REINFORCEMENT LEARNING IN INFINITE STATE SPACES: CONVERGENCE GUARANTEES VIA STABLE EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Q-learning is a simple and widely used model-free reinforcement learning algorithm with well-understood guarantees in finite state spaces. Extending it to countably infinite state spaces poses fundamental challenges due to potential instability, i.e. the state drifting away, induced by the exploration mechanism. In this paper, we present a novel framework for fully online, model-free Q-learning in infinite state spaces under discounted rewards. Our main contributions are twofold: *(i)* a proof of Q-learning convergence under stable exploration in infinite state spaces, *(ii)* an online algorithm that ensures a stable exploration. Combining *(i)* and *(ii)*, our results provide, to the best of our knowledge, the first convergence guarantees of Q-learning in infinite state spaces. We also present numerical examples to compare the performance of various variants of the framework.

1 INTRODUCTION

Reinforcement learning (RL) provides a powerful framework for solving sequential decision problems without requiring explicit knowledge of system dynamics. Among widely used and popular methods, Q-learning (20) stands out for its simplicity: the algorithm learns optimal policies purely from observed state transitions and rewards. This feature makes it particularly attractive for complex, high-dimensional systems where transition probabilities are unknown or difficult to model. A substantial body of the literature has focused on Q-learning in settings with finite state and action spaces, where the algorithm admits a relatively complete theoretical understanding (10; 18; 21). In such setting, convergence to the optimal value function has been established under mild conditions, most notably the well-know Robbins-Monro conditions. A range of finite-sample guarantees and performance bounds have been developed, see (11; 2) and references therein.

While Q-learning is well understood in finite state spaces, many practical systems naturally have countable infinite ones – for example, the number of users in communication or computing systems can grow arbitrarily large under congestion. In fact, as demonstrated in (1), conventional exploration strategies such as ε-greedy—while standard in finite state problems—can induce transience of the state process during training in infinite state problems, a phenomenon we refer to as instability, thereby preventing any meaningful convergence of Q-learning. Truncating the state space, i.e. imposing artificial boundary conditions, does not resolve the issue of instability since the state process might drift toward the boundary regardless of where it is set. This motivates our work, since to the best of our knowledge, no convergence results for Q-learning currently exist in infinite state spaces. Any such result would necessarily rely on a stable exploration that ensures that all state–action pairs must be visited infinitely often. The second challenge is precisely that of ensuring a stable exploration in the online setting.

The analysis of infinite state problems – which arise naturally in many operations research applications – such as queueing systems, inventory models, and population processes - has been addressed in several works. In (8), the authors prove that the natural policy gradient algorithm converges under certain assumptions to the optimal policy for the discounted reward criterion. These assumptions include that the algorithm must be initiated from a particular policy, which is in turn stable i.e. it generates a positive recurrent dynamics. Their setting is also model-based and offline. Previously, in (4), a deep reinforcement learning algorithm, coined proximal policy optimization, is considered

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

to deal with average reward problems for multiclass queuing networks. The algorithm proposed starts with a stable policy and iteratively improves its long-term average reward. Also, while being model-free, this approach is offline. Both approaches are based on the availability of stable policies, which cannot be guaranteed in general for model-free reinforcement learning problems. Regarding stable exploration, in (16), by means of a generative model, it is proposed to approximate the optimal policy, building on the fact that the optimal policy is stable. In (13), it is proposed to introduce into the reward structure a penalization for unstable actions. This penalization assumes the knowledge of a Lyapunov function for the system. We finally refer to the work of (12), which develops an algorithm – with resetting – for detecting (in)stability of Markov chains.

In this paper, we develop a framework for the analysis of Q-learning in countably infinite state spaces under discounted rewards. Our goal is thus to develop and analyze stabilization schemes that ensure the state–action process used to train Q-learning remains well-behaved in problems with infinite state spaces. Our main contribution is twofold.

(i) In our first contribution, see Theorem 3.1, we prove that Q-learning converges to the optimal Q-function in infinite state spaces provided the data generated by the exploration strategy satisfies the so-called Robbins-Monro conditions, which we refer to as "stable exploration". The proof extends classical techniques using weighted norms with polynomial growth.

(ii) In our second main contribution we provide an online procedure for stable exploration (Theorem 4.1). The procedure is based on sequential testing of policies. The key idea is that policies that satisfy a strong form of stability are eventually executed for a long period of time leading the system close to some ground state. This will compensates for the fact that previously executed unstable policies may drive the system far away from this ground state.

Combining the results in *(i)* and *(ii)*, we readily obtain Corollary 4.1, which is to the best of our knowledge, the first convergence guarantees of fully online model-free Q-learning in infinite state spaces. The analysis relies on weighted norms and renewal-theoretic arguments, and thus avoiding the heavy stochastic approximation machinery. The underlying idea of (12) is similar in spirit to the algorithm we propose in *(ii)*: policies (viewed as Markov chains in their setting) are executed for a certain duration in order to eventually "detect stability". Nevertheless, direct application of their approach to our framework is not possible, as it depends crucially on restart mechanisms and on prior knowledge of a Lyapunov function. In our online setting, resetting is not allowed, and we rely only on bounds for mean return times.

The paper is organized as follows. Section 2 introduces the MDP framework with countably infinite state spaces together with the main assumptions. Section 3 presents our results on the convergence of Q-learning conditioned on the exploration strategy satisfying the Robbins-Monro conditions, and Section 4 develops an online exploration strategy that guarantees Robbins-Monro to hold. Section 5 provides numerical examples illustrating the applicability of our approach. We conclude with a discussion of the results, possible extensions and open questions.

2 Model description

2.1 MDPs

A Markov decision process (MDP) models a setting in which an agent selects actions at each state, with each action determining transition probabilities over the state space and generating a corresponding reward. The objective is to identify a rule for choosing actions (a *policy*) that is optimal according to a given criterion, typically involving the maximization of accumulated rewards. In what follows, we formalize this framework and state the main assumptions under which we operate throughout the paper.

Let $\mathcal{X}$ be a discrete countably infinite state space, which we assumed is contained in $\mathbb{R}^d$ for some $d \geq 1$. Let $\mathcal{A}$ be a finite set of actions. At each state, an action can be selected, thereby determining the transition probabilities on the state space. We denote by $p(y \mid x, a)$ the probability of transitioning to state y from state x under action a.

We now introduce the central concept of *policy*. A policy is $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ where $\mathcal{P}(\mathcal{A})$ is the set of probability measures on $\mathcal{A}$. We denote $\pi(a \mid x)$ the probability of selecting action a at state x

according to policy π. Each policy π and initial condition $x_0 \in \mathcal{X}$ defines $(X_n)_{n\geq 0}$ a Markov chain on $\mathcal{X}$ such that $X_0 = x_0$ and for $n \geq 0$: $A_n \sim \pi(\cdot \,|\, X_n)$ and $X_{n+1} \sim p(\cdot \,|\, X_n, A_n)$. We denote $\mathbb{P}^\pi$ the distribution of the Markov chain defined by policy π, and $\mathbb{E}^\pi$ its corresponding expectation.

The main structural assumptions underlying our model are "locality" and "irreducibility". Locality rules out arbitrarily large transitions in a single step, and irreducibility guarantees that any state can be reached from any other state in a finite number of steps. We formalize this in the following assumption, where we denote by $|(x_1, \ldots, x_d)| = |x_1| + \cdots + |x_d|$ the ℓ^1-norm on $\mathbb{R}^d$.

Assumption 2.1. *There exists $\alpha > 0$ such that for any policy π the corresponding Markov chain $(X_n)_{n\geq 0}$ has finite range α, i.e. $|X_{n+1} - X_n| \leq \alpha$ a.s., and if $\pi(a \,|\, x) > 0$ for all (x, a) then the Markov chain is irreducible.*

As usual in RL, a reward is obtained after each transition. Let us denote by $r(x, a, y)$ the reward obtained after transition from state x to state y by selecting action a. We allow the reward to be unbounded but we impose that its growth is at most polynomial in $|\cdot|$ the ℓ_1-norm.

Assumption 2.2. *There are $\beta \geq 0$ and $C > 0$ such that if $W(x) = \max\{|x|, 1\}^\beta$ then $|r(x, a, y)| \leq CW(x)$ for all x, a and y for which $p(y \,|\, x, a) > 0$.*

Queueing problems (e.g., load balancing, server allocation, etc.) naturally fall into this set-up, and particularly satisfy Assumptions 2.1 and 2.2. In such settings, the state space is $\mathbb{N}^d$, where d stands for the number of queues, and the action set is finite. The reward function is typically given by $r(x, a, y) = -C|x|^\beta$ for some $\beta > 0$.

2.2 Online Q-learning for discounted reward problems

For a discounting parameter $0 < \gamma < 1$, the γ-discounted reward of a policy π is defined by $J^\pi_\gamma(x) := \mathbb{E}^\pi \left(\sum_{i\geq 0} \gamma^i r(X_i, A_i, X_{i+1}) \mid X_0 = x \right)$ for all x. The γ-discounted reward problem, consists of finding a policy π^* such that

$$J^{\pi^*}_\gamma(x) = J^*_\gamma(x) := \sup_\pi J^\pi_\gamma(x) \quad \text{for all } x. \tag{1}$$

A policy π^* satisfying (1) is called optimal. Under the assumptions of the previous section — in particular, Assumptions 2.1 and 2.2 — it is a classical result that optimal policies exist and can be characterized via the unique solution to the Bellman equation

$$Q^*(x, a) = \sum_y p(y \,|\, x, a) \big(r(x, a, y) + \gamma \max_b Q^*(y, b) \big). \tag{2}$$

Specifically, if $\pi^*(a|x) > 0$ only if $a \in \arg\max_{a' \in \mathcal{A}} Q^*(x, a')$ then π^* is optimal. For completeness, we provide details of these results in Appendix A, in particular the proof of existence and uniqueness of Q^* requires to work with the weighted norm $||Q||_W := \sup_{(x,a)} \frac{|Q(x,a)|}{W(x)}$, where W is as in Assumption 2.2.

The exact computation of Q^* is impossible in the model-free setting since the transition probabilities $p(y \,|\, x, a)$ are unknown. The Q-learning algorithm, first introduced in (20), is devised to iteratively approximate Q^* the solution to the Bellman equation (2) when the transitions probabilities $p(y \,|\, x, a)$ are unknown. Specifically, taking as an input an initialization $\widehat{Q}_0 : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the Q-learning performs the following update:

$$\widehat{Q}_{n+1}(x, a) := (1 - \alpha_n(x, a))\widehat{Q}_n(x, a) + \alpha_n(x, a) \left(R_{n+1} + \gamma \max_b \widehat{Q}_n(X_{n+1}, b) \right), \tag{3}$$

where

- *(i)* At step $n \geq 0$, an exploration strategy, that might depend on the entire history and time, selects an action A_n, leading to a state transition $X_{n+1} \sim p(\cdot \,|\, X_n, A_n)$ and reward $R_{n+1} = r(X_n, A_n, X_{n+1})$.
- *(ii)* α_n is called the *step-size* and takes the form

$$\alpha_n(x, a) = \frac{1}{n+1} \mathbb{1}(X_n = x, A_n = a) \quad \text{for all } n \geq 0. \tag{UD}$$

A popular exploration strategy is ε-greedy: at $n \geq 0$, with probability $\varepsilon > 0$ the action is selected uniformly from $\mathcal{A}$, and with probability $1 - \varepsilon$ the action is selected uniformly from $\arg\max_{a' \in A} \widehat{Q}_n(X_n, a')$. We also note that from *(ii)* it follows that only the Q-value of state-action pair (X_n, A_n) is updated after step n.

3 Convergence guarantees for online Q-learning

The so-called Robbins–Monro conditions play a crucial role in the convergence properties of the Q-learning algorithm. For our form of the step-size (UD), the Robbins-Monro conditions reduce to

$$\sum_{n \geq 0} \alpha_n(x, a) = +\infty \qquad \text{for all } (x, a) \quad \text{a.s.} \tag{RM}$$

Then, for finite state–space problems, it is a classical result (21; 18; 10) that if the Robbins–Monro conditions hold, then as $n \to \infty$, $\widehat{Q}_n$ converges towards Q^*. In the finite setting, it is sufficient for the Robbins-Monro conditions to hold that for any state, all actions have a strictly positive probability of being selected by the exploration strategy. It thus follows that Q-learning with ε-greedy as exploration strategy yields convergence.

In contrast, as we mentioned in the introduction, the results of (1) show that convergence of Q-learning with ε-greedy is not guaranteed in infinite state spaces. This is due to the fact that such exploration strategies might lead to a transient behavior of the state process, i.e., $X_n \to \infty$, thereby preventing Q-learning from converging.

In Theorem 3.1 below, we will show that in infinite state spaces, if the exploration strategy of Q-learning induces a sequence $(X_n, A_n)_n$ satisfying the Robbins-Monro conditions, i.e., the exploration is stable, then Q-learning does converge to the solution to the Bellman equation Q^*. To the best of our knowledge the result is missing from the literature. We provide a proof in Appendix B.

Theorem 3.1. *Impose Assumptions 2.1 and 2.2. Suppose that the Robbins-Monro conditions* (RM) *are satisfied, then, provided that* $||\widehat{Q}_0||_W < +\infty$*, we obtain for all* $0 \leq \gamma < 1$

$$\widehat{Q}_n(x, a) \xrightarrow{n \to +\infty} Q^*(x, a) \quad \textit{for all } (x, a) \quad \textit{a.s.}$$

Our proof of Theorem 3.1 builds on the finite-state argument of (10), with two main novel ingredients. First, we establish a uniform bound on the Q-learning iterates in the so-called W-norm. Second, we show that the error contribution to the Q-value at a state x coming from Q-values of states reachable in l steps from x vanishes as $l \to +\infty$.

