent behaviors across seeds. First, after each switch, the gate rapidly reallocates probability mass from the previously preferred expert to the expert that is compatible with the new segment. Second, this reallocation coincides with a prompt recovery in per-slot cost. In contrast, Single-AC ($M = 1$) cannot express segment-specific behavior. It adapts slowly and incurs elevated cost after switches. Safe-only remains stable but is suboptimal in cost because it does not exploit benign regimes (it pays a persistent conservatism premium). These observations match the mechanism suggested by our analysis.

7.2 Stability and Backlog Behavior in Queueing

Figure 2 plots the queue backlog evolution. Single-AC can suffer sustained mismatch in segments where its learned service allocation is misaligned with the active regime, which yields persistent service deficit and backlog growth. RA-MoE-AC avoids this failure because the gate can switch to the appropriate expert, and because the safety projection enforces a nontrivial minimum usage of a stabilizing expert, which prevents catastrophic excursions even when the gate is temporarily

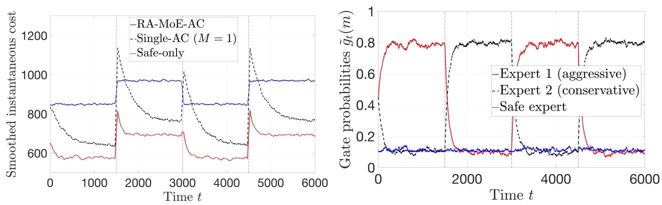

Fig. 1. Switch tracking in the queueing instantiation. Left: smoothed instantaneous cost. Right: gate probabilities for RA-MoE-AC. Vertical dashed lines mark true regime switches. RA-MoE-AC rapidly reallocates probability mass after each switch and correspondingly stabilizes the cost trajectory, while Single-AC exhibits larger and longer post-switch cost transients, and Safe-only incurs a higher cost level.

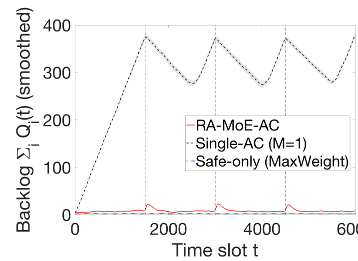

Fig. 2. Backlog and stability in the queueing instantiation. RA-MoE-AC maintains stable backlog with small post-switch transients, while Single-AC can exhibit large backlog excursions under regime mismatch. Safe-only remains stable.

Table 1. Summary metrics (mean ± standard error). RA-MoE-AC improves time-average cost and tracking regret relative to Single-AC, while remaining stability. Safe-only is stable but conservative.

Method	$V_T(\pi)$	$\mathrm{Reg}(T)/T$	$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|Q_t\|_1$	$\max_{t\le T}\|Q_t\|_1$
RA-MoE-AC (Alg. 1)	0.9541 ± 0.0049	0.1081 ± 0.0060	6.89 ± 0.28	33.66 ± 1.38
Single-AC ($M = 1$)	7.5186 ± 0.2532	6.6727 ± 0.2513	340.73 ± 12.66	577.28 ± 20.41
Safe-only (π_{safe})	1.7437 ± 0.0009	0.8977 ± 0.0021	1.25 ± 0.05	15.19 ± 0.67

uncertain. Safe-only remains stable by design, but yields larger average cost because it does not adapt service to the regime-dependent cost tradeoff (Figure 1). Overall, the backlog trajectories provide direct empirical support for our key systems claim.

7.3 Summary metrics

Table 1 summarizes the main numerical outcomes. RA-MoE-AC achieves the best overall efficiency-stability tradeoff. It improves time-average cost relative to Safe-only while preventing backlog blow-ups that can occur under Single-AC. We emphasize that these are *not* tuned-to-win comparisons, since all methods share the same parameterization and step-sizes, and we only vary whether the agent can represent and select multiple regime-specialized experts (and whether safety is enforced).

8 Conclusion

We studied *regime-switching MDPs* for performance-critical systems with latent mode changes. We show that even under benign dynamics and slow switching, any single stationary actor-critic can be misaligned with a regime-aware benchmark, and in queueing instantiations mismatch can break stability. Motivated by this, we proposed *RA-MoE-AC* with TD-residual-driven gating, per-expert critics with timescale separation, and a lightweight safety projection enforcing a stabilizing baseline. We prove switching-aware tracking bounds scaling with the number of switches and mixing time, and strong stability under safety projection. Empirically, RA-MoE-AC re-concentrates after switches, improves cost over conservative baselines, and avoids backlog blow-ups.

References

[1] Jacob Andreas, Dan Klein, and Sergey Levine. 2017. Modular multitask reinforcement learning with policy sketches. In *International conference on machine learning*. PMLR, 166–175.

[2] Peter Auer, Thomas Jaksch, and Ronald Ortner. 2008. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems* 21 (2008).

[3] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. 2017. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*. PMLR, 263–272.

[4] Michele Basseville and Igor V Nikiforov. 1993. *Detection of abrupt changes: theory and application*. Vol. 104. Prentice hall Englewood Cliffs.

[5] Omar Besbes, Yonatan Gur, and Assaf Zeevi. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems* 27 (2014).

[6] Christopher M Bishop and Hugh Bishop. 2023. *Deep learning: Foundations and concepts*. Springer Nature.

[7] Paul Bogdan and Radu Marculescu. 2011. Non-stationary traffic analysis and its implications on multicore platform design. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 30, 4 (2011), 508–519.

[8] Vivek S Borkar and Vivek S Borkar. 2008. *Stochastic approximation: a dynamical systems viewpoint*. Vol. 100. Springer.

[9] Alvaro A Cardenas, Saurabh Amin, and Shankar Sastry. 2008. Secure control: Towards survivable cyber-physical systems. In *2008 The 28th International Conference on Distributed Computing Systems Workshops*. IEEE, 495–500.

[10] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. 2019. Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 1079–1087.

[11] Nicolas Christianson, Junxuan Shen, and Adam Wierman. 2023. Optimal robustness-consistency tradeoffs for learning-augmented metrical task systems. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 9377–9399.

[12] Gregory Ditzler, Manuel Roveri, Cesare Alippi, and Robi Polikar. 2015. Learning in nonstationary environments: A survey. *IEEE Computational intelligence magazine* 10, 4 (2015), 12–25.

[13] Finale Doshi-Velez and George Konidaris. 2016. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In *IJCAI: proceedings of the conference*, Vol. 2016. 1432.

[14] Mark Eisen, Konstantinos Gatsis, George J Pappas, and Alejandro Ribeiro. 2018. Learning statistically accurate resource allocations in non-stationary wireless systems. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3559–3563.

[15] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

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

[17] Leonidas Georgiadis, Michael J Neely, and Leandros Tassiulas. 2006. *Resource allocation and cross-layer control in wireless networks*. Now Publishers Inc.

[18] Mor Harchol-Balter. 2013. *Performance modeling and design of computer systems: queueing theory in action*. Cambridge University Press.

[19] Benjamin Hindman, Andy Konwinski, Matei Zaharia, Ali Ghodsi, Anthony D. Joseph, Randy Katz, Scott Shenker, and Ion Stoica. 2011. Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center. In *Proceedings of the 8th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. USENIX Association, 295–308.

[20] Wenlong Huang, Igor Mordatch, and Deepak Pathak. 2020. One policy to control them all: Shared modular policies for agent-agnostic control. In *International Conference on Machine Learning*. PMLR, 4455–4464.

[21] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87.

[22] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. 2018. Is Q-learning provably efficient? *Advances in neural information processing systems* 31 (2018).

[23] Vijay Konda and John Tsitsiklis. 1999. Actor-critic algorithms. *Advances in neural information processing systems* 12 (1999).

[24] Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. 2021. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems* 34 (2021), 24523–24534.

[25] Xiaojun Lin, NB Shroff, and R Srikant. 2006. A tutorial on cross-layer optimization in wireless networks. *IEEE Journal on Selected Areas in Communications* 24, 8 (2006), 1452–1463.

[26] Zhenhua Liu, Iris Liu, Niangjun Chen, Christina Razon, and Adam Wierman. 2013. Data Center Demand Response: Avoiding the Coincident Peak via Workload Shifting and Local Generation. *Performance Evaluation* 70, 10 (2013), 770–791.

[27] Mohammad Hossein Manshaei, Quanyan Zhu, Tansu Alpcan, Tamer Bacşar, and Jean-Pierre Hubaux. 2013. Game theory meets network security and privacy. *Acm Computing Surveys (Csur)* 45, 3 (2013), 1–39.

[28] Hongzi Mao, Mohammad Alizadeh, Ishai Menache, and Srikanth Kandula. 2016. Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM workshop on hot topics in networks*. 50–56.
[29] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural adaptive video streaming with pensieve. In *Proceedings of the conference of the ACM special interest group on data communication*. 197–210.
[30] Jayakrishnan Nair, Adam Wierman, and Bert Zwart. 2022. *The fundamentals of heavy tails: Properties, emergence, and estimation*. Vol. 53. Cambridge University Press.
[31] Michael Neely. 2010. *Stochastic network optimization with application to communication and queueing systems*. Morgan & Claypool Publishers.
[32] Aldo Pacchiano, Christoph Dann, and Claudio Gentile. 2022. Best of both worlds model selection. *Advances in Neural Information Processing Systems* 35 (2022), 1883–1895.
[33] Ewan S Page. 1954. Continuous inspection schemes. *Biometrika* 41, 1/2 (1954), 100–115.
[34] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).
[35] Ming Shi, Yingbin Liang, and Ness Shroff. [n. d.]. Near-Optimal Adversarial Reinforcement Learning with Switching Costs. In *The Eleventh International Conference on Learning Representations*.
[36] Ming Shi, Yingbin Liang, and Ness Shroff. 2023. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. In *International Conference on Machine Learning*. PMLR, 31243–31268.
[37] Ming Shi, Yingbin Liang, and Ness B Shroff. 2024. Adversarial Online Reinforcement Learning Under Limited Defender Resources. In *Network Security Empowered by Artificial Intelligence*. Springer, 265–301.
[38] Alexander L Stolyar. 2004. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability* 14, 1 (2004), 1–53.
[39] Richard S Sutton, Andrew G Barto, et al. 2018. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.
[40] Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. 2014. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press.
[41] Abhishek Verma, Luis Pedrosa, Madhukar Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. 2015. Large-scale Cluster Management at Google with Borg. In *Proceedings of the Tenth European Conference on Computer Systems (EuroSys)*. ACM.
[42] Ge Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. 2021. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems* 34 (2021), 17084–17097.
[43] Runlong Zhou, Zhang Zihan, and Simon Shaolei Du. 2023. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *International Conference on Machine Learning*. PMLR, 42878–42914.

A Simulation Setup and Hyperparameters

A.1 Common Protocol

Horizon and switching. We run each method for horizon $T = 6000$ with $S_T = 3$ switches and minimum segment length $L_{\min} = 1500$. Switch times are $\tau_0 = 1$, $\tau_1 = 1501$, $\tau_2 = 3001$, $\tau_3 = 4501$, and $\tau_{S_T+1} = T + 1 = 6001$. Unless stated otherwise, we use equal-length segments $L_k = 1500$ and alternate regimes as $z_k = 1$ for $k \in \{1, 3\}$ and $z_k = 2$ for $k \in \{2, 4\}$.

Number of experts and safe expert. We use $M = 3$ experts, where experts $m \in \{1, 2\}$ are learned and expert $m_{\text{safe}} = 3$ is a fixed stabilizing baseline. The safety projection enforces $\tilde{g}_t(m_{\text{safe}}) \geq p_{\min}$ with $p_{\min} = 0.05$.

Seeds and confidence. We use $N_{\text{seed}} = 20$ random seeds and report mean ± one standard error. Time-series figures are smoothed using a moving average window of size 75.

Oracle regime-aware benchmark for regret. To estimate $\text{Reg}(T)$ in (4), we simulate the benchmark that applies $\pi^{*,(z)}$ on each segment using oracle regime labels. For each regime $z \in \{1, 2\}$, we approximate $\pi^{*,(z)}$ by running the same actor–critic update with the regime held fixed at z for 2×10^5 steps (using the same policy/value parameterization as the learned agents), and then deploying the resulting stationary policy whenever $z_t = z$. This yields a fair in-class regime-aware benchmark.

Table 2. Key notation.

