Trading-off Reward Maximization and Stability in Sequential Decision Making

Federico Corso
Politecnico di Milano
federico.corso@polimi.it

Marco Mussi
Politecnico di Milano
marco.mussi@polimi.it

Alberto Maria Metelli
Politecnico di Milano
albertomaria.metelli@polimi.it

Abstract

Reinforcement Learning (RL) focuses on learning policies that maximize the expected reward. This simple objective has enabled the success of RL in a wide range of scenarios. However, as emphasized by control-theoretic methods, stability is also a desired property when dealing with real-world systems. In this paper, we take a first step toward incorporating the notion of stability into RL. We focus on planning in *ergodic* Markov Decision Processes (MDPs), i.e., those that converge to a unique stationary distribution under any policy. We define the notion of stability in this context as the speed at which the induced Markov Chain (MC) converges to its stationary distribution. Noting that this property is connected to the spectral characteristics of the induced MC, we study the challenges of including a stability-related term in the RL objective function. First, we highlight how naïve approaches to trading off between reward maximization and stability lead to bilinear optimization programs, which are computationally demanding. Second, we propose an approach that bypasses this issue through a novel formulation and a surrogate objective function.

1 Introduction

Reinforcement Learning (RL, Sutton and Barto, 2018) is the branch of Artificial Intelligence (AI) aiming to solve sequential control problems. RL problems are usually formulated by means of the mathematical tool of Markov Decision Processes (MDPs, Puterman, 1994), where a problem is characterized by means of the possible states the environment can take, the actions that an agent can take, and the reward, i.e., a numerical signal to maximize through our interactions. In accordance with the *reward hypothesis* (Sutton, 2004), the AI community focused almost exclusively on performance maximization with respect to a chosen reward function (Buşoniu et al., 2018). In this context, numerous efforts have been made to study the convergence of RL algorithms to an optimal policy from a statistical perspective (Auer et al., 2008; Strehl et al., 2009). However, performance guarantees and reward maximization alone are often insufficient in safety-critical and high-stakes settings such as human-robot interaction or power plant control (Buşoniu et al., 2018), where additional properties of the learned policy, such as robustness or safety assurances, may also be required. In such contexts, traditional control-theory solutions are typically employed. Differently from RL, classical control objectives revolve around *stability* as this notion is closely related to the robustness, safety, and reliability of the systems under control (Slotine and Li, 1991; Lewis et al., 2012). Approaches that are similar to RL are known within the control community under the name of *adaptive/approximate dynamic programming* (Lewis et al., 2012; Lewis and Liu, 2013). In this context, even when optimal

18th European Workshop on Reinforcement Learning (EWRL 2025).

control approaches are used, the sole role of rewards is to represent stability requirements, such as in standard Model Predictive Control (MPC) problems, shadowing a possible trade-off between reward maximization and stability. For this reason, when formulating an optimal control problem, one cannot assume that rewards are bounded, as typically done in RL, and one can neither assume the presence of a discount factor (Postoyan et al., 2017), as rewards should reflect the behavior of the state, namely if the state grows arbitrarily.

Issues such as robustness, safety, and reliability of RL algorithms have been addressed within the context of *safe* RL (García and Fernández, 2015). Here, stability could be regarded as a specific notion of safety (Brunke et al., 2022). Safe RL problems are typically formulated using the framework of Constrained MDPs (CMDPs, Moldovan and Abbeel, 2012; Altman et al., 2019; Montenegro et al., 2024). In this context, some of the most promising results typically start with a known minimal, deterministic model of the system and use RL algorithms to learn a performance controller online. Stability during exploration is addressed leveraging the concept of *region of attraction* and the use of a *locally stabilizing* policy (Berkenkamp et al., 2017; Richards et al., 2018).

Control theory abounds in different specific notions of stability that vary, for instance, in how large the initial condition and subsequent states can be, in the speed of convergence to the asymptotic value (Khalil, 2014). Informally, stability requires that for bounded initial conditions (say, within the ball of radius δ around the origin), the system state remains bounded (say, within the ball of radius $\epsilon(\delta)$). The approach of the control community to the issue of stability in stochastic settings typically involves starting from the deterministic setting and handling stochasticity and uncertainty afterward, leveraging specific robustness tools and simplifications in order to infer stability in a proper sense, i.e., using the definitions and tools developed for the deterministic setting (Buşoniu et al., 2018).