4 Online stable exploration

In this section, we address the problem of devising stable exploration strategies, i.e. such that the sequence generated $(X_n, A_n)_n$ satisfiy Robbins-Monro. We develop Algorithm 1, a fully online model-free exploratory strategy, and then characterize in Theorem 4.1 sufficient conditions in order for Algorithm 1 satisfy Robbins-Monro.

4.1 Stable policies

We first consider Q-learning with exploration strategies consisting of persistently applying a fixed policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ (see Section 2) such that $\pi(a \,|\, x) > 0$ for all (x, a). We say a policy π is stable if the associated Markov chain $(X_n)_n$ it generates is positive recurrent, this is,

$$\mathbb{E}^{\pi}(\inf\{n \geq 1 : X_n = y \,|\, X_0 = x) < +\infty \text{ for all } x, y \in \mathcal{X}.$$

The following proposition establishes that using a stable policy π as exploration strategy is sufficient to guarantee convergence of Q-learning as defined in (3). A proof is provided in Appendix C.

Proposition 4.1. *Let* π *be a stable policy such that* $\pi(a \,|\, x) > 0$ *for all* $(x, a) \in \mathcal{X} \times \mathcal{A}$*. Then the sequence* $(X_n, A_n)_{n \geq 0}$ *generated by* π *satisfies* (RM)*.*

As in the finite state-space case, in order to have convergence, at every state, all actions must have a positive probability of being selected. Without this condition, certain state–action pairs would never be visited, leaving their Q-values unchanged.

In the model-free setting, the agent may lack prior knowledge about which policies are stable. Consequently, Proposition 4.1 cannot be applied directly. In our setting, instead of requiring knowledge of a single stable policy, we assume that the agent is given a subset of policies Λ containing at least one stable policy, see Assumption 4.1.

The challenge is to define an exploration strategy that selects policies from Λ, such that the generated sequence $(X_n, A_n)_n$ satisfies Robbins-Monro. In the rest of this section, we aim to present an online scheme based on iteratively selecting policies from Λ, ensuring that the Robbins-Monro conditions are satisfied. In order for our algorithm to work, we would need that: (C1) within Λ there are policies satisfying a —in general— stronger condition than mere stability, (C2) these policies are executed for sufficiently long time, and (C3) these policies are selected frequently enough.

4.2 THE CLASS $\Theta(c)$

In this subsection we define the class $\Theta(c)$ of stable policies, whose hitting time to a ground state satisfies a quantitative bound. This bound allows us to make a link between stable policies, see C1, and the notion of "sufficiently long time" invoked in C2 above.

Consider an arbitrary state $0 \in \mathcal{X}$ and let $\tau_0 := \inf\{i \geq 0 : X_i = 0\}$ i.e. the hitting time to state 0. We take state 0 as a reference, as it represents the empty system in queueing models, indicating the absence of jobs or customers. Informally speaking, we define $\Theta(c)$ as the set of policies for which the expected hitting time τ_0 with initial condition $X_0 = x$ scales linearly with $c|x|$ for a constant $c > 0$. Our algorithm will in turn make use of this feature in order to be able to exploit policies in $\Theta(c)$ when selected. More formally we will ask the following condition,

Definition 4.1. *For $c > 0$ we define $\Theta(c)$ as the subset of policies π for which*

$$\mathbb{P}^\pi(\tau_0 > (c+\delta)|x| \,|\, X_0 = x) \leq \exp(-u\delta|x|) \text{ for all } \delta > 0 \text{ and } x \in \mathcal{X}. \tag{4}$$

Hence, if a policy $\pi \in \Theta(c)$ is followed from state x for a time $c|x|$, there is a positive probability —increasing with $|x|$— of hitting state 0 before time $c|x|$. Thus, as we will show formally in Theorem 4.1, if policies in class $\Theta(c)$ (see C1) are executed for at least time $c|x|$ (see C2), and are selected sufficiently often (see C3), the condition of Robbins-Monro will be satisfied.

We argue now why in our online setting is natural to consider $\Theta(c)$ policies. On the one hand, we note that to the best of our knowledge, there is no general criteria that allows one to conclude whether a given policy is stable after observing it for a finite amount of time. On the other hand, while in general stronger than mere stability, $\Theta(c)$ policies provide a natural "online detection" of stability, in the sense that if one such policy is applied for a sufficiently long period of time, then by Definition 4.1, the generated trajectory will hit ground-state with high-probability.

Finally, as we detailed in Appendix G, we expect that in many applications, stable policies do indeed belong to the class of $\Theta(c)$. In particular, this holds for problems that admit a *(i)* well-defined fluid limit, *(ii)* density-dependent chains, and *(iii)* when Lyapunov functions exits, though they may be unknown to the agent.

4.3 ALGORITHM AND SUFFICIENT CONDITIONS FOR CONVERGENCE

The goal of this subsection is to construct an online exploratory strategy for training Q-learning such that the sequence $(X_n, A_n)_n$ it defines satisfies the Robbins-Monro condition. Our result requires the following assumption:

Assumption 4.1. *The agent has access to a subset of policies Λ that contains at least one policy from the class $\Theta(c)$.*

We acknowledge that, in a model-free setting, it is not apriori evident how to construct Λ. We discuss this in more detail in Appendix F.

We now present the main features of our scheme, and we refer to Algorithm 1 for a pseudo code.

The algorithm is organized in epochs, and within each epoch a policy $\pi \in \Lambda$ is selected and executed during time $\kappa|x|$ where x is the current position and $\kappa > 0$ (see lines 3-4 of Algorithm 1). Following the discussion from Section 4.2, if the currently selected policy belongs to class $\Theta(c)$ and $\kappa > c$, there is a positive probability —increasing in κ and $|x|$—of hitting state 0 before time $\kappa|x|$. Otherwise, the state process will most likely drift away. Rather than prescribing a specific way of selecting policies from Λ, in Theorem 4.1 we formulate assumptions on the selection mechanism governing Algorithm 1 that guarantee the desired stability properties. In Remark 4.1 we indicate settings where these assumptions follow. We want to emphasize that Algorithm 1 is not designed to detect stability within Λ but rather to exploit stable policies within Λ.

In case the constant c is not known a priori, every epoch κ is increased (see line 5 of Algorithm 1), in such a way that eventually $\kappa > c$. In case the c were known, κ could be to set to $c + \delta$, for arbitrary $\delta > 0$. The algorithm distinguishes two situations, depending on whether the ground state 0 is visited: *(i)* If state 0 is not visited (see lines 8-9 of Algorithm 1), and a new epoch starts, and *(ii)* If state 0 is visited (see lines 10-13 of Algorithm 1), an arbitrary exploration strategy (including the current π) is executed until the process exists $\mathcal{E}$ —a finite neighborhood of 0, moment in which a new epoch starts.

The distinction within the set $\mathcal{E}$ (lines 10-13 of Algorithm 1) is included for two main reasons. Firstly it prevents the algorithm from stalling in the (rare) case in which we end up at the ground state 0 at the end of the execution time, and secondly, it adds flexibility by for instance allowing to implement any exploration strategy within a finite set around the ground state.

Algorithm 1: Online stable exploration

Input: Initial state $X_0 = x \neq 0$; Subset of policies Λ; Finite subset of states $\mathcal{E}$ containing 0

1 Initialize $T_0 = 0$ and $\kappa_{-1} = 1$
2 **for** *epoch* $M \geq 0$ **do**
3 current state being $X_{T_M} \neq 0$
// Step 1: Select policy
4 Select a policy $\pi_M \in \Lambda$
// Step 2: Select κ
5 Select $\kappa_M > \kappa_{M-1}$ and set $t_{M+1} = \lceil \kappa_M |X_{T_M}| \rceil$
// Step 3: Execute policy
6 Construct $A_{T_M}, X_{T_M+1}, \ldots, A_{T_M+t_{M+1}-1}, X_{T_M+t_{M+1}}$ according to policy π_M, i.e.
7

$$A_{T_M+i} \sim \pi_M(\,\cdot\,|\,X_{T_M+i})$$
$$X_{T_M+i+1} \sim p(\,\cdot\,|\,X_{T_M+i}, A_{T_M+i})$$

for all $0 \leq i \leq t_{M+1} - 1$
// Step 4: Check if 0 is reached.
8 **if** $X_{T_M+i} \neq 0$ *for all* $1 \leq i \leq t_{M+1}$ **then**
9 Set $T_{M+1} = T_M + t_{M+1}$ and continue
10 **else**
11 Let i_0 be the first $1 \leq i \leq t_{M+1}$ such that $X_{T_M+i_0} = 0$
12 Execute from $X_{T_M+i_0} = 0$ any exploration strategy until leaving $\mathcal{E}$ for the first time, taking $\eta \geq 1$ steps
13 Set $T_{M+1} = T_M + i_0 + \eta$ and continue

We describe now sufficient conditions to ensure that the exploration strategy carried out by Algorithm 1 is stable, i.e. it generates a sequence $(X_n, A_n)_n$ satisfying the Robbins-Monro conditions: I) there is a deterministic number $D \geq 1$ such that for any epoch M a policy in class $\Theta(c)$ is selected between epochs $M, \ldots, M + D$; II) eventually the constants κ_M are set larger than $c > 0$; and III) eventually we select policies that assign positive probability to all actions, so as to ensure appropriate exploration. We formalize this in the following theorem, its proof is provided in Appendix E.

Theorem 4.1. *If the exploratory strategy is carried out by Algorithm 1 and the following assumptions are satisfied:*

I) There are $c > 0$ and $D \geq 1$ such that for all $M \geq 0$ there is $m \leq D$ satisfying $\pi_{M+m} \in \Theta(c)$ a.s.

II) There is $M_0 \geq 0$ such that $\kappa_M > c$ for all $M \geq M_0$.

III) There are $M_0' \geq 0$ and $q_0 > 0$ such that for all $M \geq M_0'$ we have $\pi_M(a \,|\, x) > q_0$ for all (x, a).

we obtain that the defined sequence $(X_n, A_n)_{n\geq 0}$ satisfies the Robbins-Monro condition.

Combining the results of Theorems 3.1 with those of Proposition 4.1 and Theorem 4.1, the following corollary establishes the convergence of Q-learning for discounting reward problems in infinite state spaces, and constitutes the concluding contribution of the paper.

Corollary 4.1. *Impose Assumptions 2.1 and 2.2. Suppose that Q-learning is trained with exploration strategy given by a stable policy or Algorithm 1 satisfying Assumptions I-III, then Q-learning converges to Q^*, i.e. $\widehat{Q}_n(x, a) \to Q^*(x, a)$ for all (x, a) a.s.*

Remark 4.1. *We note that Assumptions I and III will naturally hold in some settings. For instance, Assumption I is satisfied if Λ is finite and policies are selected from Λ according to a round-robin scheme, i.e., they are chosen one after another in some fixed order. Next, if needed, policies in Λ could be perturbed by a small amount $q_0 > 0$ in order to ensure Assumption III, i.e. for any $\pi \in \Lambda$ consider $\pi_{q_0}(a \,|\, x) = (1 - q_0)\pi(a \,|\, x) + q_0/\,|\mathcal{A}|$.*

While the proof of Theorem 4.1 requires Assumption I, we expect that, e.g., selecting policies uniformly from such Λ, will also work, see the next section for a practical illustration of this.

5 Numerical examples

In this section we illustrate our results by means of concrete examples. We first consider a one dimensional model where the state space is $\mathcal{X} = \mathbb{N}$, the set of actions is $\mathcal{A} = \{0, 1, 2\}$. The transition probabilities are given by $p(x + 1 \,|\, x, a) = p_a$ and $p(x - 1 \,|\, x, a) = 1 - p_a$ for $x > 0$, where $p_0 = 0.65, p_1 = 0.45$ and $p_2 = 0.4$. The reward is given by $r(x, a, y) = -|x|$. We further note that for this example the optimal policy is $\pi^*(x) = 2$ for all $x \geq 1$.

Consider the set of parametrized policies $\widehat{\Lambda} = \{(\lambda_0, \lambda_1, \lambda_2) : 0 < \lambda_i \leq 1, \lambda_0 + \lambda_1 + \lambda_2 = 1\}$ where we naturally identify $\lambda \in \widehat{\Lambda}$ with $\pi_\lambda(a \,|\, x) = \lambda_a$ for $a = 0, 1, 2$. It is not hard to see that π_λ is stable if and only if $\lambda_0 p_0 + \lambda_1 p_1 + \lambda_2 p_2 < 1/2$, and in that case it belongs to class $\Theta(c_\lambda)$ for some $c_\lambda > 0$, see Section H.

In the following Q-learning will be always initialized with $\widehat{Q}_0 = 0$, the discounting parameter is $\gamma = 0.999$. The initial condition is $x_0 = 25$. We set also $\varepsilon = 0.1$.

We consider the following exploration strategies:

A1) We perform our Algorithm 1 with: $\Lambda = \{\pi_1, \ldots, \pi_{10}\} \subset \widehat{\Lambda}$, where some π_i is stable for some i (therefore $\pi_i \in \Theta(c)$); $\mathcal{E} = \{0, \ldots, 20\}$ and while in $\mathcal{E}$ actions are selected uniformly; Outside $\mathcal{E}$ policy selection follows a Round Robin scheme, selecting one policy at each epoch; $\kappa_0 = 10$ and $\kappa_{M+1} = \kappa_M + 1$.