Symbol	Meaning
$z_t \in \mathcal{Z}$	latent regime (piecewise constant)
S_T	number of regime switches up to time T
$L_{\min}$	minimum segment length
$s_t \in \mathcal{S}$	system state (e.g., queues + exogenous variables)
$a_t \in \mathcal{A}$	control action
$P^{(z)}$	transition kernel under regime z
$c^{(z)}$	instantaneous cost under regime z
$c_{\max}$	uniform upper bound on per-slot cost
$\pi^{(m)}_{\phi_m}$	expert m actor (policy)
$g_\theta(\cdot \mid s)$	gating distribution over experts
G_π	bound on expert policy score $\|\nabla_{\phi_m} \log \pi^{(m)}_{\phi_m}(a \mid s)\|$
G_g	bound on gate score $\|\nabla_\theta \log g_\theta(m \mid s)\|$
Reg(T)	tracking regret against regime-aware benchmark
t_{mix}	(uniform) mixing time within a regime

A.2 Queueing System Instantiation

Dynamics. We simulate a two-queue single-server system (a standard extension of Section 3.4.1 used to expose regime-mismatch instability). The state is $s_t = (Q_{t,1}, Q_{t,2})$ with $Q_{t,i} \in \mathbb{R}_+$. The action is $a_t \in \{0, 1, 2\}$, where $a_t = 0$ means *idle* and $a_t = i$ means *serve queue* i. Service is one packet when nonempty, i.e.,

$$Q_{t+1,i} = \left[Q_{t,i} + A^{(z_t)}_{t,i} - \mathbb{1}\{a_t = i\}\right]^+, i \in \{1, 2\}. \tag{37}$$

Hence $\mu_{\max} = 1$. Arrivals are independent Bernoulli:

$$A^{(z)}_{t,i} \sim \text{Bernoulli}(\lambda^{(z)}_i), \tag{38}$$

with regime-dependent rates

$$(\lambda^{(1)}_1, \lambda^{(1)}_2) = (0.85, 0.25), (\lambda^{(2)}_1, \lambda^{(2)}_2) = (0.25, 0.85), \tag{39}$$

so that the identity of the bottleneck queue swaps across regimes.

Cost. We use a backlog-energy objective

$$c^{(z)}(Q, a) = w\,(Q_1 + Q_2) + \kappa^{(z)}\,\mathbb{1}\{a \neq 0\}, \tag{40}$$

with $w = 0.01$ and energy-price coefficients

$$\kappa^{(1)} = 0.02, \kappa^{(2)} = 0.20. \tag{41}$$

Thus, serving is cheap in regime 1 and expensive in regime 2, creating different cost-stability tradeoffs across segments.

Policy and critic parameterization. Each learned expert uses a softmax policy over $\{0, 1, 2\}$ with linear action scores, i.e.,

$$\pi^{(m)}_{\phi_m}(a \mid s) \propto \exp\big(u^{(m)}_a(s)\big), u^{(m)}_a(s) = (\phi^{(m)}_a)^\top \varphi_a(s), \tag{42}$$

where $\varphi_a(s) = [1,\ \tilde{Q}_1,\ \tilde{Q}_2]^\top$ and $\tilde{Q}_i = \min\{Q_i/50, 1\}$. Per-expert critics are linear, i.e.,

$$V_{w_m}^{(m)}(s) = w_m^\top \psi(s), \psi(s) = [1,\ \tilde{Q}_1,\ \tilde{Q}_2,\ \tilde{Q}_1^2,\ \tilde{Q}_2^2]^\top. \tag{43}$$

Gate parameterization. We use a softmax gate

$$g_\theta(m \mid s) \propto \exp(\theta_m^\top \varphi_g(s)), \varphi_g(s) = [1,\ \tilde{Q}_1,\ \tilde{Q}_2,\ \tilde{Q}_1 - \tilde{Q}_2]^\top. \tag{44}$$

Safe expert. We embed a stabilizing MaxWeight-style baseline as expert $m_{\text{safe}} = 3$, i.e.,

$$\pi_{\text{safe}} : a_t = \begin{cases} \arg\max_{i \in \{1,2\}} Q_{t,i}, & \text{if } Q_{t,1} + Q_{t,2} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{45}$$

This baseline satisfies the drift condition in Theorem 6 for the simulated primitives, and the safety projection enforces its minimum usage probability.

A.3 RA-MoE-AC Updates and Hyperparameters

TD residual and gate loss. We use the average-cost TD residual $\delta_t^{(m)} = c_t - \bar{c}^{(m)} + V_{w_m}^{(m)}(s_{t+1}) - V_{w_m}^{(m)}(s_t)$ and define the gate loss

$$\ell_t(m) = \text{clip}\big((\delta_t^{(m)})^2,\ 0,\ C\big) \text{ and } C = 10. \tag{46}$$

Step-sizes and timescales. We use constant step-sizes (stable finite-horizon behavior):

$$\beta = 0.05, \eta = 0.02, \alpha = 0.005, \text{ and } \rho = 0.02. \tag{47}$$

We project policy and gate parameters onto ℓ_∞ balls of radius 5 to enforce bounded scores.

B Additional Numerical Results

This appendix reports additional experiments that stress-test RA-MoE-AC beyond the basic comparisons in the main paper. Unless stated otherwise, we use the same environments, horizons, switching patterns, hyperparameters, and reporting protocol as in Appendix A. We report mean ± one standard error over N_{seed} seeds for time-average cost $V_T(\pi)$, tracking regret $\text{Reg}(T)/T$ against the oracle regime-aware benchmark, average backlog $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|Q_t\|_1$, and peak backlog $\max_{t \le T} \|Q_t\|_1$ (queueing system instantiations). When plotting time series, we use the same smoothing window as in Appendix A.

B.1 Ablations: Which Mechanism Matters?

Figure 3 summarizes ablations on the gating signal, timescale separation, and the safety projection.

Gate signal ablation. We isolate the role of the gating feedback by changing only the gate loss $\ell_t(m)$ while keeping experts, critics, step-sizes, and safety projection unchanged. We compare TD-residual losses (ours) to cost-only losses, advantage-only losses, and entropy-regularized cost losses. We report time-to-reconcentrate after each switch (time until $\max_m \tilde{g}_t(m) \ge 0.8$) and post-switch transient cost (area under the excess-cost curve over a fixed window). TD-residual losses consistently yield faster re-concentration and smaller post-switch transients.

Timescale separation ablation. We sweep (α, η, β) to violate the intended ordering (critic fast, gate intermediate, actor slow), changing only step-sizes and keeping all other components fixed. We report switching-transient cost and backlog peaks. When the critic is not the fastest component, value estimates lag the segment-wise fixed point, TD residuals become noisy, and the gate oscillates, increasing both cost and post-switch backlog spikes.

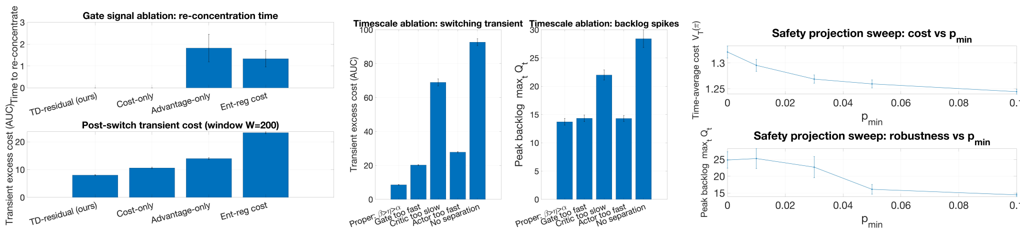

Fig. 3. Ablations. Left: gating loss variants (re-concentration and transient cost). Middle: step-size sweeps showing the role of timescale separation. Right: safety projection sweeps.

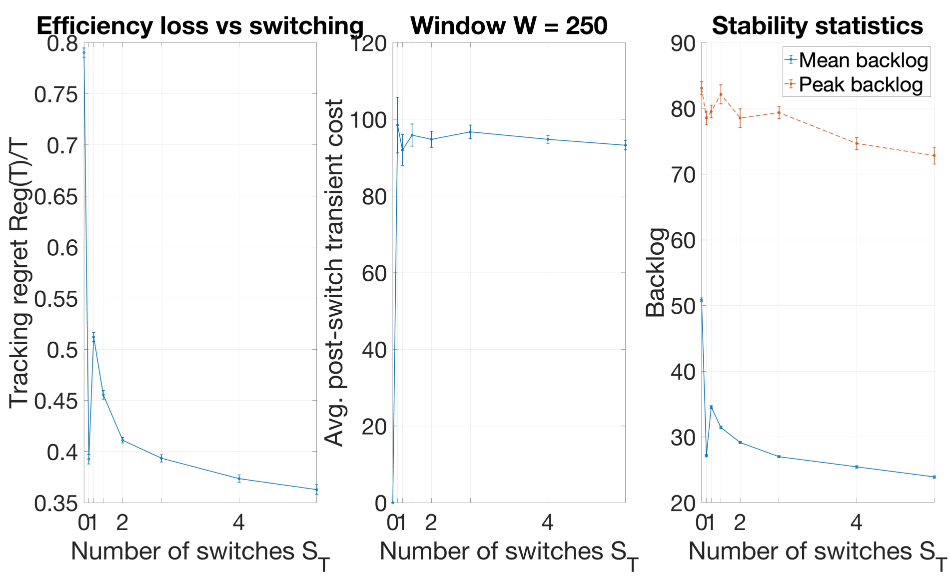

Fig. 4. Switching-frequency scaling. Tracking regret, transient cost, and backlog statistics versus S_T (and induced $L_{\min}$).

Safety projection sweep. We sweep $p_{\min} \in \{0, 0.01, 0.03, 0.05, 0.10\}$ (changing only the projection constraint) to quantify the conservatism–risk tradeoff. Larger $p_{\min}$ improves robustness (smaller peak backlog and fewer rare excursions) at the price of a mild increase in time-average cost. Setting $p_{\min} = 0$ can improve cost in benign segments but may allow catastrophic backlog excursions under prolonged gate uncertainty.

B.2 Scaling with Switching Frequency and Segment Length

We vary the number of switches S_T while holding the horizon T fixed, using equal-length segments (so $L_{\min} = T/(S_T+1)$) and the same alternating-regime pattern as in the main experiments. We report $\mathrm{Reg}(T)/T$, post-switch transient cost (fixed window after each τ_k), and backlog statistics. The trends in Figure 4 match the decomposition suggested by Theorem 5. That is, performance degrades as S_T increases (more frequent re-inference and more transients), and improves as segments lengthen relative to per-regime mixing.

Table 3. Additional baselines (mean ± s.e.). Oracle regime-aware selection and QCD+AC variants.

Method	$V_T(\pi)$	$\mathrm{Reg}(T)/T$	$\frac{1}{T}\sum_t \mathbb{E}\|Q_t\|_1$	$\max_{t\le T}\|Q_t\|_1$
Oracle regime-aware selector	0.6969 ± 0.0009	-0.0024 ± 0.0001	3.2326 ± 0.0275	18.8566 ± 0.5542
QCD+AC (reset on detection)	1.1342 ± 0.0421	0.4350 ± 0.0420	15.7246 ± 0.8973	73.3400 ± 2.1504

B.3 Additional Baselines in Table 3: Oracle and QCD-Style Switching

Oracle regime-aware selector. We include an oracle selector that observes z_t and always chooses the in-class best expert for that regime, i.e., $m_t = \arg\min_m J^{(z_t)}(\pi^{(m)})$ (ties broken arbitrarily). This isolates the performance loss due to imperfect online regime inference. Regret is estimated by Monte Carlo with an in-class regime-aware benchmark. Thus, small negative values may occur due to sampling error of the benchmark.

Quick change detection (QCD) with AC. We include a detect-then-adapt baseline that runs a standard change detector on a scalar stream (we use the selected expert's TD-residual squared loss), and upon detection resets the actor-critic state (critic weights and average-cost baseline) and restarts learning with a fixed exploration floor. This baseline separates explicit detection and reset from continuous online gating.

C Proof of Theorem 1

Proof. Fix an arbitrary stationary randomized policy π. Since π is stationary and the state space is $\{s_0, s_1\}$, define

$$p \triangleq \pi(a^+ \mid s_0) \in [0,1] \text{ and } \pi(a^- \mid s_0) = 1 - p.$$

We do not need to define $\pi(\cdot \mid s_1)$ because costs at s_1 are always zero.

C.1 Step 1: The state s_0 is visited exactly $\lfloor T/2 \rfloor$ times

By construction, the dynamics are deterministic and regime-independent: $s_{t+1} = s_1$ if $s_t = s_0$ and $s_{t+1} = s_0$ if $s_t = s_1$. Thus the chain alternates between the two states, regardless of the action choices. Assuming s_1 is the unique successor of s_0 and vice versa, the trajectory satisfies

$$s_1, s_0, s_1, s_0, \dots \quad \text{or} \quad s_0, s_1, s_0, s_1, \dots$$

depending only on the initial state. In either case, among the first T time indices, the number of visits to s_0 is exactly $\lfloor T/2 \rfloor$. Denote this number by

$$N_0(T) \triangleq \sum_{t=1}^{T} \mathbb{1}\{s_t = s_0\} = \left\lfloor \frac{T}{2} \right\rfloor. \tag{48}$$

C.2 Step 2: Expected cumulative cost under a fixed regime

Consider first the case where the regime is constant over the entire horizon, i.e., $z_t \equiv z \in \{1, 2\}$. Because costs at s_1 are always 0, only visits to s_0 contribute.