Addressing the stability issues of RL algorithms from this same approach might not be ideal. Instead, stability should be interpreted and encoded in the problem in ways that are compatible with the classical mathematical framework of RL, i.e., the MDP formalism. Moreover, given the powerful nature of RL in solving pure optimization problems, we believe that a more natural way of dealing with stability guarantees in this context would be closer to the approach developed in the context of *economic* MPC (Ellis et al., 2014). Here, the primary objective is to design a controller for a dynamical system that minimizes a chosen economic cost function (which resembles the reward in RL). Such a function typically does not encode the task of regulating the system to a stable behavior, and as a consequence, the resulting optimal controllers typically lack any kind of stability guarantees. Necessary conditions to ensure the stability of the system have been developed in this field, and also methods that trade off between economic performance and stable behavior (Rawlings et al., 2012).

As mentioned in (Gros and Zanon, 2022), stability in MDPs can be arguably analyzed in the broader context of Markov Chains (MCs, Meyn and Tweedie, 2012). In this context, instead of thinking about equilibrium points for a dynamical system, we seek instead for an equilibrium measure, commonly called an *invariant measure*, satisfying the ergodic theorem, ensuring the asymptotic convergence to such a stationary measure. This implies the existence of a *steady state* for the induced stochastic process, much like global asymptotic stability for deterministic nonlinear state-space models.

Original Contribution. In this work, we will not be specifically concerned with the conditions that ensure convergence to the stationary measure; rather, we will address the problem of finding the policy that converges the fastest to its respective stationary distribution while maximizing the reward function. The contributions are summarized as follows:

- In Section 2, we introduce the finite discrete-time MDPs, and the assumptions needed in order to derive the approaches presented in the remainder of the paper.
- In Section 3, we present our problem formulation to incorporate a notion of stability in the objective function of RL, in the most generic setting.
- In Section 4, we provide a first naïve approach and discuss the computational limitations arising from the bilinear structure of the optimization problem involved.
- In Section 5, we propose our novel approach. In particular, we show how to bypass the bilinearity by reformulating the problem in a different space and introducing a new surrogate objective function. Moreover, we discuss the guarantees on the expected average reward loss and the stability properties of the learned objective.

Finally, in Appendix A, we also characterize a simple heuristic to achieve the above-mentioned objective.

2 Average Reward MDPs and Steady State

A finite, homogeneous, and discrete-time Markov decision process (Puterman, 1994) is defined as the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r}, \boldsymbol{\beta} \rangle$, where $\mathcal{S}$ denotes the finite state space; $\mathcal{A}$ denotes the finite action space; $\mathbf{P}$ denotes the transition matrix, whose entries, denoted as $p(s'|s,a)$, specifying the probability of transitioning to the state $s' \in \mathcal{S}$ when taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, namely $\mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)$; the reward function is denoted by the vector $\mathbf{r}$, whose entries, $r(s,a)$, denote the expected immediate reward when taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ that we assume bounded $|r(s,a)| \leq R_{\max} = 1$ for every $a \in \mathcal{A}$ and $s \in \mathcal{S}$; and $\boldsymbol{\beta}$ denotes the initial distribution vector over the state space, and $\beta(s)$ is the probability that the initial state is $s \in \mathcal{S}$. We denote with $\boldsymbol{\pi} \in \Pi^{MR}$ a stationary Markovian randomized policy, and we denote with $\Pi^{MD} \subset \Pi^{MR}$ the subset of stationary Markovian deterministic policies.[1] A randomized policy specifies the probability of taking an action from a given state, namely, $\pi(a|s) = \mathbb{P}(A_t = a | S_t = s)$. A policy induces over the MDP a Markov chain (Levin et al., 2017) defined by $\mathcal{M}^{\boldsymbol{\pi}} := \langle \mathcal{S}, \mathbf{P}^{\boldsymbol{\pi}}, \boldsymbol{\beta} \rangle$ where $p^{\boldsymbol{\pi}}(s'|s) = \sum_{a \in \mathcal{A}} \pi(a|s) p(s'|s,a)$ is the state transition model. Given the induced MC, we denote as $\mathbf{P}^{\boldsymbol{\pi},t} = (\mathbf{P}^{\boldsymbol{\pi}})^t$ the t-step transition matrix, from which we can compute recursively the state-distribution of the induced MC at time step t as $\boldsymbol{\eta}_t^{\boldsymbol{\pi}} = (\mathbf{P}^{\boldsymbol{\pi},t})^\top \boldsymbol{\beta}$, and from one step to another as $\boldsymbol{\eta}_t^{\boldsymbol{\pi}} = (\mathbf{P}^{\boldsymbol{\pi}})^\top \boldsymbol{\eta}_{t-1}^{\boldsymbol{\pi}}$. The (s,s') entry of $\mathbf{P}^{\boldsymbol{\pi},t}$ is denoted as $p^{\boldsymbol{\pi},t}(s'|s)$ and corresponds to the probability of landing in state s' at time step t starting from the initial state s, equivalently, it corresponds to $\eta_t^{\boldsymbol{\pi}}(s')$ assuming $\beta(s) = 1$.