A2) We perform our Algorithm 1 with: $\Lambda = \widehat{\Lambda}$; $\mathcal{E} = \{0, \ldots, 20\}$ and π^e being ε-greedy; When a policy is selected (outside $\mathcal{E}$) and brings the process to state 0, we append this policy to Λ_0, where Λ_0 is initialized as $\varnothing$. Then, the policy selection will depend on the current state X_{T_M}: with probability $\delta_M = 1/X_{T_M}$ we choose uniformly from Λ and with probability $1 - \delta_M$ we choose uniformly from Λ_0; $\kappa_0 = 10$ and $\kappa_{M+1} = \kappa_M + 1$.

SP) We train with the fixed stable policy $\pi_{(0.3,0.4,0.3)}$ (cf. Proposition 4.1).

EG) We perform the classical version of Q-learning with ε-greedy.

We observe that scheme A1 will satisfy Assumptions I-III in Theorem 4.1, see Remark 4.1. On the other hand, A2 does not. Indeed with A2 Assumption I is not satisfied since in this case policies in class $\Theta(c)$ for some $c > 0$ are going to be selected eventually, but the number of epochs until this happens is a random variable that can take arbitrarily large values. Our numerical experiments suggest that the conditions given in Theorem 4.1 are sufficient but not strictly necessary, provided that the algorithm executes policies in $\Theta(c)$ often enough. Convergence of Q learning in the scheme SP

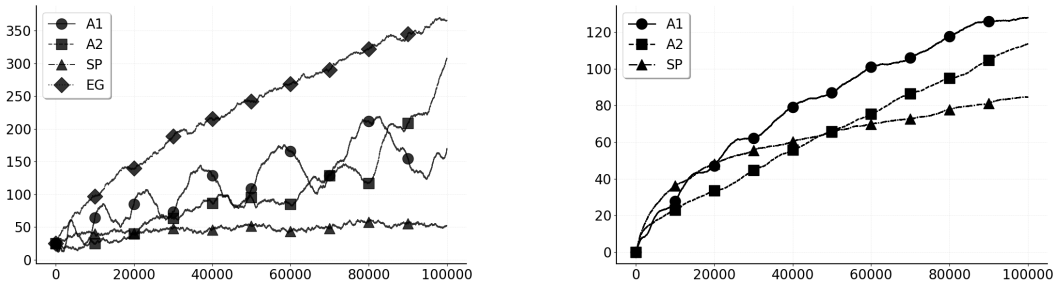

Figure 1: *(left)* Average trajectory $\frac{1}{T}\sum_i^T X^i(n)$ as a function of n, for schemes A1, A2, SP and EG. *(right)* Average number of states $\frac{1}{T}\sum_i L_5^i(n)$ for which all actions where tried at least 5 times and the current best action is the optimal by time n as a function of n.

follows from Corollary 4.1. Regarding EG, our numerical results indicate that EG yields instability in this case, the process drifts away, and no learning takes place. Instability of EG cannot be deduced from (1), since the results of (1) exclude the case in which the drift vectors lie on the same line and have opposite directions.

We simulate $T = 30$ sample paths $(X_n^i, A_n^i)_n$ for $n = 0, \dots, N$ for $N = 1000000$, for each scheme A1, A2, SP and EG, and we represent our results in Figure 1.

In Figure 1 *(left)* we plot the average trajectories obtained for each exploration strategy, i.e. $n \to \frac{1}{T}\sum_i^T X_n^i$. EG exhibits an average trajectory with positive drift indicating instability. On the other hand, SP yields trajectories that tend to a constant, indicating convergence to steady-state and stability. The trajectories of A1 and A2 show fluctuations with increasing amplitude which are a consequence of κ_M (and hence t_{M+1}) increasing, but the recurrent selection of stable policies produces a data sequence that satisfies Robbins-Monro. The latter is guaranteed for A1, and we verify it numerically for A2.

In Figure 1 *(right)* we visualize learning over time. We consider that at time n we have "learned" the optimal action at state x if (i) $\arg\max_a \widehat{Q}_n(x, \cdot) = 2$, and (ii) that all pairs (x, a) for $a = 0, 1, 2$ have been visited at least $k \geq 0$ times. The number of states at time n for which we have learned is denoted by $L_k(n)$. Thus, Figure 1 *(right)* plots the average learning $n \to \frac{1}{T}\sum_i^T L_5^i(n)$ as a function of n, and for the various strategies. We do not consider EG, because its instability implies that the same state is not visited sufficiently often. From Corollary 4.1, we know that convergence of Q-learning to the optimal values is guaranteed for A1 and SP, but not for A2. We observe that the learning increases more slowly for SP and A2 than for A1.

Comparing the curves on the *(left)* and *(right)* of Figure 1 we observe an interesting interplay between stability and learning. Indeed, for the "most stable" strategy on the left, SP, the learning increases more slowly. On the other hand, the trajectories of A1 and A2 are "less stable", in the sense that show much larger excursions, which allow them to learn at a faster rate. The characterization of the learning rate is an interesting phenomenon that we leave for future work.

We finally consider a two dimensional model. In this example we have $\mathcal{X} = \mathbb{N}^2$, $\mathcal{A} = \{0, 1\}$. Also, let for each $a = 0, 1$ a probability μ_a on $\{e_1, -e_1, e_2, -e_2\}$, the set of canonical vectors of $\mathbb{R}^2$. The transition probabilities $p(y \,|\, x, a)$ are defined by $p(x \pm e_i \,|\, x, a) = \mu_a(\pm e_i)$ for $a = 0, 1$ and $x = (x(1), x(2))$ with $x(i) \neq 0$. We will consider the parametrized set of policies $\widehat{\Lambda} = [0, 1]$ where we identify λ with π_λ such that $\pi_\lambda(0 \,|\, x) = \lambda$ for all x. We assume that the transitions on the boundary states are such that π_0 and π_1 are stable policies, and from the analysis of Appendix H, both policies belong to class $\Theta(c)$ for some $c > 0$.

We further assume that π_λ is not stable for some λ. Such a situation is represented Figure 2 (left) where we denote by $d_a := (\mu_a(e_1) - \mu_a(-e_1), \mu_a(e_2) - \mu_a(-e_2))$, the drift of action $a = 0, 1$ at interior states. This is an interesting setting, in which there are two stable policies, π_0 and π_1, but their convex combination can be unstable. This echoes the Parrondo (9) paradox from probability theory, where alternating between two individually "winning" strategies can produce a losing outcome.

Conditions for achieving such dynamics can be found in (6). As in the previous example, we consider $T = 30$ sample-paths and the following exploration strategies:

A1) Algorithm 1 with: $\Lambda = \{\pi_{\lambda_1}, \pi_{\lambda_2}, \pi_{\lambda_3}\} \subset \widehat{\Lambda}$, where π_{λ_1} and π_{λ_2} are stable but π_{λ_3} is not; $\mathcal{E} = \{0, \ldots, 10\}^2$ and while in $\mathcal{E}$ actions are selected uniformly; Outside $\mathcal{E}$ policy selection follows a Round Robin scheme, selecting one policy at each epoch; $\kappa_0 = 50$ and $\kappa_{M+1} = \kappa_M + 1$.

A2) Algorithm 1 with: $\Lambda = \{\pi_{\lambda_1}, \ldots, \pi_{\lambda_{10}}\} \subset \widehat{\Lambda}$, where some π_{λ_i} are stable and $0 < \lambda_i < 1$; $\mathcal{E} = \{0, \ldots, 10\}^2$, while in $\mathcal{E}$ actions are selected uniformly; Outside $\mathcal{E}$, when a policy is selected and brings the process to state 0, we append this policy to Λ_0, where Λ_0 is initialized as $\varnothing$. Then, the policy selection will follow round-robin while $\Lambda_0 = \varnothing$. Later, if eventually $\Lambda_0 \neq \varnothing$, policy selection will follow a variant of round-robin where at epoch M with probability $\delta_M = 1/|X_{T_M}|$ we choose the following policy according to the next policy following round-robin, and with probability $1 - \delta_M$ we choose uniformly from Λ_0; $\kappa_0 = 50$ and $\kappa_{M+1} = \kappa_M + 1$.

SP) We consider a fixed stable policy $\pi_{0.8}$.

EG) We perform the classical version of Q-learning with ε-greedy.

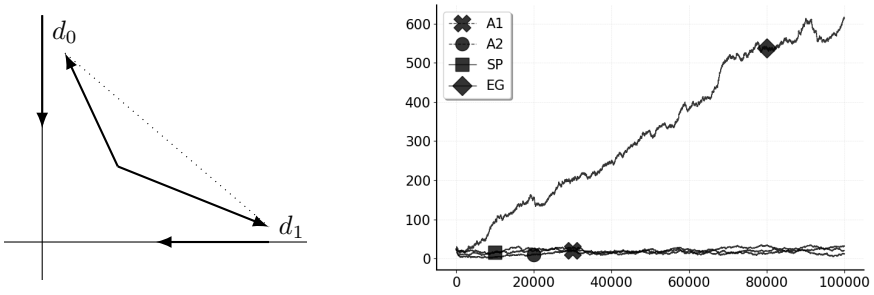

Figure 2: *(left)* Transition dynamics for 2D example, *(right)* Average trajectories over time.

From Figure 2 *(right)* we observe that while A1, A2 and SP produce stable trajectories, the trajectory of EG increases linearly on the logarithmic scale. In fact, the results of (1) imply that EG is unstable for the parameter setting under consideration.

6 Conclusion

The initial motivation for our paper was the observation that standard exploration strategies that work well in finite state-space problems, do not necessarily carry over to the infinite case. In that respect, we can cite a quote by Watkins, co-author of the foundational paper (21) on Q-learning, who in a retrospective account observes (see (19)) that "... sophisticated control of exploration is essential for more complex tasks. In a sense, reinforcement learning only provides half an algorithm: the other half is the control of exploration." Our main contribution, i.e., deriving an online exploration strategy that ensures convergence of Q-learning in infinite state spaces, can be thus seen as providing an answer to Watkins' intuition.

There are several interesting research avenues that stem from our work. All our results are for the discounted case, and a natural extension would be to study conditions such that Q-learning converge under other criteria (cf. average criterion). Our Algorithm 1 can be seen as a general recipe that leaves room for improvement. For example, it would be interesting to relax Assumption I to allow for uniform selection of policies, or more complex forms of random policy selection. Another interesting question is whether the condition that stable policies need to belong to $\Theta(c)$ can be relaxed. As mentioned in the introduction, algorithms to detect stability of Markov Chains, see (12), rely on restart mechanisms. It would be very interesting to study whether an online scheme as that of Algorithm 1 can be adapted in order to detect stability in a fully online setting.

REFERENCES

[1] U. Ayesta, S. Foss, M. Jonckheere, and V. Puricelli. When Q-learning fails: unstable behavior for infinite state spaces. *SIGMETRICS Perform. Eval. Rev. (Special issue on short papers IFIP Performance 2025)*, 53(4):79–83, Mar. 2026.

[2] C. Beck and R. Srikant. Error bounds for constant step-size Q-learning. *Systems & Control Letters*, 61(12):1203–1208, 2012.

[3] M. Bramson. *Stability of Queueing Networks*, volume 1950 of *Lecture Notes in Mathematics*. Springer, Berlin, 2008.

[4] J. G. Dai and M. Gluzman. Queueing network controls via deep reinforcement learning. *Stochastic Systems*, 12(1):30–67, 2021.

[5] R. W. R. Darling and J. R. Norris. Differential equation approximations for Markov chains. *Probability Surveys*, 5:37–79, 2008.

[6] G. Fayolle, V. A. Malyshev, and M. V. Menshikov. *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge University Press, 1995.

[7] A. Gosavi. Boundedness of iterates in Q-learning. *Systems & Control Letters*, 55(4):347–349, 2006.

[8] I. Grosof, S. T. Maguluri, and R. Srikant. Convergence of natural policy gradient for a family of infinite-state queueing mdps. *Queueing Systems*, 109(3):22, 2025.

[9] G. Harmer and D. Abbott. Parrondo's paradox. *Statistical Science*, 14:206–213, 1999.

[10] T. Jaakkola, M. I. Jordan, and S. P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Comput.*, 6(6):1185–1201, 1994.

[11] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 4868–4878, Red Hook, NY, USA, 2018. Curran Associates Inc.

[12] M. Mandjes, B. Patch, and N. Walton. Detecting Markov chain instability: a Monte Carlo approach. *Stochastic Systems*, 7(2):289–314, 2017.

[13] B. S. Pavse, M. Zurek, Y. Chen, Q. Xie, and J. P. Hanna. Learning to stabilize online reinforcement learning in unbounded state spaces. volume 235 of *Proceedings of Machine Learning Research*, pages 40015–40033. PMLR, 21–27 Jul 2024.

[14] M. L. Puterman. Markov decision processes. Stochastic models, Handb. Oper. Res. Manage. Sci. 2, 331-434 (1990)., 1990.

[15] P. Robert. *Stochastic networks and queues*, volume 52 of *Appl. Math. (N. Y.)*. Berlin: Springer, 2003.

[16] D. Shah, Q. Xie, and Z. Xu. Stable reinforcement learning with unbounded state space. *arXiv preprint: 2006.04353*, 2020.