C.2.1 In regime 1. At state s_0, action a^+ costs 0 and action a^- costs 1. Therefore, conditioned on $s_t = s_0$,

$$\mathbb{E}\left[c^{(1)}(s_t, a_t) \mid s_t = s_0\right] = 0 \cdot \pi(a^+ \mid s_0) + 1 \cdot \pi(a^- \mid s_0) = 1 - p. \tag{49}$$

C.2.2 In regime 2. The roles swap, and conditioned on $s_t = s_0$,

$$\mathbb{E}\left[c^{(2)}(s_t, a_t) \mid s_t = s_0\right] = 1 \cdot \pi(a^+ \mid s_0) + 0 \cdot \pi(a^- \mid s_0) = p. \tag{50}$$

Since the event $\{s_t = s_0\}$ is deterministic given t (Step 1), by linearity of expectation we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} c^{(1)}(s_t, a_t)\right] &= \sum_{t=1}^{T} \mathbb{E}\left[c^{(1)}(s_t, a_t)\right] = \sum_{t=1}^{T} \mathbb{1}\{s_t = s_0\}(1-p) = N_0(T)(1-p), \\
\mathbb{E}\left[\sum_{t=1}^{T} c^{(2)}(s_t, a_t)\right] &= \sum_{t=1}^{T} \mathbb{1}\{s_t = s_0\} p = N_0(T) p.
\end{aligned} \tag{51}$$

C.3 Step 3: Benchmark cost is identically zero

By definition, the regime-aware benchmark π_t^* chooses the zero-cost action at s_0 for the active regime: it chooses a^+ when $z_t = 1$ and a^- when $z_t = 2$. At s_1, both actions have cost 0. Hence, for every time t, the incurred cost under π_t^* is 0, and therefore

$$C_T^* \triangleq \mathbb{E}\left[\sum_{t=1}^{T} c^{(z_t)}(s_t^{\pi^*}, a_t^{\pi^*})\right] = 0, \tag{52}$$

for any regime sequence $\{z_t\}$.

C.4 Step 4: Choose a piecewise-constant regime sequence that maximizes the loss of π

Define a constant (hence piecewise-constant) regime sequence as follows:

$$z_t \equiv \begin{cases} 1, & \text{if } 1-p \geq p, \\ 2, & \text{if } p > 1-p. \end{cases}$$

Equivalently, choose the regime z that makes π's expected cost at s_0 equal to $\max\{p, 1-p\}$.

Under this regime sequence, combining Steps 2 and 3,

$$\mathrm{Reg}(T) = C_T(\pi) - C_T^* = C_T(\pi) \geq N_0(T) \max\{p, 1-p\} = \left\lfloor \frac{T}{2} \right\rfloor \max\{p, 1-p\}. \tag{53}$$

Finally, since $\max\{p, 1-p\} \geq \frac{1}{2}$ for all $p \in [0,1]$, we obtain the universal bound

$$\mathrm{Reg}(T) \geq \left\lfloor \frac{T}{2} \right\rfloor \cdot \frac{1}{2} \geq \frac{T}{4} - \frac{1}{2}. \tag{54}$$

This implies $\mathrm{Reg}(T) = \Omega(T)$ and $\mathrm{Reg}(T)/T = \Omega(1)$.

□

D Proof of Theorem 2

Proof. Fix any stationary randomized policy π and any $L_{\min} \geq 1$. Let $p = \pi(a^+ \mid s_0)$. As in the proof of Theorem 1, the regime-aware benchmark incurs zero cost at all times, so for any admissible regime sequence $\{z_t\}$,

$$\mathrm{Reg}(T) = C_T(\pi) - C_T^* = C_T(\pi). \tag{55}$$

We show that there exists an admissible piecewise-constant regime sequence (with minimum segment length $L_{\min}$) under which $C_T(\pi)$ grows linearly in T.

D.1 Case (i): no switching

Choose $z_t \equiv z$ constant for all $t = 1, \dots, T$, where $z \in \{1, 2\}$ is selected as in Theorem 1 to maximize π's expected cost at s_0. This regime sequence has $S_T = 0$ and a single segment of length T, hence it satisfies the minimum segment length constraint for any $L_{\min} \le T$. By Theorem 1,

$$\mathrm{Reg}(T) = C_T(\pi) \ge \left\lfloor \frac{T}{2} \right\rfloor \max\{p, 1-p\} \ge \frac{T}{4} - \frac{1}{2} = \Omega(T). \tag{56}$$

D.2 Case (ii): switching with segment length exactly $L_{\min}$

Assume T is a multiple of $2L_{\min}$ (as stated in (16)) and set $K \triangleq \frac{T}{L_{\min}}$ as an even integer. Partition the horizon into K consecutive segments of length $L_{\min}$:

$$\mathcal{I}_k \triangleq \{(k-1)L_{\min} + 1, \dots, kL_{\min}\}, \text{ for } k = 1, \dots, K. \tag{57}$$

Define a piecewise-constant regime sequence by alternating regimes:

$$z_t = \begin{cases} 1, & t \in \mathcal{I}_k \text{ with } k \text{ odd,} \\ 2, & t \in \mathcal{I}_k \text{ with } k \text{ even.} \end{cases} \tag{58}$$

Every segment has length exactly $L_{\min}$, so the minimum segment length constraint is satisfied.

D.2.1 Step 1: within each segment, s_0 is visited at least $\lfloor L_{\min}/2 \rfloor$ times. Because the state alternates deterministically between s_0 and s_1 at every time step, in any consecutive block of length $L_{\min}$, the number of indices t with $s_t = s_0$ is either $\lfloor L_{\min}/2 \rfloor$ or $\lceil L_{\min}/2 \rceil$, depending on the parity of the block start. In particular, it is always at least $\lfloor L_{\min}/2 \rfloor$. We define $N_{0,k} \triangleq \sum_{t \in \mathcal{I}_k} \mathbb{1}\{s_t = s_0\}$, then we have

$$N_{0,k} \ge \left\lfloor \frac{L_{\min}}{2} \right\rfloor, \text{ for all } k = 1, \dots, K. \tag{59}$$

D.2.2 Step 2: expected cost per segment. In a segment with regime 1, each visit to s_0 contributes expected cost $1 - p$; in a segment with regime 2, each visit contributes expected cost p. Costs at s_1 are always 0. Therefore,

$$\mathbb{E}\left[\sum_{t \in \mathcal{I}_k} c^{(z_t)}(s_t, a_t)\right] = \begin{cases} (1-p)\, N_{0,k}, & k \text{ odd,} \\ p\, N_{0,k}, & k \text{ even.} \end{cases} \tag{60}$$

D.2.3 Step 3: sum over alternating segments. Because K is even, there are exactly $K/2$ odd segments and $K/2$ even segments. Summing the lower bound $N_{0,k} \ge \lfloor L_{\min}/2 \rfloor$ and using $(1-p) + p = 1$, we get

$$\mathrm{Reg}(T) = C_T(\pi) = \mathbb{E}\left[\sum_{k=1}^{K} \sum_{t \in \mathcal{I}_k} c^{(z_t)}(s_t, a_t)\right] = \sum_{k=1}^{K} \mathbb{E}\left[\sum_{t \in \mathcal{I}_k} c^{(z_t)}(s_t, a_t)\right] \tag{61}$$

$$\ge \sum_{\substack{k \le K \\ k \text{ odd}}} (1-p) \left\lfloor \frac{L_{\min}}{2} \right\rfloor + \sum_{\substack{k \le K \\ k \text{ even}}} p \left\lfloor \frac{L_{\min}}{2} \right\rfloor \tag{62}$$

$$= \frac{K}{2} \left\lfloor \frac{L_{\min}}{2} \right\rfloor \cdot \big((1-p) + p\big) = \frac{K}{2} \left\lfloor \frac{L_{\min}}{2} \right\rfloor. \tag{63}$$

Substituting $K = T/L_{\min}$ yields

$$\mathrm{Reg}(T) \ge \frac{T}{2L_{\min}} \left\lfloor \frac{L_{\min}}{2} \right\rfloor. \tag{64}$$

Using the elementary bound $\lfloor x \rfloor \geq x - 1$ with $x = L_{\min}/2$, we obtain $\lfloor \frac{L_{\min}}{2} \rfloor \geq \frac{L_{\min}}{2} - 1$, and therefore

$$\mathrm{Reg}(T) \geq \frac{T}{2L_{\min}} \left(\frac{L_{\min}}{2} - 1 \right) = \left(\frac{1}{4} - \frac{1}{2L_{\min}} \right) T, \tag{65}$$

which matches (16).

Combining cases (i) and (ii), we conclude that for any $L_{\min} \geq 1$, there exists an admissible piecewise-constant regime sequence satisfying the minimum segment length constraint such that $\mathrm{Reg}(T) = \Omega(T)$.

□

E Proof of Theorem 3

We consider a two-queue, single-server, discrete-time system. Let $Q_{t,i} \in \mathbb{Z}_+$ denote the backlog of queue $i \in \{1,2\}$ at the beginning of slot t. In each slot, the controller chooses an action $a_t \in \{1,2\}$. If $Q_{t,a_t} > 0$, one packet departs from queue a_t; otherwise the service opportunity is wasted (the server idles). Let $A_{t,i}^{(z_t)} \in \{0,1\}$ denote arrivals to queue i during slot t under regime z_t. The queue dynamics are

$$Q_{t+1,i} = Q_{t,i} - D_{t,i} + A_{t,i}^{(z_t)} \text{ and } D_{t,i} \triangleq \mathbb{1}\{a_t = i\}\mathbb{1}\{Q_{t,i} > 0\}, \text{ for } i \in \{1,2\}. \tag{66}$$

We construct arrivals as i.i.d. Bernoulli random variables across time and queues conditioned on the regime, with means

$$\mathbb{E}[A_{t,1}^{(1)}] = \lambda_H, \mathbb{E}[A_{t,2}^{(1)}] = \lambda_L; \text{ and } \mathbb{E}[A_{t,1}^{(2)}] = \lambda_L, \mathbb{E}[A_{t,2}^{(2)}] = \lambda_H, \tag{67}$$

where $\lambda_H \in (1/2, 1)$ and $\lambda_L \in (0, 1 - \lambda_H)$. This ensures the system is stabilizable in each fixed regime.

Fixed randomized priority policies. A stationary randomized priority policy with parameter $p \in [0,1]$ is defined by

$$\Pr(a_t = 1) = p \text{ and } \Pr(a_t = 2) = 1 - p, \text{ for all } t, \tag{68}$$

independently of the state/history (thus it may waste service when the selected queue is empty).

Proof. We prove the two items in Theorem 3 by explicit construction.

E.1 Item (1): per-regime stabilizability

Fix a regime $z \in \{1,2\}$. Let (λ_1, λ_2) denote the arrival means under that regime; by (67), $(\lambda_1, \lambda_2) = (\lambda_H, \lambda_L)$ if $z = 1$ and $(\lambda_1, \lambda_2) = (\lambda_L, \lambda_H)$ if $z = 2$. By assumption, $\lambda_H + \lambda_L < 1$.

Consider the following *work-conserving* stationary policy $\pi^{\mathrm{wc},(z)}$:

- if exactly one queue is nonempty, serve the nonempty queue;
- if both queues are nonempty, serve queue i with probability α_i, where $\alpha_1, \alpha_2 > 0$, $\alpha_1 + \alpha_2 = 1$, and $\alpha_i > \lambda_i$ for both $i = 1, 2$.

Such (α_1, α_2) exist because $\lambda_1 + \lambda_2 < 1$; for example, choose $\alpha_i = \lambda_i + \frac{1-(\lambda_1+\lambda_2)}{2}$.

Let $V_t \triangleq Q_{t,1} + Q_{t,2}$. Under a work-conserving policy, whenever $V_t > 0$ the server necessarily serves one packet from a nonempty queue, so the total departure satisfies $D_{t,1} + D_{t,2} = 1$. When $V_t = 0$, we have $D_{t,1} + D_{t,2} = 0$. Summing (66) over $i \in \{1,2\}$ gives

$$V_{t+1} - V_t = (A_{t,1}^{(z)} + A_{t,2}^{(z)}) - (D_{t,1} + D_{t,2}) = (A_{t,1}^{(z)} + A_{t,2}^{(z)}) - \mathbb{1}\{V_t > 0\}. \tag{69}$$

Taking conditional expectation given V_t yields

$$\mathbb{E}[V_{t+1} - V_t \mid V_t] = (\lambda_1 + \lambda_2) - \mathbb{1}\{V_t > 0\} \leq -\epsilon \cdot \mathbb{1}\{V_t > 0\}, \tag{70}$$

where $\epsilon \triangleq 1 - (\lambda_1 + \lambda_2) > 0$. Thus, for all states with $V_t > 0$, the drift of V_t is uniformly negative by at least ϵ. By a standard Foster–Lyapunov criterion for countable-state Markov chains (using V as a Lyapunov function), this implies positive recurrence of $\{(Q_{t,1}, Q_{t,2})\}$ under regime z and hence strong stability (13). Therefore, each fixed regime is stabilizable by a stationary policy.