When studying the limiting behavior of the MC induced by a policy $\boldsymbol{\pi}$, it is natural to choose as optimality criterion that of the *average expected reward*, which can be defined as $g^{\boldsymbol{\pi}} := \lim_{T \to +\infty} \frac{1}{T} \mathbb{E}_{\boldsymbol{\beta}}^{\boldsymbol{\pi}} [\sum_{t=1}^{T} r(S_t, A_t)]$, where T denotes the horizon length (Puterman, 1994, Chapter 8). For this limit to exist in a proper sense, one may assume the underlying MDP to be *ergodic*.

Assumption 2.1 (Ergodic or Recurrent MDP — Puterman 1994, Chapter 8.3)**.** *An MDP is* ergodic *or* recurrent *if for* any *policy* $\boldsymbol{\pi} \in \Pi^{MR}$ *the induced MC consists of a single recurrent class, meaning that it is possible to reach every state from any other state in a finite number of steps.*

The most important consequence of the ergodicity assumption is that any induced MC admits a unique *stationary state distribution*, or, more formally, for every policy $\boldsymbol{\pi} \in \Pi^{MR}$ the limit $\boldsymbol{\eta}^{\boldsymbol{\pi}} = \lim_{t \to +\infty} \boldsymbol{\eta}_t^{\boldsymbol{\pi}}$ exists. Moreover, this distribution is the unique that satisfies the *invariance equation* $\boldsymbol{\eta}^{\boldsymbol{\pi}} = (\mathbf{P}^{\boldsymbol{\pi}})^\top \boldsymbol{\eta}^{\boldsymbol{\pi}}$.

Our focus revolves around controlling the rate of convergence of the sequence $(\boldsymbol{\eta}_t^{\boldsymbol{\pi}})_{t \geq 1}$ to its limit $\boldsymbol{\eta}^{\boldsymbol{\pi}}$. It is known from the literature (Seabrook and Wiskott, 2023) that quantitative aspects of convergence, such as its rate, are strictly related to the spectrum of the transition matrix $\mathbf{P}^{\boldsymbol{\pi}}$. The spectrum of an $n \times n$ square matrix $\mathbf{A}$ is the multi-set of its eigenvalues, denoted by $\Lambda(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, counting algebraic multiplicity.[2] Being $\mathbf{P}^{\boldsymbol{\pi}}$ a row-stochastic matrix, i.e., a matrix in which every row sums to 1, by Perron-Frobenius Theorem (see Meyer, 2023, Chapter 8.2), eigenvalues of $\mathbf{P}^{\boldsymbol{\pi}}$ cannot have modulus greater than 1. We denote them in decreasing order of modulus as $\lambda_1^{\boldsymbol{\pi}} = 1 \geq |\lambda_2^{\boldsymbol{\pi}}| \geq \cdots \geq |\lambda_{|\mathcal{S}|}^{\boldsymbol{\pi}}|$. For an ergodic MC, the asymptotic rate of convergence to the stationary distribution is determined by $\mu(\mathbf{P}^{\boldsymbol{\pi}}) := |\lambda_2^{\boldsymbol{\pi}}|$, namely, the Second Largest Eigenvalue Modulus (SLEM) of $\mathbf{P}^{\boldsymbol{\pi}}$. Such a statement is formalized in the following theorem.