[17] A. N. Shiryaev. *Probability*, volume 95 of *Grad. Texts Math.* New York, NY: Springer-Verlag, 2nd ed. edition, 1995.

[18] J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202, 1994.

[19] C. Watkins. Reinforcement learning: some history. `https://www.cs.rhul.ac.uk/home/chrisw/RL_some_history.html`.

[20] C. Watkins. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge United Kingdom, 1989.

[21] C. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

7 AI USE STATEMENT

In this work, we used generative AI tools only for polishing writing.

A BELLMAN EQUATION FOR INFINITE STATE SPACE PROBLEMS A.1

In this Appendix we justify the existence of the solution to the Bellman equation (2) in the set-up described in Section 2. This result follows from (14, Section 6.10.1).

First observe that (2) can be regarded as a fixed point equation for the so call Bellman operator:

$$TQ(x,a) := \sum_y p(y \,|\, x,a)(r(x,a,y) + \gamma \max_b Q(y,b)). \tag{5}$$

Proposition A.1 ((14))**.** *Suppose that Assumption 2.1 and 2.2 hold. Consider*

$$L(W) := \{f : ||f||_W := \sup_{(x,a)} \frac{|f(x,a)|}{W(x)} < +\infty\}.$$

Then for any $0 < \gamma < 1$

1. *The Bellman operator has a unique fixed point*
$$TQ^*(x,a) = Q^*(x,a) \in L(W). \tag{6}$$
2. *From any* $Q_0 \in L(W)$ *define* $Q_{k+1} = TQ_k$ *for all* $k \geq 0$*, then*
$$||Q_{k+1} - \ Q^*||_W \to \ 0.$$
3. *Optimal policies* π^* *are characterized by*
$$\pi^*(a \,|\, x) > 0 \implies a \in \arg\max_{\mathcal{A}} Q^*(x, \cdot\,)$$

Proof. We make use of Theorem 6.10.4 from (14). We check Assumptions 6.10.1 and 6.10.2 of that theorem. Assumption 6.10.1 is exactly our Assumption 2.2. Now Assumption 2.1 implies that for all (x,a)

$$\sum_y p(y \,|\, x,a) \frac{W(y)}{W(x)} \leq (1+\alpha)^\beta.$$

Denote $p_\pi^J(x,y) := \mathbb{P}^\pi(X_J = y \,|\, X_0 = x)$, we obtain for any $0 \leq \lambda < 1$ that

$$\lambda^J \sum_y p_\pi^J(x,y) \frac{W(y)}{W(x)} \leq \lambda^J (1 + J\alpha)^\beta \to 0, \ J \to \infty.$$

From this we conclude that Assumption 6.10.2 is also satisfied and that Theorem 6.10.4 holds.

□

B PROOF OF THEOREM 3.1

In order to establish Theorem 3.1, we will need a series of two lemmas coming from classical stochastic approximation theory and a third lemma where we show that the Q-learning algorithm iterates (3) are uniformly bounded in the $||\cdot||_W$ norm. This is a well known result for finite state–spaces; see (7).

The following result can be found in (18, Section 3, Lemma 1).

Lemma B.1. *For each* $n \geq 1$ *let* α_n *and* E_{n-1} *be* $\mathcal{F}_n-$*measurable random variables for which there is* $B > 0$ *such that a.s.*

1. $0 \leq \alpha_n \leq 1$;
2. $\sum_{n\geq 1} \alpha_n = +\infty$;
3. $\sum_{n\geq 1} \alpha_n^2 \leq B$;
4. $\mathcal{F}_n \subset \mathcal{F}_{n+1}$;
5. $\mathbb{E}(E_n \,|\, \mathcal{F}_n) = 0$;
6. $\mathbb{E}(E_n^2 \,|\, \mathcal{F}_n) \leq B$.

Then if $\delta_{n+1} = (1-\alpha_n)\delta_n + \alpha_n E_n$ *we obtain that* $\delta_n \to 0$ *a.s. for any* $\delta_0 \in \mathbb{R}$.

Lemma B.2. *Consider $(\alpha_n)_{n\geq 0}$ as in Lemma B.1 and let $C > 0$. Then if $(\lambda_n)_{n\geq 0}$ satisfies $\lambda_{n+1} \leq (1-\alpha_n)\lambda_n + \alpha_n C$, we have $\limsup_n \lambda_n \leq C$.*

Proof. Define $\delta_0 = \lambda_0$ and $\delta_{n+1} = (1-\alpha_n)\delta_n + \alpha_n C$. First, by induction, $\lambda_n \leq \delta_n$ for all $n \geq 0$. Second, by Lemma B.1, $\delta_n \to C$, concluding the proof. □

Before stating the third lemma needed, we define a quantity g that controls how much the reward can grow from one transition to another. Observe that from Assumption 2.2 we can conclude that

$$|r(x,a,y)| \leq C W_M(x) \quad \text{if } p(y \,|\, x,a) > 0,$$

where $W_M(x) = \max\{|x|, M\}^\beta$ for $M \geq 1$.

Now, if $p(y \,|\, x,a) > 0$ for some a we have, by Assumption 2.1, that

$$\frac{W_M(y)}{W_M(x)} \leq g_M := (1+\alpha/M)^\beta.$$

Then, given $\gamma \in (0,1)$, we can take $M_0 \geq 1$ such that $\gamma g_{M_0} < 1$. For such M_0, with a slight abuse of notation, we set till the rest of this appendix

$$g = g_{M_0} \quad \text{and} \quad W = W_{M_0}.$$

Observe that, in the notation of Appendix A, *(i)* $L(W_M) = L(W_{M'})$ for all $M, M' \geq 1$; *(ii)* the Bellman operator (5) is a contraction for $||\cdot||_{W_M}$ for all $M \geq 1$. Therefore, the solution Q^* to the Bellman equation (2) is independent of $M \geq 1$.

Lemma B.3. *Under the assumptions of Theorem 3.1, we obtain*

$$||\widehat{Q}_n||_W \leq \frac{1}{1-\gamma g} \max\left\{||r||_W, ||\widehat{Q}_0||_W\right\} \quad \textit{for all } n \geq 0. \tag{7}$$

where

$$||r||_W := \sup_{(x,a)} \frac{\max\{|r(x,a,y)| : p(y \,|\, x,a) > 0\}}{W(x)} < +\infty;$$

cf. Assumption 2.2.

Proof of Lemma B.3. The proof essentially follows that in (7). The objective is to show by induction that, with $M := \max\{||r||_W, ||\widehat{Q}_0||_W\}$,

$$||\widehat{Q}_n||_W \leq M(1 + \cdots + (\gamma g)^n), \tag{8}$$

from which the conclusion of the lemma immediately follows since $\gamma g < 1$.

Fix $n = 1$ and consider any (x,a). Supposing first that $X_0 \neq x$ or $A_0 \neq a$, we have

$$\frac{|\widehat{Q}_1(x,a)|}{W(x)} = \frac{|\widehat{Q}_0(x,a)|}{W(x)} \leq M < M(1+\gamma g)$$

Now if $X_0 = x$ and $A_0 = a$,

$$\begin{aligned}
\frac{|\widehat{Q}_1(x,a)|}{W(x)} &\leq (1-\alpha(1))\frac{|\widehat{Q}_0(x,a)|}{W(x)} + \alpha(1)\left(\frac{|r(x,a,X_1)|}{W(x)} + \gamma\frac{|\max_b \widehat{Q}_0(X_1,b)|}{W(x)}\right) \\
&\leq (1-\alpha(1))M + \alpha(1)\left(M + \gamma\frac{|\max_b \widehat{Q}_0(X_1,b)|}{W(X_1)}\frac{W(X_1)}{W(x)}\right) \\
&\leq (1-\alpha(1))M + \alpha(1)(M + M\gamma g) = M(1+\alpha(1)\gamma g) \leq M(1+\gamma g).
\end{aligned}$$

Now let $n > 1$ and assume (8) holds. Again consider any (x,a). If $X_n \neq x$ or $A_n \neq a$ we obtain

$$\frac{|\widehat{Q}_{n+1}(x,a)|}{W(x)} = \frac{|\widehat{Q}_n(x,a)|}{W(x)} \leq M(1 + \cdots + (\gamma g)^n) < M(1 + \cdots + (\gamma g)^{n+1}).$$

Now if $X_n = x$ and $A_n = a$ we have similarly as before that

$$\begin{aligned}
\frac{|\widehat{Q}_{n+1}(x,a)|}{W(x)} &\le (1-\alpha(n+1))\frac{|\widehat{Q}_n(x,a)|}{W(x)} \\
&\quad + \alpha(n+1)\left(\frac{|r(x,a,X_{n+1})|}{W(x)} + \gamma\frac{|\max_b \widehat{Q}_n(X_{n+1},b)|}{W(x)}\right) \\
&\le (1-\alpha(n+1))M \\
&\quad + \alpha(n+1)\left(M + \gamma\frac{|\max_b \widehat{Q}_n(X_{n+1},b)|}{W(X_{n+1})}\frac{W(X_{n+1})}{W(x)}\right) \\
&\le (1-\alpha(n+1))M(1+\cdots+(\gamma g)^n) + \alpha(n+1)\big(M + M(1+\cdots+(\gamma g)^n)\,\gamma g\big) \\
&\le (1-\alpha(n+1))M(1+\cdots+(\gamma g)^{n+1}) + \alpha(n+1)M\big(1+\cdots+(\gamma g)^{n+1}\big) \\
&\le M\big(1+\gamma g+\cdots+(\gamma g)^{n+1}\big).
\end{aligned}$$

This proves the lemma. □

Proof of Theorem 3.1. We analyze the dynamics of the error

$$\Delta_n(x,a) := \widehat{Q}_n(x,a) - Q^*(x,a).$$

Using the fact that Q^* satisfies the Bellman equation (2) and the definition of $\widehat{Q}_n$, we can split the error as $\Delta_n(x,a) = \Delta_n^{(1)}(x,a) + \Delta_n^{(2)}(x,a) + \Delta_n^{(3)}(x,a)$, where

$$\begin{aligned}
\Delta_{n+1}^{(1)}(x,a) &:= (1-\alpha_n(x,a))\,\Delta_n^{(1)}(x,a) \\
&\qquad + \alpha_n(x,a)\sum_y p(y\,|\,x,a)\big(r(x,a,X_{n+1}) - r(x,a,y)\big), \\
\Delta_{n+1}^{(2)}(x,a) &:= (1-\alpha_n(x,a))\,\Delta_n^{(2)}(x,a) \\
&\qquad + \alpha_n(x,a)\sum_y p(y\,|\,x,a)\,\gamma\big(J^*(X_{n+1}) - J^*(y)\big), \\
\Delta_{n+1}^{(3)}(x,a) &:= (1-\alpha_n(x,a))\,\Delta_n^{(3)}(x,a) \\
&\qquad + \alpha_n(x,a)\,\gamma\big(\widehat{J}_n(X_{n+1}) - J^*(X_{n+1})\big),
\end{aligned}$$

adopting the notation $J^*(x) := \max_a Q^*(x,a)$ and similarly $\widehat{J}_n(x) := \max_a \widehat{Q}_n(x,a)$.

The error terms $\Delta^{(1)}$ and $\Delta^{(2)}$ can be dealt with in the same way as is typically done in the framework of classical stochastic approximation theory. Specifically, Lemma B.1 implies that for $i = 1, 2$

$$\Delta_n^{(i)}(x,a) \xrightarrow{n\to+\infty} 0 \quad \text{for all } (x,a) \text{ a.s.} \tag{9}$$

Now fix some finite and nonempty subset $\mathcal{K}_0 \subset \mathcal{X}$ (for instance $\mathcal{K}_0 = \{x_0\}$) and construct $\mathcal{K}_1 \subset \cdots \subset \mathcal{K}_L \subset \ldots$ by

$$\mathcal{K}_{l+1} := \mathcal{K}_l \cup \bigcup_{(x,a)\in\mathcal{K}_l\times\mathcal{A}} \{y \in \mathcal{X} : p(y\,|\,x,a) > 0\},$$

where each $\mathcal{K}_l$ is finite because of Assumption 2.1.