E.2 Item (2): impossibility for fixed randomized priorities under switching

Fix an arbitrary $p \in [0, 1]$ and consider the fixed priority policy (68). We show that there exists a piecewise-constant regime sequence for which the system is not strongly stable. Define the "bad" regime for this p as

$$z_{\text{bad}}(p) \triangleq \begin{cases} 1, & \text{if } p < \lambda_H, \\ 2, & \text{otherwise.} \end{cases} \tag{71}$$

This choice is always valid because if $p \geq \lambda_H$, then $1 - p \leq 1 - \lambda_H < \lambda_H$ (since $\lambda_H > 1/2$), so regime 2 makes queue 2 heavy with arrival λ_H but service attempt probability $1 - p < \lambda_H$.

Now consider the (piecewise-constant) regime sequence $z_t \equiv z_{\text{bad}}(p)$ for all t. Let i^* be the heavy queue under that regime: $i^* = 1$ if $z_{\text{bad}}(p) = 1$ and $i^* = 2$ if $z_{\text{bad}}(p) = 2$. Under the fixed-p policy, the action attempt probability for serving queue i^* is

$$s(p) \triangleq \begin{cases} p, & i^* = 1, \\ 1 - p, & i^* = 2. \end{cases} \tag{72}$$

By construction, $s(p) < \lambda_H$.

From the queue update (66), we always have the inequality

$$Q_{t+1,i^*} \geq Q_{t,i^*} - \mathbb{1}\{a_t = i^*\} + A_{t,i^*}^{(z_{\text{bad}}(p))}. \tag{73}$$

Indeed, when $Q_{t,i^*} > 0$, the departure is exactly $\mathbb{1}\{a_t = i^*\}$; when $Q_{t,i^*} = 0$, the true departure is 0 and thus subtracting $\mathbb{1}\{a_t = i^*\}$ only makes the right-hand side smaller, so the inequality holds.

Taking expectation of (73) conditional on Q_{t,i^*} and using $\mathbb{E}[\mathbb{1}\{a_t = i^*\}] = s(p)$ and $\mathbb{E}[A_{t,i^*}^{(z_{\text{bad}}(p))}] = \lambda_H$, we obtain

$$\mathbb{E}[Q_{t+1,i^*} \mid Q_{t,i^*}] \geq Q_{t,i^*} + (\lambda_H - s(p)). \tag{74}$$

Taking total expectation and iterating yields, for all $t \geq 1$,

$$\mathbb{E}[Q_{t,i^*}] \geq \mathbb{E}[Q_{1,i^*}] + (t-1)(\lambda_H - s(p)). \tag{75}$$

Since $\lambda_H - s(p) > 0$, $\mathbb{E}[Q_{t,i^*}]$ grows at least linearly in t.

Finally, strong stability (13) fails because

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Q_{t,1} + Q_{t,2}] \geq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Q_{t,i^*}] \overset{(75)}{\geq} \frac{1}{T}\sum_{t=1}^{T}(t-1)(\lambda_H - s(p)) = \frac{(T-1)}{2}(\lambda_H - s(p)) \xrightarrow[T\to\infty]{} \infty. \tag{76}$$

Therefore, the fixed randomized priority policy with parameter p is not strongly stable under the (piecewise-constant) regime sequence $z_t \equiv z_{\text{bad}}(p)$. This proves item (2) and completes the proof. □

F Proof of Theorem 4

Proof. Fix $p \in [0, 1]$ and consider the fixed randomized priority policy (68). Define

$$s_{\min}(p) \triangleq \min\{p, 1-p\} \in [0, 1/2] \text{ and } \delta(p) \triangleq \lambda_H - s_{\min}(p). \tag{77}$$

Since $\lambda_H > 1/2$, we have $\delta(p) \geq \lambda_H - \frac{1}{2} > 0$.

F.1 Step 1: construct a regime sequence with minimum segment length $L_{\min}$

Define the "bad" regime as the regime in which the heavy queue is the *less frequently attempted* queue under the fixed-p policy:

$$z_{\text{bad}} \triangleq \begin{cases} 1, & \text{if } p \le 1/2 \quad (\text{queue 1 is attempted with prob. } p = s_{\min}(p)), \\ 2, & \text{if } p > 1/2 \quad (\text{queue 2 is attempted with prob. } 1-p = s_{\min}(p)). \end{cases} \tag{78}$$

Let $i^* \in \{1,2\}$ denote the heavy queue under regime z_{bad} (so $i^* = 1$ if $z_{\text{bad}} = 1$ and $i^* = 2$ if $z_{\text{bad}} = 2$). Under the fixed-p policy, the attempt probability to serve queue i^* equals $s_{\min}(p)$.

Now define a piecewise-constant regime process by alternating regimes in segments of length exactly $L_{\min}$:

$$z_t = \begin{cases} z_{\text{bad}}, & t \in \{(2k-2)L_{\min}+1, \dots, (2k-1)L_{\min}\}, \\ 3 - z_{\text{bad}}, & t \in \{(2k-1)L_{\min}+1, \dots, 2kL_{\min}\}, \end{cases} \qquad k = 1, 2, \dots \tag{79}$$

Every segment has length $L_{\min}$, hence the minimum segment length constraint is satisfied. The endpoints of the *bad-regime segments* occur at times $t_k \triangleq (2k-1)L_{\min}$, $k = 1, 2, \dots$, which are infinitely many regime-segment endpoints.

F.2 Step 2: backlog increase over any bad segment is linear in $L_{\min}$

Fix any bad segment $[t_0, t_0 + L_{\min} - 1]$ in which $z_t = z_{\text{bad}}$ for all $t \in [t_0, t_0 + L_{\min} - 1]$. For the heavy queue i^*, the same inequality as in (73) holds for every slot in this segment:

$$Q_{t+1,i^*} \;\ge\; Q_{t,i^*} - \mathbb{1}\{a_t = i^*\} + A_{t,i^*}^{(z_{\text{bad}})}. \tag{80}$$

Summing from $t = t_0$ to $t_0 + L_{\min} - 1$ and telescoping gives

$$Q_{t_0+L_{\min},i^*} - Q_{t_0,i^*} \ge \sum_{t=t_0}^{t_0+L_{\min}-1} A_{t,i^*}^{(z_{\text{bad}})} - \sum_{t=t_0}^{t_0+L_{\min}-1} \mathbb{1}\{a_t = i^*\}. \tag{81}$$

Taking expectation and using $\mathbb{E}[A_{t,i^*}^{(z_{\text{bad}})}] = \lambda_H$ and $\mathbb{E}[\mathbb{1}\{a_t = i^*\}] = s_{\min}(p)$, we obtain

$$\mathbb{E}\big[Q_{t_0+L_{\min},i^*} - Q_{t_0,i^*}\big] \ge (\lambda_H - s_{\min}(p))L_{\min} = \delta(p)L_{\min}. \tag{82}$$

Since $\mathbb{E}[Q_{t_0,i^*}] \ge 0$, (82) implies

$$\mathbb{E}[Q_{t_0+L_{\min},i^*}] \ge \delta(p)L_{\min}. \tag{83}$$

F.3 Step 3: infinitely many regime-segment endpoints

Apply (83) to each bad-regime segment. In the constructed regime sequence, the endpoint of the k-th bad segment is $t_k = (2k-1)L_{\min}$. Thus, for all $k \ge 1$,

$$\mathbb{E}[Q_{t_k,i^*}] \ge \delta(p)L_{\min}. \tag{84}$$

Therefore,

$$\mathbb{E}[Q_{t_k,1} + Q_{t_k,2}] \ge \mathbb{E}[Q_{t_k,i^*}] \ge \delta(p)L_{\min}, \tag{85}$$

for infinitely many regime-segment endpoints t_k. This proves the claimed $\Theta(L_{\min})$ lower bound (with an explicit constant $\delta(p) > 0$).

The above argument yields a clean bound without an $O(1)$ slack because we used the inequality $Q_{t+1,i^*} \ge Q_{t,i^*} - \mathbb{1}\{a_t = i^*\} + A_{t,i^*}$, which holds even when the queue is empty and the service attempt is wasted.

□

G Proof of Lemma 1

Proof. Fix a regime z and a stationary policy π. To simplify notation, write $P \equiv P_\pi^{(z)}$ and $\mu \equiv \mu_\pi^{(z)}$.

G.1 Step 1: Express the distribution of s_t via the $t-1$-step kernel

Conditioned on $s_1 = s$, the distribution of s_t is given by the $(t-1)$-step transition kernel:

$$\mathcal{L}(s_t \mid s_1 = s) = P^{t-1}(s, \cdot), \tag{86}$$

because P^{t-1} is the $(t-1)$-fold composition of the one-step kernel P. Therefore,

$$\mathbb{E}[f(s_t) \mid s_1 = s] - \mathbb{E}_{x\sim\mu}[f(x)] = \int_{\mathcal{S}} f(x) P^{t-1}(s, dx) - \int_{\mathcal{S}} f(x)\mu(dx). \tag{87}$$

G.2 Step 2: Use the dual characterization of total variation

Define the signed measure $\nu \triangleq P^{t-1}(s, \cdot) - \mu$. Then, (87) becomes

$$\mathbb{E}[f(s_t) \mid s_1 = s] - \mathbb{E}_{x\sim\mu}[f(x)] = \int_{\mathcal{S}} f(x)\, \nu(dx). \tag{88}$$

Recall the standard inequality (a consequence of the definition of total variation): for any measurable f with $\|f\|_\infty \leq 1$,

$$\left| \int_{\mathcal{S}} f(x)\, \nu(dx) \right| \leq 2\mathrm{TV}\left(P^{t-1}(s, \cdot), \mu\right). \tag{89}$$

For completeness, we justify (89): for any two probability measures α, β on $\mathcal{S}$,

$$\mathrm{TV}(\alpha, \beta) \triangleq \sup_{A \subseteq \mathcal{S}} |\alpha(A) - \beta(A)|, \tag{90}$$

and one equivalent dual form is

$$\mathrm{TV}(\alpha, \beta) = \frac{1}{2} \sup_{\|g\|_\infty \leq 1} \left| \int g\, d(\alpha - \beta) \right|. \tag{91}$$

Applying this with $\alpha = P^{t-1}(s, \cdot)$, $\beta = \mu$, and $g = f$ gives (89).

G.3 Step 3: Apply the geometric mixing assumption

By (27) with $t-1$ in place of t, we have

$$\mathrm{TV}\left(P^{t-1}(s, \cdot), \mu\right) \leq C_{\mathrm{mix}} \rho^{t-1}. \tag{92}$$

Combining with (89) yields

$$\left| \mathbb{E}[f(s_t) \mid s_1 = s] - \mathbb{E}_{x\sim\mu}[f(x)] \right| \leq 2\, C_{\mathrm{mix}}\, \rho^{t-1}, \tag{93}$$

which is exactly (28).

G.4 Step 4: The ϵ-burn-in result

If $C_{\mathrm{mix}} \rho^{t-1} \leq \epsilon$, then (28) implies the bias is at most 2ϵ. This completes the proof.

□

H Proof of Lemma 2

Proof. Fix a segment $\mathcal{I}_k = \{\tau_{k-1}, \ldots, \tau_k - 1\}$ on which the regime is constant and equal to $z_k \equiv z$. Fix an expert $m \in [M]$ and, within this proof, suppress the superscript (m) when no confusion arises. Let $\mathcal{F}_t$ be the natural filtration generated by all randomness up to time t.

H.1 Objects, fixed points, and the within-segment "frozen-ϕ" viewpoint

For each time $t \in \mathcal{I}_k$, denote the expert policy by $\pi_t(\cdot \mid s) \equiv \pi^{(m)}_{\phi_m^{(t)}}(\cdot \mid s)$ and the induced Markov kernel on $\mathcal{S}$ (under regime z) by

$$P_t(s' \mid s) \triangleq \sum_{a \in \mathcal{A}} \pi_t(a \mid s)\, P^{(z)}(s' \mid s, a). \tag{94}$$

Let μ_t be the stationary distribution of P_t (existence/uniqueness is guaranteed by Definition 2). Define the steady-state average cost of π_t under regime z by

$$J_t \triangleq J^{(z)}(\pi_t) = \mathbb{E}_{s \sim \mu_t, a \sim \pi_t(\cdot|s)} \left[c^{(z)}(s, a)\right]. \tag{95}$$

For the critic, recall the linear differential-value approximation $V_w(s) = \psi(s)^\top w$. Under regime z and policy π_t, define the (average-cost) TD feature matrix and vector:

$$A_t^* \triangleq \mathbb{E}_{(s,a,s') \sim \mu_t \times \pi_t \times P^{(z)}} \left[\psi(s)\left(\psi(s) - \psi(s')\right)^\top\right], \tag{96}$$

$$b_t^* \triangleq \mathbb{E}_{(s,a) \sim \mu_t \times \pi_t} \left[(J_t - c^{(z)}(s, a))\psi(s)\right]. \tag{97}$$

Assumption 2 (realizability) implies that the projected average-cost Bellman equation has a unique solution w_t^* (up to an additive constant in the differential value; the linear parameter w_t^* is unique under the standard convention that removes this constant), equivalently

$$A_t^* w_t^* = b_t^*. \tag{98}$$

Moreover, Assumption 2 yields uniform conditioning, i.e., there exists $\lambda_A > 0$ (depending only on $\lambda_{\min}$ and boundedness of ψ) such that

$$\sigma_{\min}(A_t^*) \geq \lambda_A \text{ for all } t \in \mathcal{I}_k, \tag{99}$$

and hence $\|(A_t^*)^{-1}\| \leq 1/\lambda_A$ uniformly.