Theorem 2.1 (Spectral Conditions For Ergodicity — Meyn 2022, Theorem 6.2)**.** *Given the transition matrix* $P^{\boldsymbol{\pi}}$ *of the MC induced by a policy* $\boldsymbol{\pi}$*, suppose that* $\lambda_1^{\boldsymbol{\pi}} = 1$ *is the only eigenvalue satisfying* $|\lambda^{\boldsymbol{\pi}}| = 1$*. Then, the chain is ergodic, and the convergence rate of:*

$$\lim_{t \to +\infty} p^{\boldsymbol{\pi},t}(s' \mid s) = \eta^{\boldsymbol{\pi}}(s'), \quad \forall s, s' \in \mathcal{S}, \tag{1}$$

is geometric:

$$\lim_{t \to +\infty} \frac{1}{t} \log \left(\max_{s,s' \in \mathcal{S}} \left| p^{\boldsymbol{\pi},t}(s' \mid s) - \eta^{\boldsymbol{\pi}}(s') \right| \right) = \log \left(\mu(\mathbf{P}^{\boldsymbol{\pi}}) \right) < 0, \tag{2}$$

where $\mu(\mathbf{P}^{\boldsymbol{\pi}}) = |\lambda_2^{\boldsymbol{\pi}}|$ *is the SLEM.*

[1] From now on, we will omit the adjective "Markovian" when referring to a policy for the sake of simplicity, as we consider only this type of policies.

[2] For generic $n \times n$ matrices $\mathbf{A}$, each λ_i might belong to the set of complex numbers $\mathbb{C}$.

As stated in (Boyd et al., 2004), there are numerous indices that quantify the convergence rate of a MC, e.g., the *mixing rate*, the *mixing time*, and the *spectral gap*. Notably, most of them show a monotonically increasing dependence in the SLEM; thus, it is convenient to focus on that. Before proceeding, it is important to stress that most of the literature concerned with SLEM optimization is built around the assumption that the MC is reversible with respect to its own stationary distribution.

Definition 2.1 (Reversible Markov Chain — Levin et al. 2017)**.** *A MC is* reversible *if it satisfies the detailed balance equation with respect to its stationary distribution:*

$$\eta^{\boldsymbol{\pi}}(s')p^{\boldsymbol{\pi}}(s'|s) = \eta^{\boldsymbol{\pi}}(s)p^{\boldsymbol{\pi}}(s|s'), \quad \forall s, s'. \tag{3}$$

While SLEM optimization has been extensively studied in the MC literature, at least in the reversible case, in the context of MDPs, such a problem has only been introduced combined with other objectives, but never on its own (Tarbouriech and Lazaric, 2019; Mutti and Restelli, 2020).