Consider for any subset $\mathcal{K} \subset \mathcal{X}$ the restricted $||\cdot||_W$ norm, defined by

$$||f||_W^{\mathcal{K}} := \sup_{x\in\mathcal{K}, a\in\mathcal{A}} \frac{|f(x,a)|}{W(x)}.$$

Fix $L \ge 1$ and $\varepsilon > 0$. Since $\mathcal{K}_L$ is finite and $W(x) \ge 1$ for all x, we obtain from (9) that a.s. there is $N_0 \ge 1$ such that

$$||\Delta_n^{(1)} + \Delta_n^{(2)}||_W^{\mathcal{K}_L} < \varepsilon \text{ for all } n \ge N_0.$$

According to Proposition A.1 and Lemma B.3 we can take a constant $M > 0$ such that

$$\|\Delta_n\|_W^{\mathcal{K}_L} = \|\widehat{Q}_n - Q^*\|_W^{\mathcal{K}_L} \leq \|\widehat{Q}_n - Q^*\|_W \leq M \text{ for all } n \geq 0 \text{ a.s.}$$

Considering $(x, a) \in \mathcal{K}_{L-1} \times \mathcal{A}$, we obtain

$$\begin{aligned}
\frac{|\Delta_{n+1}^{(3)}(x,a)|}{W(x)} &\leq (1-\alpha_n(x,a))\frac{|\Delta_n^{(3)}(x,a)|}{W(x)} + \alpha_n(x,a)\,\gamma\,\frac{\left|\widehat{J}_n(X_{n+1}) - J^*(X_{n+1})\right|}{W(x)} \\
&= (1-\alpha_n(x,a))\frac{|\Delta_n^{(3)}(x,a)|}{W(x)} + \alpha_n(x,a)\,\gamma g\,\frac{\left|\widehat{J}_n(X_{n+1}) - J^*(X_{n+1})\right|}{W(X_{n+1})} \\
&\leq (1-\alpha_n(x,a))\frac{|\Delta_n^{(3)}(x,a)|}{W(x)} + \alpha_n(x,a)\,\gamma g\,\|\Delta_n\|_W^{\mathcal{K}_L} \\
&\leq (1-\alpha_n(x,a))\frac{|\Delta_n^{(3)}(x,a)|}{W(x)} + \alpha_n(x,a)\,\gamma g M.
\end{aligned}$$

This recurrence inequality implies (see Lemma B.2) that a.s. for each $(x, a) \in \mathcal{K}^{L-1} \times \mathcal{A}$ there is $N_1(x, a) \geq 1$ such that

$$\frac{|\Delta_n^{(3)}(x,a)|}{W(x)} \leq \gamma g M \text{ for all } n \geq N_1(x,a).$$

Then, recalling that $\mathcal{K}_{L-1}$ is finite, we let $N_1 := \max\{N_1(x,a), N_0\}$ to obtain that a.s.

$$\|\widehat{Q}_n - Q^*\|_W^{\mathcal{K}_{L-1}} \leq \|\Delta_n^{(1)} + \Delta_n^{(2)}\|_W^{\mathcal{K}_{L-1}} + \|\Delta_n^{(2)}\|_W^{\mathcal{K}_{L-1}} \leq \varepsilon + \gamma g M \text{ for all } n \geq N_1.$$

Now considering $(x, a) \in \mathcal{K}_{L-2} \times \mathcal{A}$ and $n \geq N_1$,

$$\begin{aligned}
\frac{|\Delta_{n+1}^{(3)}(x,a)|}{W(x)} &\leq (1-\alpha_n(x,a))\frac{|\Delta_{n+1}^{(3)}(x,a)|}{W(x)} + \alpha_n(x,a)\gamma g\|\Delta_n\|_W^{\mathcal{K}_{L-1}} \\
&\leq (1-\alpha_n(x,a))\frac{|\Delta_{n+1}^{(3)}(x,a)|}{W(x)} + \alpha_n(x,a)\gamma g(\varepsilon + \gamma g M)
\end{aligned}$$

Arguing exactly as before, we claim the existence of $N_2 \geq N_1$ such that

$$\|\widehat{Q}_n - Q^*\|_W^{\mathcal{K}_{L-2}} \leq \|\Delta_n^{(1)} + \Delta_n^{(2)}\|_W^{\mathcal{K}_{L-2}} + \|\Delta_n^{(3)}\|_W^{\mathcal{K}_{L-2}} \leq \varepsilon + \gamma g(\varepsilon + \gamma g M) \text{ for all } n \geq N_2.$$

After repeating this argument for L steps, we obtain that a.s. there is $N_L \geq N_{L-1}$ such that

$$\|\widehat{Q}_n - Q^*\|_W^{\mathcal{K}_0} \leq \frac{1}{1-\gamma g}\varepsilon + (\gamma g)^L M \text{ for all } n \geq N_L.$$

We conclude by recalling that by assumption $\gamma g < 1$. □

C Proof of Proposition 4.1

We show first that the Markov chain $(X_n, A_n)_{n\geq 0}$ is irreducible and positive recurrent. Irreducibility is due to the fact that $(X_n)_n$ is irreducible (Assumption 2.1) and since it is assumed that $\pi(a \,|\, x) > 0$ for all (x, a).

Now, positive recurrence is equivalent to the existence of an invariant probability measure. Therefore, since π is stable, the Markov chain $(X_n)_{n\geq 0}$ has an invariant probability measure μ_π on $\mathcal{X}$. From this it is not hard to check that $\nu_\pi(x, a) = \pi(\, a \,|\, x)\mu_\pi(x)$ is an invariant probability measure for $(X_n, A_n)_n$.

Now, since $(X_n, A_n)_n$ is irreducible and positive recurrent, $\nu_\pi(x, a) > 0$ and by the ergodic theorem

$$\lim_n \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}(X_i = x, A_i = a) = \nu_\pi(x,a) \text{ a.s.}$$

Fix then (x, a) and $0 < \delta < \nu(x, a)$, there is a.s. an $N \geq 1$ such that for all $n \geq N$

$$\frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}(X_i = x, A_i = a) \geq \delta > 0.$$

Thus, a.s. for $n \geq N$

$$\begin{aligned} \alpha_n(x, a) &= \frac{1 + \sum_{i=0}^{n} \mathbb{1}(X_i = x, A_i = a)}{n+1} \frac{\mathbb{1}(X_n = x, A_n = a)}{1 + \sum_{i=0}^{n} \mathbb{1}(X_i = x, A_i = a)} \\ &\geq \delta \, \frac{\mathbb{1}(X_n = x, A_n = a)}{1 + \sum_{i=0}^{n} \mathbb{1}(X_i = x, A_i = a)}. \end{aligned}$$

Because of irreducibility and positive recurrence for each (x, a), the random times

$$\begin{aligned} \eta(1) &:= \inf\{i \geq 0 : X_i = x, A_i = a\} \\ \eta(l+1) &:= \inf\{i > \eta(l) : X_i = x, A_i = a\} \end{aligned}$$

are finite and $\eta(l) \to +\infty$ a.s.

As a consequence, we conclude that

$$\sum_{n \geq 1} \alpha_n(x, a) \geq \delta \sum_{n \geq N} \frac{\mathbb{1}(X_n = x, A_n = a)}{1 + \sum_{i=0}^{n} \mathbb{1}(X_i = x, A_i = a)} \geq \delta \sum_{l:\, \eta(l) \geq N} \frac{1}{l+1} = +\infty \text{ a.s.}$$

□

D Proof of Proposition D.1

The proof of Theorem 4.1 relies on the proposition stated below (see Appendix E). This proposition provides moment bounds for the visit times to each state–action pair under the dynamics induced by Algorithm 1.

Proposition D.1. *Suppose Assumptions I), II) and III) are satisfied, consider*

$$\begin{aligned} \tau_{(x,a)}(1) &:= \inf\{t \geq 0 : X_i = x, A_i = a\} \\ \tau_{(x,a)}(l+1) &:= \inf\{t > \tau_{(x,a)}(l) : X_i = x, A_i = a\}. \end{aligned}$$

Then, for each $\sigma \geq 1$ there is a constant $\beta_{(x,a)}(\sigma) > 0$ such that

$$\mathbb{E}\left(\left(\tau_{(x,a)}(l+1) - \tau_{(x,a)}(l)\right)^{\sigma}\right) \leq \beta_{(x,a)}(\sigma) \text{ for all } l \geq 1.$$

The proof relies on demonstrating that the first hitting time of state 0 for the proposed scheme has all the required moments. We then extend this result to the first and subsequent visits to any state–action pair, thereby concluding the proof of Proposition D.1. The proposition will be established through a sequence of claims.

Recall Assumption II) and III) of Theorem 4.1. Without loss of generality, we assume throughout this proof that $M_0 = M_0' = 0$. Otherwise, we repeat the same arguments as bellow after epoch $M_0'' = \max\{M_0, M_0'\}$. Therefore, for all M we have $\kappa_M > c$ and $\pi_M(a \,|\, x) > q_0$ for all (x, a).

Consider the sequence of epochs M_l where a policy in class $\Theta(c)$ is selected. This can be recursively defined via

$$\begin{aligned} M_1 &:= \inf\{M \geq 0 : \pi_M \in \Theta(c)\}, \text{ and} \\ M_{l+1} &:= \inf\{M > M_l : \pi_M \in \Theta(c)\}, \quad \text{for } l \geq 1. \end{aligned}$$

For convenience of notation we set $M_0 := 0$. Because of Assumption I), as given in Theorem 4.1, we obtain that $M_{l+1} - M_l \leq D$ a.s., and therefore in particular $M_l < +\infty$ for all $l \geq 0$.

Define $\tau_0^0 := \tau_0 = \inf\{t \geq 0 : X_t = 0\}$ and for each $j \geq 1$

$$\tau_0^j := \inf\{t > T_j : X_t = 0\}.$$

Claim D.1. *For any $l \geq 1$*

$$\mathbb{E}_x\left(\mathbb{1}(\tau_0 > T_{M_l+1})F(x, X_{T_{M_1}}, \kappa_{M_1}, \ldots, X_{T_{M_l}}, \kappa_{M_l})\right)$$
$$\leq \sum_{\substack{x_1,\ldots,x_l\neq 0\\ c<k_1\leq\cdots\leq k_l}} \exp\left(-u\sum_{i=1}^{l}(k_i-c)|x_i|\right) F(x, x_1, k_1, \ldots, x_l, k_l)\mu_l(x_1, k_1, \ldots, x_l, k_l)$$

where μ_l is the joint distribution of $(X_{T_{M_1}}, \kappa_{M_1}, \ldots, X_{T_{M_l}}, \kappa_{M_l})$ conditioned on $\{\tau_0 > T_{M_l}\}$.

Proof of Claim D.1. In the event $\{\tau_0 > T_{M_l+1}\}$ we have $X_{T_{M_1}}, \ldots, X_{T_{M_l}} \neq 0$. It will be convenient to consider for $x_1, \ldots, x_l \neq 0$ and $c < k_1 \leq \cdots \leq k_l$ the event

$$\Omega(x_1, k_1, \ldots, x_l, k_l) = \{X_{T_{M_1}} = x_1, \kappa_{M_1} = k_1, \ldots, X_{T_{M_l}} = x_l, \kappa_{M_l} = k_l\}.$$

Thus,

$$\mathbb{E}_x\left(\mathbb{1}(\tau_0 > T_{M_{l+1}})F(x, X_{T_{M_1}}, \kappa_{M_1}, \ldots, X_{T_{M_l}}, \kappa_{M_l})\right)$$
$$= \sum_{\substack{x_1,\ldots,x_l\neq 0\\ c<k_1\leq\cdots\leq k_l}} \mathbb{P}\left(\tau_0 > T_{M_l+1}, \Omega(x_1, k_1, \ldots, x_l, k_l)\right) F(x, x_1, k_1, \ldots, x_l, k_l).$$

Now, since $\{\tau_0 > T_{M_l+1}\} = \{\tau_0^{M_l} > t_{M_l+1}, \tau_0 > T_{M_l}\}$ we have

$$\mathbb{P}\left(\tau_0 > T_{M_l+1}, \Omega(x_1, k_1, \ldots, x_l, k_l)\right)$$
$$= \mathbb{P}\left(\tau_0^{M_l} > t_{M_l+1} \,|\, \tau_0 > T_{M_l}, \Omega(x_1, k_1, \ldots, x_l, k_l)\right) \mathbb{P}\left(\tau_0 > T_{M_l}, \Omega(x_1, k_1, \ldots, x_l, k_l)\right).$$

By definition $\pi_{M_l} \in \Theta(c)$, therefore,

$$\mathbb{P}\left(\tau_0^{M_l} > t_{M_l+1} \,|\, \tau_0 > T_{M_l}, \Omega(x_1, k_1, \ldots, x_l, k_l)\right) \leq \exp(-a(k_l - c)|x_l|)$$

Now,

$$\mathbb{P}\left(\tau_0 > T_{M_l}, \Omega(x_1, k_1, \ldots, x_l, k_l)\right)$$
$$= \mathbb{P}\left(X_{T_{M_l}} = x_l, \kappa_{M_l} = k_l \,|\, \tau_0 > T_{M_l}, \Omega(x_1, k_1, \ldots, x_{l-1}, k_{l-1})\right) =: \delta_l(x_l, k_l \,|\, x_1, k_1, \ldots, x_{l-1}, k_{l-1})$$

We obtain finally that

$$\mathbb{P}\left(\tau_0 > T_{M_l+1}, \Omega(x_1, k_1, \ldots, x_l, k_l)\right)$$
$$\leq \exp(-a(k_l - c)|x_l|)\delta_l(x_l, k_l \,|\, x_1, k_1, \ldots, x_{l-1}, k_{l-1})\mathbb{P}\left(\tau_0 > T_{M_l}, \Omega(x_1, k_1, \ldots, x_{l-1}, k_{l-1})\right)$$
$$\leq \exp(-a(k_l - c)|x_l|)\delta_l(x_l, k_l \,|\, x_1, k_1, \ldots, x_{l-1}, k_{l-1})\mathbb{P}\left(\tau_0 > T_{M_{l-1}+1}, \Omega(x_1, k_1, \ldots, x_{l-1}, k_{l-1})\right)$$
$$\leq \exp\left(-u\sum_{j=1}^{l}(k_j - c)|x_j|\right) \delta_l(x_l, k_l \,|\, x_1, k_1, \ldots, x_{l-1}, k_{l-1}) \ldots \delta_1(x_1, k_1 \,|\, x),$$

where from the second to the third step we use the fact that $\{\tau_0 > T_{M_l}\} \subset \{\tau_0 > T_{M_{l-1}+1}\}$.