H.2 Full-information critic/baseline updates analyzed in this lemma

Within the segment, we analyze the standard full-information (per-expert) average-cost TD(0) recursions:

$$\bar{c}^{(t+1)} = (1 - \rho_t)\bar{c}^{(t)} + \rho_t c_t, \tag{100}$$

$$w^{(t+1)} = w^{(t)} - \beta_t \delta_t \psi(s_t), \tag{101}$$

where $c_t = c^{(z)}(s_t, a_t)$ and the TD residual is

$$\delta_t \triangleq c_t - \bar{c}^{(t)} + \psi(s_{t+1})^\top w^{(t)} - \psi(s_t)^\top w^{(t)}. \tag{102}$$

(The statement of the lemma concerns the full-information iterates; this recursion is exactly the per-expert recursion used in the analysis. All constants below are uniform in (z, m).)

H.3 Burn-in and a clean bias bound after mixing

Fix $b \geq 1$ (to be chosen as $b = \Theta(t_{\text{mix}})$). By Lemma 1, for any bounded measurable $f : \mathcal{S} \to [-1, 1]$ and any $t \geq \tau_{k-1} + b$,

$$\left|\mathbb{E}[f(s_t) \mid \mathcal{F}_{\tau_{k-1}}] - \mathbb{E}_{s \sim \mu_{\tau_{k-1}}}[f(s)]\right| \leq 2C_{\text{mix}}\rho^{b-1}. \tag{103}$$

In particular, for any bounded $g(s, a, s') \in [-1, 1]$, the same bound holds for g evaluated along one-step transitions by applying (103) to the Markov chain on the augmented state space (or by conditioning on s_t and using boundedness); we will use this informally as

$$\left|\mathbb{E}[g(s_t, a_t, s_{t+1})] - \mathbb{E}_{(s,a,s')\sim\mu_t\times\pi_t\times P^{(z)}}[g(s,a,s')]\right| \le O(C_{\text{mix}}\rho^b). \tag{104}$$

H.4 Tracking of the average-cost baseline $\bar{c}^{(t)}$

Define the baseline error

$$e_t \triangleq \bar{c}^{(t)} - J_t. \tag{105}$$

From (100), we have

$$\begin{aligned} e_{t+1} &= \bar{c}^{(t+1)} - J_{t+1} \\ &= (1-\rho_t)\bar{c}^{(t)} + \rho_t c_t - J_{t+1} \\ &= (1-\rho_t)(\bar{c}^{(t)} - J_t) + \rho_t(c_t - J_t) + (1-\rho_t)(J_t - J_{t+1}). \end{aligned} \tag{106}$$

Taking absolute values and expectations yields

$$\mathbb{E}[|e_{t+1}|] \le (1-\rho_t)\mathbb{E}[|e_t|] + \rho_t\mathbb{E}[|c_t - J_t|] + (1-\rho_t)\mathbb{E}[|J_{t+1} - J_t|]. \tag{107}$$

Since $0 \le c_t \le c_{\max}$, we have $|c_t - J_t| \le c_{\max}$ and thus

$$\rho_t\mathbb{E}[|c_t - J_t|] \le \rho_t c_{\max} \le c_{\max}\rho_{\max}. \tag{108}$$

Because the policy parameters move on the slow timescale α_t and the policy-score is bounded (Assumption 1), standard smoothness/perturbation arguments under uniform mixing imply that $J^{(z)}(\pi_\phi^{(m)})$ is Lipschitz in ϕ over the (compact) parameter set used in the analysis, i.e., there exists $L_J < \infty$ (depending only on $(c_{\max}, C_{\text{mix}}, \rho, G_\pi)$) such that

$$|J_{t+1} - J_t| \le L_J\|\phi_m^{(t+1)} - \phi_m^{(t)}\|. \tag{109}$$

Moreover, the actor update magnitude is uniformly bounded: $\|\phi_m^{(t+1)} - \phi_m^{(t)}\| \le O(\alpha_t)$ (bounded score and bounded advantage/TD signal under clipping/bounded costs). Therefore

$$\mathbb{E}[|J_{t+1} - J_t|] \le O(\alpha_t). \tag{110}$$

Plugging (108)–(110) into (107) gives, for all $t \in \mathcal{I}_k$, we have

$$\mathbb{E}[|e_{t+1}|] \le (1-\rho_t)\mathbb{E}[|e_t|] + O(\rho_{\max}) + O(\alpha_t). \tag{111}$$

Iterating (111) over $t \ge \tau_{k-1} + b$ and using $\rho_t \le \rho_{\max}$ yields

$$\begin{aligned} \mathbb{E}[|e_t|] &\le (1-\rho_{\max})^{t-(\tau_{k-1}+b)}\mathbb{E}[|e_{\tau_{k-1}+b}|] + \sum_{u=\tau_{k-1}+b}^{t-1}(1-\rho_{\max})^{t-1-u}\left(O(\rho_{\max}) + O(\alpha_u)\right) \\ &\le O(\rho_{\max}) + O\left(\sup_{u\in\mathcal{I}_k}\frac{\alpha_u}{\rho_{\max}}\right) + (1-\rho_{\max})^{t-(\tau_{k-1}+b)}\mathbb{E}[|e_{\tau_{k-1}+b}|]. \end{aligned} \tag{112}$$

Finally, by the mixing/burn-in bound (104) applied to the cost (bounded by $c_{\max}$), the bias in $\mathbb{E}[e_{\tau_{k-1}+b}]$ due to a non-stationary initial state contributes at most $O(C_{\text{mix}}\rho^b)$. Hence, uniformly for $t \ge \tau_{k-1} + b$,

$$\mathbb{E}[|\bar{c}^{(t)} - J_t|] \le O(\rho_{\max}) + O\left(\sup_{u\in\mathcal{I}_k}\frac{\alpha_u}{\rho_{\max}}\right) + O\left(C_{\text{mix}}\rho^b\right). \tag{113}$$

Since the baseline is updated on the fast timescale (in our algorithmic choices ρ_t is of the same order as β_t), we may replace $\rho_{\max}$ by $\beta_{\max}$ in the ratio term up to constants, yielding the lemma's stated $O(\sup \alpha/\beta)$ dependence.

H.5 Step 2: tracking of the critic parameter $w^{(t)}$

Define the critic error

$$\Delta_t \triangleq w^{(t)} - w_t^*. \tag{114}$$

From (101)-(102), we can rewrite the TD recursion as

$$\begin{aligned} w^{(t+1)} &= w^{(t)} - \beta_t \left(c_t - \bar{c}^{(t)} + \psi(s_{t+1})^\top w^{(t)} - \psi(s_t)^\top w^{(t)} \right) \psi(s_t) \\ &= w^{(t)} + \beta_t \left((\bar{c}^{(t)} - c_t)\psi(s_t) - \psi(s_t) \left(\psi(s_t) - \psi(s_{t+1})\right)^\top w^{(t)} \right). \end{aligned} \tag{115}$$

Introduce the sample quantities

$$A_t \triangleq \psi(s_t) \left(\psi(s_t) - \psi(s_{t+1})\right)^\top \text{ and } b_t \triangleq (\bar{c}^{(t)} - c_t)\psi(s_t), \tag{116}$$

so (115) is $w^{(t+1)} = w^{(t)} + \beta_t \left(b_t - A_t w^{(t)}\right)$. Subtract w_{t+1}^* from both sides and add/subtract the mean quantities A_t^*, b_t^*:

$$\begin{aligned} \Delta_{t+1} &= w^{(t+1)} - w_{t+1}^* \\ &= w^{(t)} - w_t^* + \beta_t \left(b_t - A_t w^{(t)}\right) + (w_t^* - w_{t+1}^*) \\ &= \Delta_t + \beta_t \left((b_t - b_t^*) - (A_t - A_t^*)w^{(t)}\right) + \beta_t \left(b_t^* - A_t^* w^{(t)}\right) + (w_t^* - w_{t+1}^*) \\ &= (I - \beta_t A_t^*)\Delta_t + \beta_t \underbrace{\left((b_t - b_t^*) - (A_t - A_t^*)w^{(t)}\right)}_{\text{martingale + mixing bias}} + \beta_t \underbrace{(b_t^* - A_t^* w_t^*)}_{=0 \text{ by } (98)} + (w_t^* - w_{t+1}^*) \\ &= (I - \beta_t A_t^*)\Delta_t + \beta_t \xi_t + (w_t^* - w_{t+1}^*), \end{aligned} \tag{117}$$

where $\xi_t \triangleq (b_t - b_t^*) - (A_t - A_t^*)w^{(t)}$. By (99), for $\beta_t \le 1/\|A_t^*\|$ (which holds for all sufficiently large t, we have the operator norm bound

$$\|I - \beta_t A_t^*\| \le 1 - \beta_t \lambda_A / 2. \tag{118}$$

First note $\|\psi(s)\| \le 1$ and $0 \le c_t \le c_{\max}$ imply $\|A_t\| \le 2$ and $\|b_t\| \le |\bar{c}^{(t)} - c_t| \le |\bar{c}^{(t)}| + c_{\max}$. Under boundedness and the fact that $\bar{c}^{(t)}$ is a convex combination of bounded costs, we have $|\bar{c}^{(t)}| \le c_{\max}$, hence $\|b_t\| \le 2c_{\max}$. Similarly $\|b_t^*\| \le 2c_{\max}$ and $\|A_t^*\| \le 2$. Using these and $\|w^{(t)}\| \le \|w_t^*\| + \|\Delta_t\|$,

$$\begin{aligned} \|\xi_t\| &\le \|b_t - b_t^*\| + \|A_t - A_t^*\| \|w^{(t)}\| \\ &\le \|b_t - b_t^*\| + \|A_t - A_t^*\| \left(\|w_t^*\| + \|\Delta_t\|\right). \end{aligned} \tag{119}$$

By the burn-in mixing bound (104) applied to the bounded functions defining A_t and b_t, for all $t \ge \tau_{k-1} + b$,

$$\mathbb{E}[\|A_t - A_t^*\|] + \mathbb{E}[\|b_t^* - (J_t - c_t)\psi(s_t)\|] \le O(C_{\text{mix}}\rho^b). \tag{120}$$

Moreover,

$$b_t - (J_t - c_t)\psi(s_t) = (\bar{c}^{(t)} - J_t)\psi(s_t), \tag{121}$$

so $\|b_t - (J_t - c_t)\psi(s_t)\| \le |\bar{c}^{(t)} - J_t|$. Combining with (113) yields

$$\mathbb{E}[\|b_t - b_t^*\|] \le \mathbb{E}[|\bar{c}^{(t)} - J_t|] + O(C_{\text{mix}}\rho^b) \le O(\rho_{\max}) + O\left(\sup_{u \in \mathcal{I}_k} \frac{\alpha_u}{\beta_u}\right) + O(C_{\text{mix}}\rho^b). \tag{122}$$

Substituting (120)-(122) into (119), and using that $\|w_t^*\|$ is uniformly bounded (a consequence of $\|(A_t^*)^{-1}\| \le 1/\lambda_A$ and bounded $\|b_t^*\|$), we obtain

$$\mathbb{E}[\|\xi_t\|] \le O(\rho_{\max}) + O\left(\sup_{u\in\mathcal{I}_k} \frac{\alpha_u}{\beta_u}\right) + O(C_{\text{mix}}\rho^b) + O\left(\mathbb{E}[\|\Delta_t\|]\cdot C_{\text{mix}}\rho^b\right). \tag{123}$$

The last term is higher order once $b = \Theta(t_{\text{mix}})$ is chosen so that $C_{\text{mix}}\rho^b$ is a small constant; we absorb it into constants in big-O. As in (109), uniform mixing plus bounded score functions imply that (A_t^*, b_t^*) are Lipschitz in ϕ_m, and by the matrix inverse perturbation identity,

$$w^*(\phi) = A^*(\phi)^{-1}b^*(\phi) \Rightarrow \|w^*(\phi) - w^*(\phi')\| \le L_w\|\phi - \phi'\| \tag{124}$$

for some $L_w < \infty$ depending only on $(c_{\max}, \lambda_A, C_{\text{mix}}, \rho, G_\pi)$. Thus

$$\|w_{t+1}^* - w_t^*\| \le O(\alpha_t). \tag{125}$$

Taking norms in (117), applying (118), and then taking expectations, for $t \ge \tau_{k-1} + b$,

$$\begin{aligned}\mathbb{E}[\|\Delta_{t+1}\|] &\le (1-\beta_t\lambda_A/2)\,\mathbb{E}[\|\Delta_t\|] + \beta_t\mathbb{E}[\|\xi_t\|] + \mathbb{E}[\|w_{t+1}^* - w_t^*\|] \\ &\le (1-\beta_t\lambda_A/2)\,\mathbb{E}[\|\Delta_t\|] + \beta_t\left(O(\rho_{\max}) + O\left(\sup_{u\in\mathcal{I}_k}\frac{\alpha_u}{\beta_u}\right) + O(C_{\text{mix}}\rho^b)\right) + O(\alpha_t). \end{aligned}\tag{126}$$