Matrix Notation. At occurrence, the above-mentioned objects will be referred using a convenient matrix notation. Importantly, all vectors are intended as *column vectors*, and will be denoted in bold. Matrices will also be denoted with bold uppercase letters. The initial distribution $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{S}|}$ will be interpreted as a stochastic column-vector, whose components will be denoted as $\beta(s) = \mathbb{P}(S_0 = s)$. The transition model of the MDP will be denoted with $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$, a row-stochastic matrix representation of the transition kernel of the MDP. The $((s,a),s')$ component of such a matrix will be denoted with the usual $p(s' \mid s,a)$. The policy will be denoted with $\boldsymbol{\Pi} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$ and its components by $\boldsymbol{\Pi}(s,(s,a)) = \pi(a|s)$. Additionally, the transition matrix induced by policy π, will be written as $\mathbf{P}^{\boldsymbol{\pi}} = \boldsymbol{\Pi}\mathbf{P}$, a $|\mathcal{S}| \times |\mathcal{S}|$ row-stochastic matrix as well. We denote the (s,s') entry of this matrix as $p^{\pi}(s'|s)$. When dealing with the stationary distribution, we denote its minimum-valued entry as $\eta^{\boldsymbol{\pi}}_{\min} = \min_{s\in\mathcal{S}} \eta^{\boldsymbol{\pi}}(s)$. We use $\mathbf{D}_{\boldsymbol{\eta}} \coloneqq \operatorname{diag}(\boldsymbol{\eta})$ to indicate the diagonal matrix whose non-zero entries are the elements of vector $\boldsymbol{\eta}$. We use the symbol $\mathbb{S}_n$ for the set of $n \times n$ row-stochastic matrices. We represent the simplex over a set $\mathcal{X}$ by $\Delta(\mathcal{X})$, and when working with random quantities, we use uppercase letters to denote random variables, while we use lowercase letters to denote the values that these random elements take on. The outer product is denoted as $\otimes$, while the Hadamard product, i.e., the element-wise matrix multiplication, is denoted through the symbol $\odot$. The symbol $\preceq$ denotes matrix inequality, i.e., $\mathbf{X} \preceq \mathbf{Y}$ means $\mathbf{Y} - \mathbf{X}$ is positive semidefinite. The symbol $\leq$ instead denotes element-wise inequality. The orthogonal complement of a vector $\boldsymbol{v}$ is denoted as $\operatorname{range}(\boldsymbol{v})^{\perp}$. The symbol $\mathbf{1}$ denotes a column vector of all ones, while $\mathbf{I}$ represents the identity matrix.

3 Problem Formulation

In this work, we aim to find the policy $\boldsymbol{\pi}$ that, given the underlying MDP structure, achieves the best trade-off between average reward maximization and fastest rate of convergence toward its stationary distribution. In more formal terms, the above problem can be posed as the following optimization problem:[3]

$$\begin{array}{ll} \underset{\boldsymbol{\pi} \in \Pi^{MR}}{\text{maximize}} & g^{\boldsymbol{\pi}} - \mu(\mathbf{P}^{\boldsymbol{\pi}}). \end{array} \tag{4}$$

Here, $g^{\boldsymbol{\pi}}$ is a scalar denoting the average reward achieved under policy $\boldsymbol{\pi}$. Importantly, under the ergodicity assumption, such a scalar does not depend on the initial state (Puterman, 1994), and can be expressed as $g^{\boldsymbol{\pi}} = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \eta^{\boldsymbol{\pi}}(s)\pi(a \mid s)r(s,a)$. The latter, is a continuous function of $\boldsymbol{\pi}$, from continuity of the eigenvector $\boldsymbol{\eta}^{\boldsymbol{\pi}}$ (Lax, 2014, Theorem 8). Moreover, from perturbation theory (Kato, 2013, Chapter 2), it is known that every eigenvalue is a continuous function of the entries of its matrix. Thus, the function $\mu(\mathbf{P}^{\boldsymbol{\pi}})$ is a continuous function of the entries of $\mathbf{P}^{\boldsymbol{\pi}} = \boldsymbol{\Pi}\mathbf{P}$, thus continuous in π. Finally, the set $\Pi^{MR} = \prod_{s\in\mathcal{S}} \Delta_{|\mathcal{A}|}$ is compact. This allows us to conclude, by the *extreme-value theorem* (Rudin, 1976, Theorem 4.16), that our objective achieves both global maximum and minimum over Π^{MR}.

[3]For the sake of generality, one could consider a weight $w > 0$ in the objective ($g^{\boldsymbol{\pi}} - w\mu(\mathbf{P}^{\boldsymbol{\pi}})$) to trade off between reward maximization and stability. However, we are more interested in studying the properties of such a problem, which are not affected by the value of w. As such, for the sake of simplicity, we set $w = 1$.