We end by noticing that $\mu_l(x_1, k_1, \ldots, x_l, k_l) = \delta_l(x_l, k_l \,|\, x_1, k_1, \ldots, x_l, k_l) \ldots \delta_1(x_1, k_1 \,|\, x)$ is by definition the distribution of $(X_{T_{M_1}}, \kappa_{M_1}, \ldots, X_{T_{M_l}}, \kappa_{M_l})$ conditioned to $\{\tau_0 > T_{M_l}\}$. □

Claim D.2. *In the event $\{T_{M_l+1} < \tau_0 \leq T_{M_{l+1}+1}\}$ we have*

$$\tau_0 \leq (1+\alpha)^D \sum_{i=0}^{l} |X_{T_{M_i}}|\kappa_{M_{i+1}}^D \tag{10}$$

Proof of Claim D.2. Fix $l \geq 1$ we first observe that

$$\{T_{M_l+1} < \tau_0 \leq T_{M_{l+1}+1}\} = \bigcup_{j=1}^{M_{l+1}-M_l} \{T_{M_l+j} < \tau_0, \tau_0^{M_l+j} \leq t_{M_l+j+1}\}.$$

Now, for each $j = 1, \dots, M_{l+1} - M_l$, in the event $\{T_{M_l+j} < \tau_0, \tau_0^{M_l+j} \leq t_{M_l+j+1}\}$ we obtain

$$\begin{aligned}\tau_0 &\leq T_{M_l+1} + t_{M_l+2} + \cdots + t_{M_l+j+1} \\ &= T_{M_l+1} + \kappa_{M_l+1}|X_{T_{M_l+1}}| + \cdots + \kappa_{M_l+j}|X_{T_{M_l+j}}|.\end{aligned}$$

Now, because of Assumption 2.1 for any $h = 0, \dots, j-1$

$$\begin{aligned}|X_{T_{M_l+h+1}}| \leq |X_{T_{M_l+h}}| + \alpha t_{M_l+h+1} &= |X_{T_{M_l+h}}| + \alpha\kappa_{M_l+h}|X_{T_{M_l+h}}| \\ &= |X_{T_{M_l+h}}|(1 + \alpha\kappa_{M_l+h}) \\ &\leq \cdots \leq |X_{T_{M_l}}| \prod_{i=0}^{h}(1 + \alpha\kappa_{M_l+i}).\end{aligned}$$

We have that $1 \leq j \leq M_{l+1} - M_l \leq D$, due to Assumption I), and $\kappa_{M_{l+1}} \geq \kappa_{M_l+j} \geq 1$, then, we obtain for each event $\{T_{M_l+j} < \tau_0, \tau_0^{M_l+j} \leq t_{M_l+j+1}\}$ that

$$\begin{aligned}\tau_0 &\leq T_{M_l+1} + \kappa_{M_{l+1}}|X_{T_{M_l}}| \sum_{h=1}^{M_{l+1}-M_l} \prod_{i=0}^{h-1}(1 + \alpha\kappa_{M_{l+1}}) \\ &\leq T_{M_l+1} + \kappa_{M_{l+1}}|X_{T_{M_l}}| \frac{(1 + \alpha\kappa_{M_{l+1}})^{M_{l+1}-M_l} - 1}{\alpha\kappa_{M_{l+1}}} \\ &\leq T_{M_l+1} + (1+\alpha)^D |X_{T_{M_l}}| \kappa_{M_{l+1}}^D\end{aligned}$$

where in the last step we assume without loss of generality that $\alpha \geq 1$. Therefore, we conclude that in the event $\{T_{M_l+1} < \tau_0 \leq T_{M_{l+1}+1}\}$

$$\tau_0 \leq T_{M_l+1} + (1+\alpha)^D |X_{T_{M_l}}| \kappa_{M_{l+1}}^D.$$

In the event $\{\tau_0 > T_{M_l+1}\}$ we have by definition

$$T_{M_l+1} = \sum_{j=0}^{M_l} t_{j+1} = \sum_{j=0}^{M_l} \kappa_j |X_j|.$$

We conclude by arguing similarly as before, to obtain that

$$\begin{aligned}T_{M_l+1} &\leq \kappa_{M_l}|X_{T_{M_l}}| + \sum_{i=0}^{l-1} \sum_{j=M_i}^{M_{i+1}-1} \kappa_j |X_{T_j}| \\ &\leq \kappa_{M_l}|X_{T_{M_l}}| + \sum_{i=0}^{l-1} \kappa_{M_{i+1}} \sum_{j=M_i}^{M_{i+1}-1} |X_{T_j}| \\ &\leq \kappa_{M_l}|X_{T_{M_l}}| + \sum_{i=0}^{l-1} \kappa_{M_{i+1}}|X_{T_{M_i}}| \sum_{j=0}^{M_{i+1}-M_i-1} \prod_{h=0}^{j-1}(1 + \alpha\kappa_{M_i+h}) \\ &\leq \kappa_{M_l}|X_{T_{M_l}}| + \sum_{i=0}^{l-1} \kappa_{M_{i+1}}|X_{T_{M_i}}| \sum_{j=0}^{M_{i+1}-M_i-1} \prod_{h=0}^{j-1}(1 + \alpha\kappa_{M_{i+1}}) \\ &\leq \kappa_{M_l}|X_{T_{M_l}}| + \sum_{i=0}^{l-1} \kappa_{M_{i+1}}|X_{T_{M_i}}| \frac{(1 + \alpha\kappa_{M_{i+1}})^{M_{i+1}-M_i} - 1}{\alpha\kappa_{M_{i+1}}} \\ &\leq \kappa_{M_l}|X_{T_{M_l}}| + \sum_{i=0}^{l-1} |X_{T_{M_i}}|(1 + \alpha\kappa_{M_{i+1}})^{M_{i+1}-M_i} \\ &\leq \sum_{i=0}^{l} |X_{T_{M_i}}|(1 + \alpha\kappa_{M_i})^{M_{i+1}-M_i} \leq (1+\alpha)^D \sum_{i=0}^{l} |X_{T_{M_i}}| \kappa_{M_i}^D\end{aligned}$$

□

Claim D.3. *For all $x \in \mathcal{X}$ and $\sigma \geq 1$ we have $\mathbb{E}_x(\tau_0^\sigma) < +\infty$.*

Proof of Claim D.3. We show first that there is a.s. $l \geq 1$ such that $\tau_0 \leq T_{M_l+1}$. In particular, $\tau_0 < +\infty$ a.s. This follows by an application of Claim D.1 with $F = 1$, indeed

$$\begin{aligned}\mathbb{P}(\tau_0 > T_{M_l+1}) &\leq \sum_{\substack{x_1,\dots,x_l \neq 0 \\ c<k_1 \leq \cdots \leq k_l}} \exp\left(-u\sum_{i=1}^{l}(k_i - c)|x_i|\right)\mu_l(x_1,k_1,\dots,x_l,k_l) \\ &\leq \exp(-u\delta l)\end{aligned}$$

since $|x_i| \geq 1$ and where $\delta = \kappa_0 - c > 0$.

Therefore, $\sum_l \mathbb{P}(\tau_0 > T_{M_l+1}) < +\infty$ and by the Borel-Cantelli lemma we conclude that a.s. $\tau_0 \leq T_{M_{l+1}}$. □

We now obtain that

$$\begin{aligned}\mathbb{E}_x(\tau_0^\sigma) &= \mathbb{E}_x(\mathbb{1}(\tau_0 \leq T_{M_1+1})\tau_0^\sigma) + \mathbb{E}_x\left(\sum_{l=1}^{+\infty}\mathbb{1}(T_{M_l+1} < \tau_0 \leq T_{M_{l+1}+1})\tau_0^\sigma\right) \\ &\leq \mathbb{E}_x\left(\mathbb{1}(\tau_0 \leq T_{M_1+1})T_{M_1+1}^\sigma\right) \\ &\qquad + \sum_{l=1}^{+\infty}\mathbb{E}_x(\mathbb{1}(T_{M_l+1} < \tau_0 \leq T_{M_{l+1}+1})F(x, X_{T_{M_1}}, \kappa_{M_1}, \dots, X_{T_{M_l}}, \kappa_{M_l})^\sigma)\end{aligned}$$

where $F(x, x_1, k_1, \dots, x_l, k_l) = (1+\alpha)^D \sum_{i=0}^{l} |x_i| k_{i+1}^D$ because of Claim D.2.

Now, Claim D.1 implies

$$\begin{aligned}&\sum_{l=1}^{+\infty}\mathbb{E}_x\left(\mathbb{1}(T_{M_l+1} < \tau_0 \leq T_{M_{l+1}+1})F(x, X_{T_{M_1}}, \kappa_{M_1}, \dots, X_{T_{M_l}}, \kappa_{M_l})^\sigma\right) \\ &\leq \sum_{l=1}^{+\infty}\mathbb{E}_x\left(\mathbb{1}(\tau_0 > T_{M_l+1})F(x, X_{T_{M_1}}, \kappa_{M_1}, \dots, X_{T_{M_l}}, \kappa_{M_l})^\sigma\right) \\ &\leq \sum_{l=1}^{+\infty}\sum_{\substack{x_1,\dots,x_l \neq 0 \\ c<k_1 \leq \cdots \leq k_l}} \exp\left(-u\sum_{i=0}^{l}(k_i - c)|x_i|\right)F(x, x_1, k_1, \dots, x_l, k_l)^\sigma \mu_l(x_1, k_1, \dots, x_l, k_l).\end{aligned}$$

□

We now extend this last claim to subsequent visits to state 0.

Claim D.4. *Let $\tau_0(0) = 0$, $\tau_0(1) = \tau_0$ and for $h \geq 1$*

$$\tau_0(h+1) = \inf\{t > \tau_0(h) : X_t = 0\}.$$

Then, there is $\beta_0(\sigma) > 0$ such that

$$\mathbb{E}_x((\tau_0(h+1) - \tau_0(h))^\sigma) \leq \beta_0(\sigma) \text{ for all } h \geq 0.$$

Proof of Claim D.4. Subsequent to the time $\tau_0(1)$, the process enters an exploratory phase until it exits $\mathcal{E}$, a finite neighborhood of state 0. Let $\eta(1)$ be the time to exit $\mathcal{E}$, is not hard to see that $\eta(1)$ has all finite moments, since $\mathcal{E}$ is finite.

We argue then as follows. Either there is a visit to state 0 prior to time $\eta(1)$ or not. In the event where there was such a visit, then $\tau_0(2) - \tau_0(1) \leq \eta(1)$ has all required moments. Otherwise, in the event where there were no visits during this time, we consider $Y_n = X_{n+\tau_0(1)+\eta(1)}$. The process $(Y_n)_n$ can be seen as an instance of Algorithm 1 itself. Therefore, all previous claims can be applied to $(Y_n)_n$, so that in particular $\tau_0(2) - \tau_0(1) = \inf\{t > 0 : Y_t = 0\}$ has all required moments.

We conclude, by means of repeating this argument, that the increment $\tau_0(h+1) - \tau_0(h)$ possesses finite moments of all orders for all $h \geq 1$. □

Proof of Proposition D.1. For each $h \geq 0$, because of Assumption 2.1 and III), there is $q(x,a) > 0$ such that for each $h \geq 1$

$$\mathbb{P}(\inf\{t > \tau_0(h) : X_t = x, A_t = a\} \leq \tau_0(h+1)) \geq q(x,a) > 0.$$

This is, there is a lower bound, independent of h, of the probability of hitting (x,a) before hitting 0 again.

From this, with $\tau_{(x,a)}(1) := \inf\{t \geq 0 : X_t = x, A_t = a)\}$, we obtain

$$\begin{aligned}
\mathbb{E}\left(\tau^{\sigma}_{(x,a)}(1)\right) &\leq \sum_{j\geq 1}(1-q(x,a))^{j-1}\mathbb{E}\left(\tau_0(j)^{\sigma}\right) \\
&\leq \sum_{j\geq 1}(1-q(x,a))^{j-1}\mathbb{E}\left((\tau_0(j)-\tau_0(j-1)+\cdots+\tau_0(1)-\tau_0(0))^{\sigma}\right) \\
&\leq \sum_{j\geq 1}(1-q(x,a))^{j-1} j^{\sigma}\beta_0^{\max}(\sigma) < +\infty
\end{aligned}$$

where $\beta_0^{\max}(\sigma) \leq \max\{\beta_0(i)^{\sigma} : \text{ for } i \leq \sigma\}$. A similar computation can be performed for $\mathbb{E}\left((\tau_{(x,a)}(h+1)-\tau_{(x,a)}(h))^{\sigma}\right)$ for all $h \geq 1$. □

E Proof of Theorem 4.1

Fix any (x,a) and denote $\eta(l) = \tau_{(x,a)}(l)$ for any $l \geq 1$. These are the subsequent times where a visit to state x and a selection of action a takes place.

We will first show that there is $\delta > 0$ such that

$$\liminf_n \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}(X_i = x, A_i = a) > \delta \quad \text{a.s.} \tag{11}$$

According to Proposition D.1, a.s. $0 < \eta(l) < +\infty$ and $\eta(l) \to +\infty$ as $l \to +\infty$.

We argue first that

$$\liminf_n \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}(X_i = x, A_i = a) = \liminf_l \frac{l}{\eta(l+1)} \text{ a.s.} \tag{12}$$

Indeed, let

$$S_n = \frac{1}{n}\sum_{i=0}^{n-1} \mathbb{1}(X_i = x, A_i = a) = \frac{1}{n}\#\{i \leq n-1 : X_i = x, A_i = a\}.$$

We observe that $S_{\eta(l)+1} = \frac{l}{\eta(l)+1}$ by definition of $\eta(l)$. Now, the sequence S_n is nonincreasing from $n = \eta(l)+1$ to $n = \eta(l+1)$, since there are no visits to (x,a) from time $\eta(l)+1$ to $\eta(l+1)-1$. Then, from $n = \eta(l)+1$ to $n = \eta(l+1)$ the quantity S_n takes its lowest value $l/\eta(l+1)$ at $n = \eta_{l+1}$. From this we derive (12).