Using again that $\alpha_t \le \beta_t \cdot \sup_{u\in\mathcal{I}_k}(\alpha_u/\beta_u)$, we can rewrite (126) as

$$\mathbb{E}[\|\Delta_{t+1}\|] \le (1-\beta_t\lambda_A/2)\,\mathbb{E}[\|\Delta_t\|] + O(\beta_t\rho_{\max}) + O\left(\beta_t \sup_{u\in\mathcal{I}_k}\frac{\alpha_u}{\beta_u}\right) + O(\beta_t C_{\text{mix}}\rho^b). \tag{127}$$

Let $\beta_{\max} = \sup_{u\in\mathcal{I}_k}\beta_u$. A standard discrete Grönwall argument for sequences of the form $x_{t+1} \le (1-c\beta_t)x_t + \beta_t u$ yields (uniformly over $t \ge \tau_{k-1}+b$)

$$\mathbb{E}[\|\Delta_t\|] \le O(\beta_{\max}) + O(\rho_{\max}) + O\left(\sup_{u\in\mathcal{I}_k}\frac{\alpha_u}{\beta_u}\right) + O(C_{\text{mix}}\rho^b) + \exp\left(-\frac{\lambda_A}{2}\sum_{u=\tau_{k-1}+b}^{t-1}\beta_u\right)\mathbb{E}[\|\Delta_{\tau_{k-1}+b}\|]. \tag{128}$$

Finally, as in Appendix H.4, the burn-in/mixing lemma implies that the initialization error at time $\tau_{k-1}+b$ contributes at most an additional $O(C_{\text{mix}}\rho^b)$ bias term in expectation, which we absorb. Dropping the exponentially decaying term completes the desired bound for the critic.

H.6 Choosing the burn-in length $b = \Theta(t_{\text{mix}})$

Pick any constant $\varepsilon \in (0,1)$ and set

$$b \triangleq \min\{t \ge 1 :\ C_{\text{mix}}\rho^t \le \varepsilon\}. \tag{129}$$

Then $b = \Theta(t_{\text{mix}}(\varepsilon)) = \Theta(t_{\text{mix}})$ and the burn-in contribution becomes $O(\varepsilon)$. Substituting this choice into (113) and (128) yields the lemma statement:

$$\mathbb{E}\left[|\bar{c}^{(m,t)} - J^{(z)}(\pi^{(m)}_{\phi_m^{(t)}})|\right] \le O(\rho_{\max}) + O(\beta_{\max}) + O\Big(\sup_{u\in\mathcal{I}_k}\frac{\alpha_u}{\beta_u}\Big) + O\big(C_{\text{mix}}\rho^b\big), \tag{130}$$

$$\mathbb{E}\left[\|w_m^{(t)} - w^{*,(z,m)}(\phi_m^{(t)})\|\right] \le O(\beta_{\max}) + O\Big(\sup_{u\in\mathcal{I}_k}\frac{\alpha_u}{\beta_u}\Big) + O\big(C_{\text{mix}}\rho^b\big), \tag{131}$$

with constants depending only on $(c_{\max}, \lambda_{\min}, C_{\text{mix}}, \rho)$ (and the implied uniform conditioning constant λ_A).

□

I Proof of Lemma 3

Proof. We prove for a fixed-share gate under full-information losses $\ell_t(m) \in [0, C]$. We then relate the bound to the *post-projection* distribution $\tilde{g}_t$ used for sampling.

I.1 Step 0: Setup and the switching-aware gate update

Let $\ell_t \in [0, C]^M$ denote the loss vector at time t. Consider the fixed-share update on the simplex with parameters: learning rate $\eta > 0$ and share parameter $\alpha \in (0, 1)$. Initialize $w_1(m) = 1/M$ for all $m \in [M]$ and define $p_t(m) \triangleq w_t(m)/\sum_{j=1}^M w_t(j)$. Given ℓ_t, define the exponentiated update

$$\hat{w}_{t+1}(m) = p_t(m) \exp\left(-\eta \ell_t(m)\right) \text{ and } \hat{p}_{t+1}(m) = \frac{\hat{w}_{t+1}(m)}{\sum_{j=1}^M \hat{w}_{t+1}(j)}. \tag{132}$$

The fixed-share (switching-aware) distribution is then

$$p_{t+1}(m) = (1-\alpha)\hat{p}_{t+1}(m) + \alpha \cdot \frac{1}{M} \text{ for } m \in [M]. \tag{133}$$

This is the classical fixed-share Hedge algorithm (Herbster–Warmuth), which competes with expert sequences that switch a limited number of times.

For the present lemma, we first analyze the distribution generated by the switching-aware update (133); denote it by $g_t(\cdot)$ (to avoid overloading), and later relate it to $\tilde{g}_t(\cdot)$.

I.2 Step 1: A standard upper bound on the log-partition potential

Define the log-partition (potential)

$$W_t \triangleq \sum_{m=1}^M w_t(m) \text{ and } \Phi_t \triangleq \log W_t. \tag{134}$$

Using $p_t(m) = w_t(m)/W_t$ and the exponentiated update,

$$W_{t+1} = \sum_{m=1}^M w_{t+1}(m) = \sum_{m=1}^M ((1-\alpha)\hat{p}_{t+1}(m) + \alpha/M) \cdot W_{t+1}. \tag{135}$$

It is standard to analyze the *intermediate* normalization after exponentiation:

$$Z_t \triangleq \sum_{m=1}^M p_t(m) \exp(-\eta \ell_t(m)).$$

Then, $\log Z_t$ is the one-step change of the potential for the pure Hedge update (before sharing). By Hoeffding's lemma (or the convexity of exp), since $\ell_t(m) \in [0, C]$, we have

$$\log Z_t = \log \mathbb{E}_{m\sim p_t}\left[\exp(-\eta \ell_t(m))\right] \leq -\eta \mathbb{E}_{m\sim p_t}[\ell_t(m)] + \frac{\eta^2 C^2}{8}. \tag{136}$$

Summing (136) over $t = 1, \dots, T$ yields the usual Hedge upper bound:

$$\sum_{t=1}^T \mathbb{E}_{m\sim p_t}[\ell_t(m)] \leq \frac{\Phi_1 - \Phi_{T+1}^{(\text{Hedge})}}{\eta} + \frac{\eta C^2}{8} T, \tag{137}$$

where $\Phi_{T+1}^{(\text{Hedge})}$ denotes the log-partition after T pure-Hedge exponentiated steps. Fixed-share differs only in that it mixes $\hat{p}_{t+1}$ with the uniform distribution. The standard way to handle this is to lower bound the total weight assigned to a comparator expert sequence under the fixed-share dynamics, which we do next.

I.3 Step 2: A lower bound on the weight of a comparator switching sequence

Let $\mathbf{m}_{1:T} = (m_1, \dots, m_T)$ be any expert sequence with at most S switches, i.e.,

$$S(\mathbf{m}_{1:T}) \triangleq \sum_{t=2}^{T} \mathbb{1}\{m_t \neq m_{t-1}\} \leq S. \tag{138}$$

For fixed-share, one can interpret $p_t(\cdot)$ as the marginal of a Markov prior over expert indices with switch probability α: stay with probability $1-\alpha$, switch uniformly to one of M experts with probability α. Under this interpretation, the (unnormalized) weight assigned to $\mathbf{m}_{1:T}$ after observing losses is proportional to

$$\Pr(\mathbf{m}_{1:T}) \cdot \exp\left(-\eta \sum_{t=1}^{T} \ell_t(m_t)\right), \tag{139}$$

where

$$\Pr(\mathbf{m}_{1:T}) = \frac{1}{M} \cdot (1-\alpha)^{T-1-S(\mathbf{m}_{1:T})} \cdot \left(\frac{\alpha}{M}\right)^{S(\mathbf{m}_{1:T})}. \tag{140}$$

Consequently, the total normalizer (the total weight summed over all sequences) is at least the weight of the single sequence $\mathbf{m}_{1:T}$, i.e.,

$$W_{T+1}^{(\mathrm{FS})} \geq \frac{1}{M} \cdot (1-\alpha)^{T-1-S(\mathbf{m}_{1:T})} \cdot \left(\frac{\alpha}{M}\right)^{S(\mathbf{m}_{1:T})} \cdot \exp\left(-\eta \sum_{t=1}^{T} \ell_t(m_t)\right), \tag{141}$$

where $W_{T+1}^{(\mathrm{FS})}$ is the normalizer induced by the fixed-share recursion. Taking logs and using $S(\mathbf{m}_{1:T}) \leq S$ yields

$$\log W_{T+1}^{(\mathrm{FS})} \geq -\log M - (T-1-S)\log\frac{1}{1-\alpha} - S\log\frac{M}{\alpha} - \eta\sum_{t=1}^{T} \ell_t(m_t). \tag{142}$$

I.4 Step 3: Combine the upper and lower bounds to obtain switching regret

A standard fixed-share analysis (see Herbster–Warmuth) combines the one-step bound (136), which upper bounds the evolution of the normalizer for exponentiated updates, and the lower bound (142), which ensures the normalizer cannot be too small because it must include the comparator path.

Concretely, one obtains the following regret bound for the fixed-share prediction sequence $g_t(\cdot)$ generated by (133): for any comparator sequence $\mathbf{m}_{1:T}$ with $S(\mathbf{m}_{1:T}) \leq S$,

$$\sum_{t=1}^{T}\sum_{m=1}^{M} g_t(m)\ell_t(m) - \sum_{t=1}^{T} \ell_t(m_t) \leq \frac{\log M + S\log\frac{M}{\alpha} + (T-1-S)\log\frac{1}{1-\alpha}}{\eta} + \frac{\eta C^2}{8}T. \tag{143}$$

Equation (143) is the standard fixed-share bound. Its proof is exactly the potential argument summarized above, with the Markov-prior lower bound (141) playing the role of the "best expert" lower bound in classical Hedge.

I.5 Step 4: Specialize to the piecewise-constant in-class selector

In Lemma 3, the comparator is the piecewise-constant in-class selector $m_t^{\mathrm{ic}} \equiv m_k^{\mathrm{ic}}$ for $t \in \mathcal{I}_k$, which switches exactly S_T times: $S(\mathbf{m}_{1:T}^{\mathrm{ic}}) = S_T$. Applying (143) with $S = S_T$ gives

$$\sum_{t=1}^{T}\sum_{m=1}^{M} g_t(m)\ell_t(m) - \sum_{t=1}^{T} \ell_t(m_t^{\mathrm{ic}}) \leq \frac{\log M + S_T\log\frac{M}{\alpha} + (T-1-S_T)\log\frac{1}{1-\alpha}}{\eta} + \frac{\eta C^2}{8}T. \tag{144}$$

I.6 Step 5: Choose (α, η) and simplify

A convenient choice is $\alpha = \frac{S_T}{T-1}$ when $S_T \geq 1$ if $S_T = 0$ take any small constant α and the bound reduces to the standard Hedge bound). With this choice,

$$(T - 1 - S_T) \log \frac{1}{1-\alpha} = (T - 1 - S_T) \log \frac{T-1}{T-1-S_T} \leq S_T, \tag{145}$$

and

$$S_T \log \frac{M}{\alpha} = S_T \log \left(\frac{M(T-1)}{S_T} \right) \leq S_T \log(MT). \tag{146}$$

Thus, (144) becomes

$$\sum_{t=1}^{T} \sum_{m=1}^{M} g_t(m) \ell_t(m) - \sum_{t=1}^{T} \ell_t(m_t^{\text{ic}}) \leq \frac{\log M + S_T \log(MT) + S_T}{\eta} + \frac{\eta C^2}{8} T. \tag{147}$$

Choose

$$\eta = \sqrt{\frac{8 \left(\log M + S_T \log(MT) + S_T \right)}{C^2 T}}. \tag{148}$$

Plugging into (147) yields

$$\sum_{t=1}^{T} \sum_{m=1}^{M} g_t(m) \ell_t(m) - \sum_{t=1}^{T} \ell_t(m_t^{\text{ic}}) \leq O \left(C\sqrt{T \log M} + C S_T \log(MT) \right). \tag{149}$$

In the main text, it is common to suppress the additional $\log T$ factor with $\tilde{O}(\cdot)$ notation, in which case (149) is reported as $O\left(C\sqrt{T \log M} + C S_T \log M\right)$.

I.7 Step 6: From g_t to the post-projection sampling distribution $\tilde{g}_t$

In Algorithm 1, the distribution used to sample the expert is $\tilde{g}_t(\cdot)$, obtained by projecting (or modifying) $g_t(\cdot)$ to enforce $\tilde{g}_t(m_{\text{safe}}) \geq p_{\min}$. For arbitrary loss vectors, projection can only change the expected loss by at most C times the total mass moved. In particular, for the common "raise safe coordinate then renormalize" projection used in your paper, one can show for every t,

$$\sum_{m=1}^{M} \tilde{g}_t(m) \ell_t(m) \leq \sum_{m=1}^{M} g_t(m) \ell_t(m) + C p_{\min}, \tag{150}$$

because the projection increases the safe expert probability by at most $p_{\min}$ (if $g_t(m_{\text{safe}}) < p_{\min}$) and the loss range is $[0, C]$. Summing (150) over $t = 1, \ldots, T$ yields

$$\sum_{t=1}^{T} \sum_{m=1}^{M} \tilde{g}_t(m) \ell_t(m) - \sum_{t=1}^{T} \ell_t(m_t^{\text{ic}}) \leq \left(\sum_{t=1}^{T} \sum_{m=1}^{M} g_t(m) \ell_t(m) - \sum_{t=1}^{T} \ell_t(m_t^{\text{ic}}) \right) + C p_{\min} T. \tag{151}$$

Combining (151) with (149) yields the post-projection bound

$$\sum_{t=1}^{T} \sum_{m=1}^{M} \tilde{g}_t(m) \ell_t(m) - \sum_{t=1}^{T} \ell_t(m_t^{\text{ic}}) \leq O \left(C\sqrt{T \log M} + C S_T \log(MT) \right) + C p_{\min} T. \tag{152}$$

I.8 Conclusion

Ignoring the stability floor (or using $\tilde{O}(\cdot)$ notation to suppress $\log T$), the fixed-share gate satisfies

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\tilde{g}_t(m)\ell_t(m)-\sum_{t=1}^{T}\ell_t(m_t^{\mathrm{ic}})\le O\left(C\sqrt{T\log M}+C\,S_T\log M\right), \tag{153}$$

which is the claimed form in Lemma 3 (with the standard caveat discussed above regarding the safety projection and the possible additional $\log T$ factor under explicit parameter choices).

□

J Proof of Lemma 4

Proof. Recall that the regret $\mathrm{Reg}(T)=C_T(\pi_{1:T})-C_T^*$, the cost $C_T(\pi_{1:T})\triangleq\mathbb{E}\Big[\sum_{t=1}^{T}c^{(z_t)}(s_t,a_t)\Big]$, and the cost $C_T^*\triangleq\mathbb{E}\Big[\sum_{t=1}^{T}c^{(z_t)}(s_t^*,a_t^*)\Big]$, where $\pi_{1:T}$ is the (possibly history-dependent) algorithmic policy sequence, and $\pi_t^*=\pi^{*,(z_t)}$ is the regime-aware benchmark from (3).

J.1 Step 1: Insert the in-class regime-aware comparator

For each regime z, let the in-class best stationary policy (over the union of expert families) be $\pi^{\mathrm{ic},(z)}\in\arg\min_{\pi\in\cup_{m\in[M]}\Pi^{(m)}}J^{(z)}(\pi)$ and $J^{\mathrm{ic}}(z)\triangleq\min_{\pi\in\cup_{m\in[M]}\Pi^{(m)}}J^{(z)}(\pi)$. Let $m^{\mathrm{ic}}(z)$ be any expert index achieving $J^{\mathrm{ic}}(z)$, and define the piecewise-constant selector $m_t^{\mathrm{ic}}\equiv m_k^{\mathrm{ic}}\triangleq m^{\mathrm{ic}}(z_k)$ for $t\in\mathcal{I}_k$.

Define the (idealized) in-class regime-aware policy sequence $\pi_t^{\mathrm{ic}}\triangleq\pi^{\mathrm{ic},(z_t)}$ and its finite-horizon cost $C_T^{\mathrm{ic}}\triangleq\mathbb{E}\Big[\sum_{t=1}^{T}c^{(z_t)}(s_t^{\mathrm{ic}},a_t^{\mathrm{ic}})\Big]$, where $a_t^{\mathrm{ic}}\sim\pi^{\mathrm{ic},(z_t)}(\cdot\mid s_t^{\mathrm{ic}})$ and $s_{t+1}^{\mathrm{ic}}\sim P^{(z_t)}(\cdot\mid s_t^{\mathrm{ic}},a_t^{\mathrm{ic}})$. Then, by adding and subtracting C_T^{ic}, we have

$$\mathrm{Reg}(T)=\underbrace{\left(C_T(\pi_{1:T})-C_T^{\mathrm{ic}}\right)}_{\triangleq\,\mathrm{Reg}_{\mathrm{alg}\to\mathrm{ic}}(T)}+\underbrace{\left(C_T^{\mathrm{ic}}-C_T^*\right)}_{\triangleq\,\mathrm{Reg}_{\mathrm{ic}\to*}(T)}. \tag{154}$$

J.2 Step 2: Bound the approximation gap $\mathrm{Reg}_{\mathrm{ic}\to*}(T)$

By definition of Approx_π in (24), we have

$$J^{\mathrm{ic}}(z)-J^{(z)}(\pi^{*,(z)})\le\mathrm{Approx}_\pi,\quad\text{for all } z\in\mathcal{Z}. \tag{155}$$

If each regime were held fixed forever and both chains were initialized in stationarity, then the per-step gap between $\pi^{\mathrm{ic},(z)}$ and $\pi^{*,(z)}$ would be exactly $J^{\mathrm{ic}}(z)-J^{(z)}(\pi^{*,(z)})\le\mathrm{Approx}_\pi$. Over a finite horizon with switching, one must also account for the segment burn-in bias after each switch. We isolate this bias as $\mathrm{Reg}_{\mathrm{switch}}(T)$ (defined below).

Concretely, fix a burn-in length b (e.g., $b=t_{\mathrm{mix}}(\epsilon)$ from Definition 2) and decompose each segment $\mathcal{I}_k=\{\tau_{k-1},\dots,\tau_k-1\}$ into its burn-in part $\mathcal{I}_k^{\mathrm{burn}}\triangleq\{\tau_{k-1},\dots,\min\{\tau_{k-1}+b-1,\tau_k-1\}\}$ and its post-burn-in part $\mathcal{I}_k^{\mathrm{stat}}\triangleq\mathcal{I}_k\setminus\mathcal{I}_k^{\mathrm{burn}}$. Using bounded costs ($0\le c^{(z)}\le c_{\max}$), we can always upper bound the burn-in contribution by $c_{\max}$ per step and write

$$\mathrm{Reg}_{\mathrm{switch}}(T)\triangleq c_{\max}\sum_{k=1}^{S_T+1}|\mathcal{I}_k^{\mathrm{burn}}|\le c_{\max}(S_T+1)b. \tag{156}$$

Then, on the post-burn-in portions $\mathcal{I}_k^{\mathrm{stat}}$, Lemma 1 justifies replacing time averages by steady-state averages up to an $O(\epsilon)$ bias; absorbing these $O(\epsilon T)$ terms into (156) (by choosing ϵ as a fixed constant) yields

$$\mathrm{Reg}_{\mathrm{ic}\to *}(T) = C_T^{\mathrm{ic}} - C_T^* \le T \cdot \mathrm{Approx}_\pi + \mathrm{Reg}_{\mathrm{switch}}(T). \tag{157}$$

J.3 Step 3: Decompose $\mathrm{Reg}_{\mathrm{alg}\to\mathrm{ic}}(T)$ into gate-selection and within-expert learning

At each time t, Algorithm 1 produces (after safety projection) a sampling distribution $\tilde{g}_t(\cdot \mid s_t)$ over experts, samples $m_t \sim \tilde{g}_t(\cdot \mid s_t)$, then samples $a_t \sim \pi^{(m_t)}_{\phi^{(t)}_{m_t}}(\cdot \mid s_t)$. Define the conditional expected one-step cost if expert m were used at time t (given the realized s_t and the current parameters):

$$\bar{c}_t(m) \triangleq \mathbb{E}\left[c^{(z_t)}(s_t, a) \middle| s_t, z_t, a \sim \pi^{(m)}_{\phi^{(t)}_m}(\cdot \mid s_t) \right]. \tag{158}$$

Then, the algorithm's conditional expected one-step cost equals $\sum_m \tilde{g}_t(m \mid s_t)\bar{c}_t(m)$, so

$$\mathrm{Reg}_{\mathrm{alg}\to\mathrm{ic}}(T) = \sum_{t=1}^{T} \mathbb{E}\Big[\sum_m \tilde{g}_t(m \mid s_t)\bar{c}_t(m) \Big] - \sum_{t=1}^{T} \mathbb{E}\Big[\bar{c}_t(m_t^{\mathrm{ic}})\Big] \tag{159}$$

$$+ \underbrace{\sum_{t=1}^{T} \mathbb{E}\Big[\bar{c}_t(m_t^{\mathrm{ic}})\Big] - \sum_{t=1}^{T} \mathbb{E}\Big[c^{(z_t)}(s_t^{\mathrm{ic}}, a_t^{\mathrm{ic}})\Big]}_{\triangleq\ \mathrm{Reg}_{\mathrm{AC}}(T)}. \tag{160}$$

The last bracket is exactly a within-expert learning/modeling term. It measures how far the current parameterized policy $\pi^{(m_t^{\mathrm{ic}})}_{\phi^{(t)}_{m_t^{\mathrm{ic}}}}$ (and its induced trajectory) is from the ideal in-class stationary comparator $\pi^{\mathrm{ic},(z_t)}$. This is the term denoted $\mathrm{Reg}_{\mathrm{AC}}(T)$ in the lemma statement. It is controlled by Lemma 2 and standard average-cost policy-gradient arguments.

It remains to upper bound the first difference in (159), which is purely an expert-selection error:

$$\sum_{t=1}^{T} \mathbb{E}\Big[\sum_m \tilde{g}_t(m \mid s_t)\bar{c}_t(m) - \bar{c}_t(m_t^{\mathrm{ic}})\Big]. \tag{161}$$

J.4 Step 4: Relate expert-selection cost to the gate surrogate loss

By construction of the gate, we assume a calibration (or domination) relationship between instantaneous selection suboptimality and the surrogate gate loss $\ell_t(m) \in [0, C]$. Specifically, assume there exists $\kappa_1 \ge 1$ and an additive bias $\mathrm{Approx}_V \ge 0$ such that for all t and all distributions $q \in \Delta_M$,

$$\mathbb{E}\Big[\sum_m q(m)\bar{c}_t(m) - \bar{c}_t(m_t^{\mathrm{ic}})\Big] \le \kappa_1 \mathbb{E}\Big[\sum_m q(m)\ell_t(m) - \ell_t(m_t^{\mathrm{ic}})\Big] + \mathrm{Approx}_V. \tag{162}$$

Heuristically, Approx_V captures the fact that TD-residual losses are computed using approximate differential values, so the surrogate need not be perfectly aligned with true cost. In realizable settings, Approx_V can be taken as 0. Apply (162) with $q = \tilde{g}_t(\cdot \mid s_t)$ and sum over t:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\Big[\sum_m \tilde{g}_t(m \mid s_t)\bar{c}_t(m) - \bar{c}_t(m_t^{\mathrm{ic}})\Big] \\
&\le \kappa_1 \left(\sum_{t=1}^{T} \mathbb{E}\Big[\sum_m \tilde{g}_t(m \mid s_t)\ell_t(m)\Big] - \sum_{t=1}^{T} \mathbb{E}[\ell_t(m_t^{\mathrm{ic}})] \right) + T \cdot \mathrm{Approx}_V \\
&= \kappa_1 \mathrm{Reg}_{\mathrm{gate}}(T) + T \cdot \mathrm{Approx}_V,
\end{aligned} \tag{163}$$

where $\text{Reg}_{\text{gate}}(T)$ is exactly as defined in the lemma statement (note $\ell_t(m_t^{\text{ic}})$ is deterministic given the losses).

Combining (163) with (159) yields

$$\text{Reg}_{\text{alg}\to\text{ic}}(T) \leq \kappa_1 \text{Reg}_{\text{gate}}(T) + \text{Reg}_{\text{AC}}(T) + T \cdot \text{Approx}_V. \tag{164}$$

J.5 Step 5: Combine the pieces

Plugging (157) and (164) into (154) gives

$$\text{Reg}(T) \leq \kappa_1 \text{Reg}_{\text{gate}}(T) + \text{Reg}_{\text{AC}}(T) + \text{Reg}_{\text{switch}}(T) + T \cdot \text{Approx}_\pi + T \cdot \text{Approx}_V, \tag{165}$$

which is exactly (32).

□

K Proof of Theorem 5

Proof. We prove (34) by combining the three structural lemmas proved in the appendix: (i) the gate regret bound (Lemma 3), (ii) the critic/baseline tracking bound within a regime segment (Lemma 2), and (iii) the regret decomposition (Lemma 4). We also bound the switching transient term $\text{Reg}_{\text{switch}}(T)$ using the per-regime mixing property (Definition 2).