Let $e_i = \mathbb{E}(\eta(i+1)-\eta(i)\,|\,\mathcal{F}_{\eta(i)})$ and $\xi_i = \eta(i+1)-\eta(i)-e_i$ for each $i \geq 0$. We obtain from Proposition D.1 that a.s.

$$\begin{aligned}
&\mathbb{E}(\xi_i\,|\,\mathcal{F}_{\eta(i)}) = 0 \text{ and} \\
&\mathbb{E}(\xi_i^2\,|\,\mathcal{F}_{\eta(i)}) \leq \beta(2) := \beta_{(x,a)}(2) \text{ for all } i \geq 0.
\end{aligned}$$

We note that conditioning on $\mathcal{F}_{\eta(i)}$ introduces no additional difficulty, and that the reasoning from Proposition D.1 extends to this conditional expectation.

Therefore, by the Law of Large Numbers for Martingales (17, Chapter 7, Section 3, Corollary 2), we have a.s.

$$0 = \lim \frac{1}{l}\sum_{i=0}^{l-1}\xi_i = \lim \frac{\eta(l)}{l} - \frac{1}{l}\left(\sum_{i=0}^{l-1} e_i\right).$$

From this one can not claim that each term separately has a limit, but it follows that

$$\limsup_l \frac{\eta(l)}{l} = \limsup \frac{1}{l}\left(\sum_{i=0}^{l-1} e_i\right) \leq \beta(1) := \beta_{(x,a)}(1) < +\infty \quad \text{a.s.}$$

where we make use again of Proposition D.1. We finally conclude (11) since

$$\liminf \frac{l}{\eta(l)} \geq 1/\beta(1) > 0 \quad \text{a.s.}$$

Now with (11) we conclude, by the same arguments as in the proof of Proposition 4.1, that

$$\sum_{n\geq 0} \alpha_n(x,a) = +\infty \quad \text{a.s.}$$

Since (x,a) is arbitrary and $\mathcal{X}\times\mathcal{A}$ is countable, we deduce

$$\sum_{n\geq 0} \alpha_n(x,a) = +\infty \quad \text{for all } (x,a) \text{ a.s.}$$

□

F How to construct Λ

Theorem 4.1 requires that the set Λ contains policies in class $\Theta(c)$ (see Assumption I). In this appendix we address how to obtain such a subset of policies Λ in a model-free setting.

Our perspective is that many problems naturally admit a parametrized family of policies $\widehat{\Lambda}$. For such families, it is often reasonable to expect the existence of a "stability region" with nonempty interior, although its exact characterization may be unknown. Suppose that $\widehat{\Lambda}$ contains a (possibly small) open subset of policies belonging to the class $\Theta(c)$. In Appendix G we show that this is the case in many practical applications and we give sufficient conditions for a policy to pertain to class $\Theta(c)$.

Then, if a discretization Λ of $\widehat{\Lambda}$ is chosen to be sufficiently fine, so that Λ includes policies in $\Theta(c)$, we obtain a finite set Λ containing a policy in $\Theta(c)$. Thus, following Remark 4.1, Assumptions I-III are satisfied by a round-robin selection mechanism. If the discretization is not fine enough, it is not guaranteed that the resulting finite policy set contains a policy in $\Theta(c)$.

If the discretization is not fine enough, the resulting finite policy might not contain a policy in $\Theta(c)$. To address this, we could consider the following dynamic construction. Consider a sequence of discretizations $\Lambda^{(1)} \subset \Lambda^{(2)} \subset \cdots \subset \Lambda^{(h)} \subset \cdots \subset \widehat{\Lambda}$, where each $\Lambda^{(h)}$ is finite and the sequence $\{\Lambda^{(h)}\}_{h\geq 1}$ provides increasingly accurate approximations of $\widehat{\Lambda}$ as $h \to \infty$. Algorithm 1 could then be instantiated in the following manner: the agent initially performs a round-robin exploration over $\Lambda^{(1)}$ for a prescribed duration. If stability is not detected, i.e., there were no returns to ground-state 0, it proceeds to $\Lambda^{(2)}$, and so on, increasing h iteratively. This process continues until ground-state is first reached while the agent is executing a round-robin scheme over some $\Lambda^{(h)}$ with $h \geq 1$. Reaching state 0 is guaranteed with high probability because, for sufficiently large h, the set $\Lambda^{(h)}$ contains policies in $\Theta(c)$. At the first such occurrence, the agent halts the refinement of the discretization and subsequently continues to operate with the round-robin scheme restricted to $\Lambda^{(h)}$.

We leave for future work a rigorous mathematical analysis of the above scheme.

G Sufficient conditions for belonging in $\Theta(c)$

As we explain in Section 4.2, the class $\Theta(c)$ is natural to consider in our online setting. By definition, all policies in $\Theta(c)$ are stable, and in this appendix we provide settings and conditions under which stable policies further belong to class $\Theta(c)$. This further illustrates the interest of class $\Theta(c)$ when it comes to stabilizing a system.

G.1 Fluid limit criteria

We recall that the fluid limit is the (usually) deterministic trajectory obtained as the limit of a rescaled process: starting from state $|x|$, one rescales time by $|x|$ and space by $1/|x|$, and lets $|x| \to \infty$.

A widely applicable sufficient condition is the existence of a policy whose associated fluid limit hits 0 before time c (15, Chapter 9), which is closely related to condition (4) defining class $\Theta(c)$. A rigourous proof of this implication would need to be carried out case by case.

The condition for existence of fluid limits and its properties is a classical and well-studied problem, for instance in the case of queueing networks (15; 3), population processes, chemical reaction networks, and random graph explorations; see (5) for a comprehensive treatment and illustrative examples.

G.2 Density-dependent chains

A family of examples where belonging in $\Theta(c)$ can be ensured is that of density-dependent chains in the sense of Kurtz. Let $(X^N)_{N\geq 1}$ be the family with transitions $x \mapsto x + \ell/N$ at rate $N\beta_\ell(x)$, $\ell \in \mathcal{L}$ finite, with β_ℓ Lipschitz and bounded, and let $F(x) := \sum_{\ell\in\mathcal{L}} \ell\, \beta_\ell(x)$ be the associated vector field. Assume the mean-field ODE $\dot{x} = F(x)$ drains at uniform speed towards the ground state, i.e. $\langle F(x), \nabla f(x)\rangle \leq -\rho < 0$ for a norm-like f and all $x \neq 0$; this is a property of the *limit* and involves no N-dependent quantity. The transfer to the pre-limit chain is quantitative: by the exponential deviation bound for density-dependent chains (see (5, Theorem 4.3)), for every horizon T and every $\varepsilon > 0$,

$$\mathbb{P}\Big(\sup_{t\leq T} \big|X^N(t) - x(t)\big| \geq \varepsilon\Big) \;\leq\; C_1(T)\, e^{-C_2(T)\, N\varepsilon^2}, \tag{13}$$

where $x(\cdot)$ solves the ODE started at $X^N(0)$ and C_1, C_2 depend only on T, $\mathcal{L}$ and the Lipschitz constants of the β_ℓ. Using these bounds, one can derive a drift inequality

$$\mathbb{E}_x[f(X^N(T))] - f(x) \leq -\rho T/2,$$

and therefore $X^N \in \Theta(c)$ with $c = c_2/\rho + O(1)$

Observe that the uniformity is the operative point: the detection horizon of the Algorithm 1 scales as $c|x|$, so a constant degrading with the system size would defeat the procedure on precisely the large-N systems it is designed for; the mean-field structure guarantees instead that a single policy family and a single detection budget serve every system size $N \geq N_0$.

G.3 Drift criteria

In this appendix, we analyze conditions under which ergodic Markov chains in $\mathbb{Z}^d_{\geq 0}$ belong to class $\Theta(c)$ for some $c > 0$. Then, in Appendix H, we will apply Theorem G.1, to Markov chains arising in queueing examples that cover in particular the models considered in the numerics section of this work (Section 5).

Let $(X_n)_n$ be a Markov chain in $\mathbb{Z}^d_{\geq 0}$ for $d \geq 1$, irreducible and with bounded range $\alpha > 0$, i.e.

$$|X_{n+1} - X_n| \leq \alpha \text{ for all } n \geq 0 \text{ a.s.}$$

where $|\cdot|$ stands for the ℓ^1-norm of a vector, see also Assumption 2.1.

We define the drift of the Markov chain at state $x \in \mathbb{Z}^d_{\geq 0}$ by

$$d(x) := \mathbb{E}_x(X_1 - x) \in \mathbb{R}^d.$$

We now state the main result of this appendix.

Theorem G.1. *Assume that there is a vector v with non-negative entries and a constant $\gamma > 0$ such that*

$$\langle d(x), v\rangle \leq -\gamma \text{ for all } x \neq 0. \tag{14}$$

Then, the Markov chain is positive recurrent and moreover, there is $c > 0$, depending only on $|v|$,γ and α such that

$$\mathbb{P}_x(\tau_0 > (c+\delta)|x|) \leq \exp(-u\delta|x|) \text{ for all } x. \tag{15}$$

Suppose that $(X'_n)_n$ is another Markov chain with bounded range $\alpha' = \alpha$ and satisfying (14) *for $v' = v$ but $\gamma' > \gamma$. Then* (15) *also holds for this Markov chain with the same c and u.*

Proof of Theorem G.1. Let $f(x) = \langle x, v\rangle$, then for $x \neq 0$,

$$\mathbb{E}_x(f(X_1) - f(x)) = \langle d(x), v\rangle \leq -\gamma \text{ for all } x \neq 0. \tag{16}$$

Since $f(x) > 0$ for all $x \in \mathbb{Z}^d_{\geq 0} \setminus \{0\}$ together with (16) imply that the Markov chain is positive recurrent.

We also have that, since $||\cdot||_2 \leq |\cdot|$

$$f(x) - f(y) = \langle x - y, v\rangle \leq ||v||_2||x - y||_2 \leq |v||x - y|. \tag{17}$$

We conclude the proof of (15) by means of the following claim, adapted from Proposition 8.8 in (15). For the reader's convenience, and in view of the specific features of our setting, we reproduce the proof at the end of this subsection.

Claim: Consider $\lambda_1, \lambda_2 > 0$ such that if $h(\lambda) = -\gamma\lambda + (\lambda^2/2)\exp(|v|\alpha\lambda)$

$$-1 < h(\lambda_2) < -\gamma\lambda_2/2 \text{ and } \lambda_1 = -\log(1 - \gamma\lambda_2/2).$$

Then, for all x

$$\mathbb{E}_x(\exp(\lambda_1\tau_0)) \leq \exp(\lambda_2 f(x)) \tag{18}$$

Observe that λ_1 and λ_2 depend only on $|v|$, γ and α. Next, (17) gives $f(x) \leq |v||x|$. Thus, by Markov's inequality

$$\begin{aligned}\mathbb{P}_x(\tau_0 > (c+\delta)|x|) &\leq \frac{\mathbb{E}_x(\exp(\lambda_1\tau_0))}{\exp(\lambda_1(c+\delta)|x|)} \\ &\leq \frac{\exp(\lambda_2 f(x))}{\exp(\lambda_1(c+\delta)|x|)} \\ &\leq \exp(((\lambda_2|v| - \lambda_1 c)/\delta - \lambda_1)\delta|x|)\end{aligned}$$

Therefore, we conclude (15) for $a = \lambda_1$ and $\lambda_2|v| - \lambda_1 c < 0$, or equivalently,

$$c > |v|\frac{\lambda_2}{\lambda_1}. \tag{19}$$

where c and a depend only on $|v|$,γ and α.

□

Observe that if the Markov chain has bounded range α, then $\tau_0 \geq |x|/\alpha$ a.s. if $X_0 = x$. This indicates that we must obtain $c > 1/\alpha$ for c as in (15). From the proof of Theorem G.1, particularly, from (19), on the one hand we have,

$$\frac{\lambda_2}{\lambda_1} = \frac{\lambda_2}{-\log(1 - \gamma\lambda_2/2)} = \frac{2\lambda_2}{\gamma\lambda_2 g(\lambda_2)} = \frac{2}{\gamma g(\lambda_2)}$$

where $g(\lambda) := 1 - (\gamma\lambda/2)/2 + (\gamma\lambda/2)^2/3 - \ldots$. For λ_2 small enough we have $g(\lambda_2) \leq 2$. Thus,

$$\frac{\lambda_2}{\lambda_1} \geq \frac{1}{\gamma}.$$

On the other hand, (16) and (17) imply that $\gamma < |\langle d(x), v\rangle| \leq \alpha|v|$. Therefore,

$$c > |v|\frac{\lambda_2}{\lambda_1} \geq \frac{\gamma}{\alpha}\frac{1}{\gamma} = \frac{1}{\alpha}.$$

Proof of Claim (18)**.** Consider $Z_n = \exp(\lambda f(X_n)) \geq 1$ for all $n \geq 0$ and $\lambda > 0$. First observe that

$$\mathbb{E}(Z_{n+1} - Z_n \,|\, \mathcal{F}_n) = \mathbb{E}(\exp(\lambda(f(X_{n+1}) - f(X_n))) - 1 \,|\, \mathcal{F}_n)Z_n,$$

where $\mathcal{F}_n$ denotes the σ-algebra generated by $X_0, \ldots, X_n$.