K.1 Step 1: Decomposition

By Lemma 4, under bounded costs and the calibration relation defining κ_1,

$$\text{Reg}(T) \leq \kappa_1 \text{Reg}_{\text{gate}}(T) + \text{Reg}_{\text{AC}}(T) + \text{Reg}_{\text{switch}}(T) + T \cdot \text{Approx}_\pi + T \cdot \text{Approx}_V. \tag{166}$$

We now bound the three regret components $\text{Reg}_{\text{gate}}(T)$, $\text{Reg}_{\text{AC}}(T)$, and $\text{Reg}_{\text{switch}}(T)$.

K.2 Step 2: Gate regret term

In the full-information variant, the gate observes bounded losses $\ell_t(m) \in [0, C]$ for all m. Applying Lemma 3 yields

$$\text{Reg}_{\text{gate}}(T) = \sum_{t=1}^{T} \sum_{m=1}^{M} \tilde{g}_t(m)\ell_t(m) - \sum_{t=1}^{T} \ell_t(m_t^{\text{ic}}) \leq c_0\Big(C\sqrt{T \log M} + CS_T \log M\Big), \tag{167}$$

for some absolute constant $c_0 > 0$ (e.g., $c_0 = O(1)$ depending on the exact fixed-share/Hedge variant). Multiplying by κ_1 gives

$$\kappa_1 \text{Reg}_{\text{gate}}(T) \leq O\Big(\kappa_1 C\sqrt{T \log M} + \kappa_1 CS_T \log M\Big). \tag{168}$$

K.3 Step 3: Switching transient term

We upper bound the transient cost incurred immediately after each regime switch, before the state distribution re-mixes within the new regime and the critic/baseline recenters.

Fix a constant accuracy level $\epsilon \in (0, 1/4]$ and let $b \triangleq t_{\text{mix}}(\epsilon)$ be the corresponding uniform mixing time (from Definition 2 plus the definition of $t_{\text{mix}}(\epsilon)$). Partition time into the $S_T + 1$ segments $\{\mathcal{I}_k\}_{k=1}^{S_T+1}$ with switch times $\{\tau_k\}$. For each segment k, define its first b steps as the "burn-in" subset

$$\mathcal{I}_k^{\text{burn}} \triangleq \{\tau_{k-1}, \tau_{k-1} + 1, \ldots, \min\{\tau_{k-1} + b - 1, \tau_k - 1\}\}. \tag{169}$$

By bounded costs ($0 \le c^{(z)} \le c_{\max}$), the maximal per-step contribution to regret in these burn-in steps is at most $c_{\max}$. Therefore, the total burn-in contribution across all segments is bounded by

$$\mathrm{Reg}_{\mathrm{switch}}(T) \le c_{\max} \sum_{k=1}^{S_T+1} |\mathcal{I}_k^{\mathrm{burn}}| \le c_{\max}(S_T+1)\, b = O\big(c_{\max} S_T\, t_{\mathrm{mix}}(\epsilon)\big), \tag{170}$$

where we used $b = t_{\mathrm{mix}}(\epsilon)$ and absorbed the additive $(+1)$ into the big-O. Taking ϵ as a fixed constant (e.g., $\epsilon = 1/4$) yields the stated scaling

$$\mathrm{Reg}_{\mathrm{switch}}(T) \le O\big(c_{\max} S_T\, t_{\mathrm{mix}}\big), \tag{171}$$

where t_{mix} is shorthand for $t_{\mathrm{mix}}(1/4)$.

K.4 Step 4: Within-expert learning term $\mathrm{Reg}_{\mathrm{AC}}(T)$

By definition in Lemma 4, $\mathrm{Reg}_{\mathrm{AC}}(T)$ collects the loss due to imperfect advantage surrogates and slow actor updates. We bound it by a standard "stochastic approximation under Markov noise" argument, using Lemma 2 to control the bias of the TD residual $\delta_t^{(m)}$ (as an advantage surrogate) within each stationary regime segment.

Fix a segment $\mathcal{I}_k$ with regime z_k, and consider the in-class comparator expert m_k^{ic} on this segment. Let $m \triangleq m_k^{\mathrm{ic}}$ for brevity. For the average-cost actor-critic update in Algorithm 1, the actor update for expert m uses the score $\nabla_{\phi_m} \log \pi_{\phi_m}^{(m)}(a_t \mid s_t)$ multiplied by the TD residual $\delta_t^{(m)}$. Under Assumption 1, we have a uniform score bound

$$\big\|\nabla_{\phi_m} \log \pi_{\phi_m}^{(m)}(a \mid s)\big\| \le G_\pi. \tag{172}$$

Moreover, under Assumption 2 and Lemma 2, after burn-in $b = \Theta(t_{\mathrm{mix}})$ within the segment, the critic/baseline tracking errors satisfy, for all $t \in \mathcal{I}_k$ with $t \ge \tau_{k-1} + b$,

$$\mathbb{E}\left[\big|\bar{c}^{(m,t)} - J^{(z_k)}(\pi_{\phi_m^{(t)}}^{(m)})\big|\right] \le \xi_{c,k}, \tag{173}$$

$$\mathbb{E}\left[\|w_m^{(t)} - w^{*,(z_k,m)}(\phi_m^{(t)})\|\right] \le \xi_{w,k}, \tag{174}$$

where (matching Lemma 2) one can take

$$\xi_{c,k} = O(\rho_{\max}) + O(\beta_{\max}) + O\Big(\sup_{u \in \mathcal{I}_k} \frac{\alpha_u}{\beta_u}\Big) + O(C_{\mathrm{mix}} \rho^b), \tag{175}$$

$$\xi_{w,k} = O(\beta_{\max}) + O\Big(\sup_{u \in \mathcal{I}_k} \frac{\alpha_u}{\beta_u}\Big) + O(C_{\mathrm{mix}} \rho^b). \tag{176}$$

These tracking errors imply that the TD residual $\delta_t^{(m)}$ is an approximately centered advantage surrogate within the segment. Its conditional expectation differs from the ideal average-cost advantage by at most a bias proportional to $\xi_{c,k} + \xi_{w,k}$. Because the actor update is scaled by step size α_t, the cumulative performance loss contributed by this bias over the segment is bounded by

$$\sum_{t \in \mathcal{I}_k : t \ge \tau_{k-1}+b} \alpha_t \mathbb{E}\Big[\big|(\text{bias in } \delta_t^{(m)})\big| \cdot \big\|\nabla_{\phi_m} \log \pi_{\phi_m}^{(m)}(a_t \mid s_t)\big\|\Big] \le G_\pi(\xi_{c,k} + \xi_{w,k}) \sum_{t \in \mathcal{I}_k} \alpha_t. \tag{177}$$

In addition, the martingale (noise) part of the actor update contributes the usual $\sum_t \alpha_t^2$ term. Concretely, because $|\delta_t^{(m)}| \le O(c_{\max}) + O(\|w_m\|)$ and the score is bounded by G_π, one obtains

$$\sum_{t \in \mathcal{I}_k} \alpha_t^2 \mathbb{E}\big[\|\delta_t^{(m)} \nabla_{\phi_m} \log \pi_{\phi_m}^{(m)}(a_t \mid s_t)\|^2\big] \le c_1 \sum_{t \in \mathcal{I}_k} \alpha_t^2 \tag{178}$$

for some finite c_1 depending only on $(c_{\max}, G_\pi)$ and the projection radius for (w_m). Summing (177) and (178) over all segments and using $\sum_k \sum_{t\in\mathcal{I}_k} \alpha_t = \sum_{t=1}^T \alpha_t$ and $\sum_k \sum_{t\in\mathcal{I}_k} \alpha_t^2 = \sum_{t=1}^T \alpha_t^2$ yields

$$\mathrm{Reg}_{\mathrm{AC}}(T) \le c_2 \sum_{t=1}^{T} \alpha_t + c_3 \sum_{t=1}^{T} \alpha_t^2 + c_4 \sum_{k=1}^{S_T+1} (\xi_{c,k} + \xi_{w,k}) \sum_{t\in\mathcal{I}_k} \alpha_t, \tag{179}$$

for constants c_2, c_3, c_4 depending only on boundedness parameters.

Finally, choose standard diminishing step sizes $\alpha_t = \alpha/\sqrt{t}$:

$$\sum_{t=1}^{T} \alpha_t = O(\sqrt{T}) \text{ and } \sum_{t=1}^{T} \alpha_t^2 = O(\log T), \tag{180}$$

and under two-timescale separation ($\sup_{u\in\mathcal{I}_k} \alpha_u/\beta_u \to 0$ and $\beta_{\max}, \rho_{\max} \to 0$), the tracking-error factors $\xi_{c,k}, \xi_{w,k}$ are $o(1)$ (or can be treated as constants absorbed into $\tilde{O}(\cdot)$ at finite T). Therefore, (179) gives the claimed sublinear rate

$$\mathrm{Reg}_{\mathrm{AC}}(T) \le \tilde{O}(\sqrt{T}), \tag{181}$$

where $\tilde{O}(\cdot)$ hides logarithmic factors (from $\sum_t \alpha_t^2$) and the constant tracking-error terms from Lemma 2. This matches the $\tilde{O}(\sqrt{T})$ term in (34).

K.5 Step 5: Combine all bounds

Substitute (168), (171), and (181) into (166):

$$\mathrm{Reg}(T) \le O\Big(\kappa_1 C\sqrt{T\log M} + \kappa_1 C S_T \log M\Big) + \tilde{O}(\sqrt{T}) + O\big(c_{\max} S_T t_{\mathrm{mix}}\big) + T(\mathrm{Approx}_\pi + \mathrm{Approx}_V), \tag{182}$$

which is exactly (34).

K.6 Vanishing average regret

Divide both sides by T. If $S_T = o(T)$ and $\mathrm{Approx}_\pi + \mathrm{Approx}_V = o(1)$, then each term on the right divided by T converges to 0:

$$\frac{\sqrt{T\log M}}{T} \to 0, \quad \frac{S_T \log M}{T} \to 0, \quad \frac{\tilde{O}(\sqrt{T})}{T} \to 0, \quad \frac{S_T t_{\mathrm{mix}}}{T} \to 0, \tag{183}$$

hence $\mathrm{Reg}(T)/T \to 0$. This completes the proof.

□

L Proof of Theorem 6

Proof. We use a standard Foster-Lyapunov drift argument.

L.1 Step 1: Drift inequality ensured by the safety projection

By construction of the safety projection in Algorithm 1, the post-projection sampling distribution satisfies

$$\tilde{g}_t(m_{\mathrm{safe}}) \ge p_{\min} > 0, \text{ for all } t. \tag{184}$$

For the queueing instantiations (and more generally, whenever a stabilizing baseline policy exists), we assume the following baseline drift property: there exist constants $B < \infty$ and $\epsilon > 0$ such that, if the stabilizing expert m_{safe} is selected with probability at least $p_{\min}$ at every time, then the induced queueing process satisfies the one-step conditional Lyapunov drift bound

$$\mathbb{E}\left[L(Q_{t+1}) - L(Q_t) \mid Q_t\right] \leq B - \epsilon \|Q_t\|_1, \text{ for all } t, \tag{185}$$

with $L(Q) \triangleq \frac{1}{2}\|Q\|_2^2$. Thus, under the stated condition $\tilde{g}_t(m_{\text{safe}}) \geq p_{\min}$, inequality (185) holds.

L.2 Step 2: Unconditional drift and telescoping

Taking total expectation of (185) and using the tower property yields, for every t,

$$\begin{aligned}\mathbb{E}\left[L(Q_{t+1})\right] - \mathbb{E}\left[L(Q_t)\right] &= \mathbb{E}\left[\mathbb{E}\left[L(Q_{t+1}) - L(Q_t) \mid Q_t\right]\right] \\ &\leq B - \epsilon \mathbb{E}\left[\|Q_t\|_1\right].\end{aligned} \tag{186}$$

Summing (186) over $t = 1, 2, \ldots, T$ gives a telescoping sum:

$$\mathbb{E}\left[L(Q_{T+1})\right] - \mathbb{E}\left[L(Q_1)\right] \leq BT - \epsilon \sum_{t=1}^{T} \mathbb{E}\left[\|Q_t\|_1\right]. \tag{187}$$

Since $L(\cdot) \geq 0$, we have $\mathbb{E}[L(Q_{T+1})] \geq 0$, and therefore

$$\epsilon \sum_{t=1}^{T} \mathbb{E}\left[\|Q_t\|_1\right] \leq BT + \mathbb{E}\left[L(Q_1)\right]. \tag{188}$$

Divide both sides by $T\epsilon$:

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|Q_t\|_1\right] \leq \frac{B}{\epsilon} + \frac{\mathbb{E}[L(Q_1)]}{\epsilon T}. \tag{189}$$

Taking $\limsup_{T\to\infty}$ on both sides and using $\frac{\mathbb{E}[L(Q_1)]}{\epsilon T} \to 0$ yields

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|Q_t\|_1\right] \leq \frac{B}{\epsilon}. \tag{190}$$

This is exactly the desired bound.

L.3 Step 3: Strong stability

Under the standard definition of strong stability (finite time-average expected backlog),

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\|Q_t\|_1\right] < \infty, \tag{191}$$

the bound above implies that the queue component $\{Q_t\}$ is strongly stable.

This completes the proof.

□

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009