Then, since $e^{\Delta} - 1 \leq \Delta + (\Delta^2/2)e^{|\Delta|}$, we obtain the bound

$$\begin{aligned}
&\mathbb{E}(\exp(\lambda(f(X_{n+1}) - f(X_n))) - 1 \,|\, \mathcal{F}_n) \\
&\leq \mathbb{E}(\lambda(f(X_{n+1}) - f(X_n)) \,|\, \mathcal{F}_n) \\
&+ \mathbb{E}((\lambda^2/2)(f(X_{n+1}) - f(X_n))^2 \exp(\lambda|f(X_{n+1}) - f(X_n)|) \,|\, \mathcal{F}_n).
\end{aligned}$$

In addition, because of (16), (17) and the bounded range assumption, in the event $\{\tau_0 > n\}$ the following inequality holds

$$\mathbb{E}(\exp(\lambda(f(X_{n+1}) - f(X_n))) - 1 \,|\, \mathcal{F}_n) \leq -\lambda\gamma + (\lambda^2/2)\exp(|v|\alpha\lambda).$$

Therefore, in the event $\{\tau_0 > n\}$ and for $h(\lambda) := -\lambda\gamma + (\lambda^2/2)\exp(|v|\alpha\lambda)$ we have

$$\mathbb{E}(Z_{n+1} \,|\, \mathcal{F}_n) \leq (1 + h(\lambda))Z_n.$$

Set $\lambda = \lambda_2 > 0$ such that $-1 < h(\lambda_2) < -\lambda\gamma/2$, and $\lambda_1 = -\log(1 - \lambda\gamma/2) > 0$. We obtain that, in the event $\{\tau_0 > n\}$,

$$\begin{aligned}
\mathbb{E}(\exp(\lambda_1(n+1))Z_{n+1} \,|\, \mathcal{F}_n) &\leq \exp(\lambda_1(n+1))\exp(-\lambda_1)Z_n \\
&\leq \exp(\lambda_1 n)Z_n.
\end{aligned}$$

Therefore, $(\exp(\lambda_1(n \wedge \tau_0)Z_{n\wedge\tau_0})_{n\geq 0}$ is a supermartingale, from where we derive that

$$\mathbb{E}(\exp(\lambda_1(n \wedge \tau_0)) \leq \mathbb{E}(\exp(\lambda_1(n \wedge \tau_0)Z_{n\wedge\tau_0}) \leq \mathbb{E}(Z_0) = \exp(\lambda_2 f(x)).$$

We complete the proof of Claim (18) by passing to the limit as $n \to \infty$ and invoking the monotone convergence theorem. □

H CLASS $\Theta(c)$ IN QUEUEING SYSTEM EXAMPLES

In this appendix we consider queueing models arising from load balancing and server allocation, we show that they admit a natural parametrized family of policies containing policies in class $\Theta(c)$. In particular these models cover the examples considered in the numerics, see Section 5.

In the queueing systems considered below, each $x = (x_1, \ldots, x_d) \in \mathbb{Z}_+^d$ indicates the lengths of the queues (how many jobs are in each queue) where $d \geq 1$ is the total number of queues in the system. Thus, the space of actions is finite $\mathcal{A} = \{0, 1, \ldots, M\}$. From the arrival and serving rates, and other parameters of each particular model, we derive discrete time transitions probabilities $p(y \,|\, x, a)$. The reward considered is typically $r(x) = -|x|$. Assumptions 2.1 and 2.2 are satisfied.

We denote the drift of action a at state x by

$$d_a(x) = \sum_y (y - x)p(y \,|\, x, a) \in \mathbb{R}^d.$$

Example 1: load balancing. Jobs arrive to a two-queue system with rate λ. Upon arrival, a job is sent by the dispatcher to queue 1 with probability $p \in [0, 1]$ and to queue 2 with probability $1 - p$. Jobs leave queue i with rate μ_i for $i = 1, 2$. Further, the dispatcher allows only two possibilities for the parameter p, namely $p_0, p_1 \in [0, 1]$, for assigning the job to queue 1. So the agent has only two available actions at each state, i.e. $\mathcal{A} = \{0, 1\}$.

For this example, the drift is constant on the interior states, the $x-$axis and $y-$axis. Let $x = (x_1, x_2)$, then

$$\begin{aligned}
d_{\text{interior},a} := d_a(x) &= \left(\frac{\lambda p_a - \mu_1}{\lambda + \mu_1 + \mu_2}, \frac{\lambda(1 - p_a) - \mu_2}{\lambda + \mu_1 + \mu_2}\right) \quad \text{for } x_1, x_2 > 0, \\
d_{x-\text{axis},a} := d_a(x) &= \left(\frac{\lambda p_a - \mu_1}{\lambda + \mu_1}, \frac{\lambda(1 - p_a)}{\lambda + \mu_1}\right) \quad \text{for } x_1 > 0 \text{ and } x_2 = 0, \\
d_{y-\text{axis},a} := d_a(x) &= \left(\frac{\lambda p_a}{\lambda + \mu_2}, \frac{\lambda(1 - p_a) - \mu_2}{\lambda + \mu_2}\right) \quad \text{for } x_1 = 0 \text{ and } x_2 > 0.
\end{aligned}$$

Consider $\varrho_i = \lambda/\mu_i$ for $i = 1, 2$. Assume that

$$\varrho_1 p_0 + \varrho_2(1 - p_0) < 1 \text{ and } \varrho_1 p_1 > 1 \tag{20}$$

Let π_a the policy that executes action a at every state x, for $a = 0, 1$. The first inequality of (20) implies that π_0 is stable, while the second inequality implies that π_1 is unstable.

Consider the family of policies $\Lambda = \{\pi_t := (1-t)\pi_0 + t\pi_1 \text{ s.t. } t \in [0,1]\}$.

Proposition H.1. *Assume* (20)*. Then, there are $t^* \in (0,1)$ and $c > 0$ such that $\pi_t \in \theta(c)$ for all $t < t^*$.*

Proof. Consider $v = (1, 1)$, because of the left-hand side of (20), we obtain

$$\langle w, v\rangle < 0 \text{ for } w = d_{\text{interior},0}, d_{x-\text{axis},0}, d_{y-\text{axis},0}.$$

Therefore, there is $\gamma > 0$ such that

$$\langle d_0(x), v\rangle < -\gamma \text{ for all } x \neq 0$$

where d_0 is the drift of the Markov chain associated to policy π_0.

Let $f(x) = \langle x, v\rangle$, then for $x \neq 0$,

$$\begin{aligned}
\mathbb{E}_x^{\pi_t}(f(X_1) - f(x)) &= (1-t)\langle d_0(x), v\rangle + t\langle d_1(x), v\rangle \\
&\leq -(1-t)\gamma + t|v| \qquad (21)\\
&= -(1-t)\gamma + 2t := -\gamma(t), \qquad (22)
\end{aligned}$$

where we used the fact that the range $\alpha = 1$, and that $|v| = 2$.

Finally, set $t^* = \frac{2}{\gamma+2} - \varepsilon_0 \in (0,1)$, for $\varepsilon_0 > 0$ small. Then, if $t < t^*$ we have $\gamma(t) > 0$, so that Theorem G.1 can be applied to the Markov chain defined by π_t for all $t < t^*$. Therefore we conclude the proposition by considering $c = c(t^*)$ and noticing that $\gamma(t)$ increases as $t \to 0$.

□

Example 2: Server allocation. Two types of jobs arrive at a single server with the same rate $\lambda > 0$. Jobs of type i arrive at queue i for $i = 0, 1$. The server chooses to serve a job from queue 1 (resp. queue 2) with probability p (resp. $1 - p$). The server does not serve an empty queue. If the selected job is of type i, it leaves the system with rate μ_i. As in the previous example, we assume the server has two possibilities for p, namely, $p_0, p_1 \in [0, 1]$, for selecting to serve a job from queue 1. So again, the agent has only two available actions at each state, i.e. $\mathcal{A} = \{0, 1\}$ (one one queue is empty both actions define the same transitions).

As for Example 1, the drift is constant on the interior states, the $x-$axis and $y-$axis. Let $x = (x_1, x_2)$, then

$$\begin{aligned}
d_{\text{interior},a} := d_a(x) &= \left(\frac{\lambda - \mu_1 p_a}{2\lambda + p_a\mu_1 + \mu_2(1-p_a)}, \frac{\lambda - \mu_2(1-p_a)}{2\lambda + p_a\mu_1 + \mu_2(1-p_a)}\right) \quad \text{for } x_1, x_2 > 0,\\
d_{x-\text{axis},a} := d_a(x) &= \left(\frac{\lambda - \mu_1}{2\lambda + \mu_1}, \frac{\lambda}{2\lambda + \mu_1}\right) \quad \text{for } x_1 > 0 \text{ and } x_2 = 0,\\
d_{y-\text{axis},a} := d_a(x) &= \left(\frac{\lambda}{2\lambda + \mu_2}, \frac{\lambda - \mu_2}{2\lambda + \mu_2}\right) \quad \text{for } x_1 = 0 \text{ and } x_2 > 0.
\end{aligned}$$

Consider again $\varrho_i = \lambda/\mu_i$ for $i = 1, 2$. Assume that

$$2\varrho_1 < p_0,\ 2\varrho_2 < 1 - p_0 \text{ and } \varrho_1 p_1 > 1 \tag{23}$$

As in the previous example, let π_a the policy that executes action a at every state x, for $a = 0, 1$. The two first inequalities of (23) imply that π_0 is stable, while the last inequality implies that π_1 is unstable. Consider also $\pi_t := (1-t)\pi_0 + t\pi_1$ such that $t \in [0, 1]$. The proof of the following proposition is analogous to the one of Proposition H.1.

Proposition H.2. *Assume* (23). *Then, there are* $t^* \in (0,1)$ *and* $c > 0$ *such that* $\pi_t \in \theta(c)$ *for all* $t < t^*$.

Example 3: Server allocation with state-dependent service capacity. The setting is the same as in Example 2. The main difference is that we assume $\mu = \mu_1 = \mu_2$ when both queues are busy while, the service capacity is $\widetilde{\mu}$ when one queue is empty. We also assume that $p_0 = 1$ and $p_1 = 0$, meaning that action 0 (resp. 1) serves queue 1 (resp. 2).

As before, the drift is constant on the interior states, the $x-$axis and $y-$axis. Let $x = (x_1, x_2)$, then

$$
\begin{aligned}
d_{\text{interior},0} := d_0(x) &= \left(\frac{\lambda-\mu}{2\lambda+\mu}, \frac{\lambda}{2\lambda+\mu}\right) \quad \text{for } x_1, x_2 > 0,\\
d_{\text{interior},1} := d_1(x) &= \left(\frac{\lambda}{2\lambda+\mu}, \frac{\lambda-\mu}{2\lambda+\mu}\right) \quad \text{for } x_1, x_2 > 0,\\
d_{x-\text{axis},a} := d_a(x) &= \left(\frac{\lambda-\widetilde{\mu}}{2\lambda+\widetilde{\mu}}, \frac{\lambda}{2\lambda+\widetilde{\mu}}\right) \quad \text{for } x_1 > 0 \text{ and } x_2 = 0,\\
d_{y-\text{axis},a} := d_a(x) &= \left(\frac{\lambda}{2\lambda+\widetilde{\mu}}, \frac{\lambda-\widetilde{\mu}}{2\lambda+\widetilde{\mu}}\right) \quad \text{for } x_1 = 0 \text{ and } x_2 > 0.
\end{aligned}
$$

Assume that

$$
\frac{\mu\widetilde{\mu}}{\mu+\widetilde{\mu}} > \lambda > \frac{\mu}{2} \tag{24}
$$

These inequalities imply that π_0 and π_1 are stable, but that π_t is unstable for some $t \in (0,1)$.

Proposition H.3. *There are* $0 < t_1^* < t_2^* < 1$ *and* c_0, c_1 *such that* $\pi_t \in \theta(c_0)$ *for all* $0 \le t \le t_1^*$ *and* $\pi_t \in \theta(c_1)$ *for all* $t_2^* \le t \le 1$.

Proof. We start by showing that there are $v = (v_1, v_2)$ with $v_i > 0$ and $\gamma > 0$ such that

$$
\langle d_0(x), v\rangle < -\gamma \text{ for } x \neq 0
$$

Once we stablish this inequality, to conclude, we repeat the same argument as for Proposition H.1. We also notice that by symmetry the same inequality, for different v and γ, can be achieved for d_1.

Consider the the lines perpendicular to $d_{\text{interior},0}, d_{y-\text{axis},0}$ and $d_{x-\text{axis},0}$, these are, respectively defined by

$$
\mathrm{l}_1 :\ y = \frac{\mu-\lambda}{\lambda}x, \quad \mathrm{l}_2 :\ y = \frac{\lambda}{\widetilde{\mu}-\lambda}x, \quad \mathrm{l}_3 :\ y = \frac{\widetilde{\mu}-\lambda}{\lambda}x.
$$

We need to show that the slope of l_1 is less than the one of l_3 but greater than the one of l_3. This is,

$$
\frac{\widetilde{\mu}-\lambda}{\lambda} > \frac{\mu-\lambda}{\lambda} > \frac{\lambda}{\widetilde{\mu}-\lambda},
$$

which is implied by (24). □