4 Explicit Characterizations of the SLEM

In the following, Problem (4) is first made explicit through the standard spectral norm characterization of the SLEM under the reversibility assumption of $\mathbf{P}^{\boldsymbol{\pi}}$ (Boyd et al., 2004). Moreover, a possible characterization of the same quantity in the more general case of non-reversible $\mathbf{P}^{\boldsymbol{\pi}}$ is discussed.

While the asymptotic expected average reward $g^{\boldsymbol{\pi}}$ is easily made explicit using the standard dual formulation of the average reward objective (Puterman, 1994), expressing the SLEM requires more care.

Reversible Case. When it is assumed either that for every policy $\boldsymbol{\pi} \in \Pi^{MR}$ the induced MC is reversible, or when one explicitly restrict the policy space to the set of "reversible" policies, denoted as Π^{MRR}, the expression of $\mu(\mathbf{P}^{\boldsymbol{\pi}})$ simplifies.[4] Indeed, reversibility allows to work with the symmetric matrix $\mathbf{D}_{\boldsymbol{\eta}^{\boldsymbol{\pi}}}^{1/2}\mathbf{P}^{\boldsymbol{\pi}}\mathbf{D}_{\boldsymbol{\eta}^{\boldsymbol{\pi}}}^{-1/2}$, as the spectrum is invariant under similarity transformation, $\Lambda(\mathbf{D}_{\boldsymbol{\eta}^{\boldsymbol{\pi}}}^{1/2}\mathbf{P}^{\boldsymbol{\pi}}\mathbf{D}_{\boldsymbol{\eta}^{\boldsymbol{\pi}}}^{-1/2}) = \Lambda(\mathbf{P}^{\boldsymbol{\pi}})$. Because of symmetry, and knowing the dominant eigenvector $\sqrt{\boldsymbol{\eta}^{\boldsymbol{\pi}}}$, it is possible to express the SLEM through its variational characterization (Horn and Johnson, 2012, Chapter 4.2). Specifically to express it as the spectral norm of the matrix $\mathbf{D}_{\boldsymbol{\eta}^{\boldsymbol{\pi}}}^{1/2}P^{\boldsymbol{\pi}}\mathbf{D}_{\boldsymbol{\eta}^{\boldsymbol{\pi}}}^{-1/2}$ restricted to the orthogonal complement of the dominant eigenvector $\mathrm{range}(\sqrt{\boldsymbol{\eta}^{\boldsymbol{\pi}}})^{\perp}$:

$$\mu(\mathbf{P}^{\boldsymbol{\pi}}) = \left\| \mathbf{D}_{\boldsymbol{\eta}^{\boldsymbol{\pi}}}^{1/2}\mathbf{P}^{\boldsymbol{\pi}}\mathbf{D}_{\boldsymbol{\eta}^{\boldsymbol{\pi}}}^{-1/2} - \sqrt{\boldsymbol{\eta}^{\boldsymbol{\pi}}}\sqrt{\boldsymbol{\eta}^{\boldsymbol{\pi}}}^{\top} \right\|_2 . \tag{5}$$

Non-Reversible Case. When the induced MC is non-reversible, to the best of our knowledge, no exact and explicit expression of the SLEM exists. The most reasonable workaround is then to leverage *ergodicity coefficients* (Ipsen and Selee, 2011). These mathematical objects have been introduced to estimate how fast inhomogeneous products of irreducible stochastic matrices converge to a rank-one matrix. As a consequence, these "coefficients" help in providing approximate but explicit expressions of the set of subdominant eigenvalues, $\{|\lambda_i| : i \geq 2\}$, of irreducible row-stochastic matrices.

Definition 4.1 (p-norm ergodicity coefficient — Ipsen and Selee 2011)**.** *For any integer $p \geq 1$ the p-norm ergodicity coefficient of a matrix $\mathbf{S}$ is defined as:*

$$\tau_p(\mathbf{1}, \mathbf{S}) \coloneqq \max_{\substack{\|\mathbf{z}\|_p = 1 \\ \mathbf{z}^{\top}\mathbf{1} = 0}} \|\mathbf{S}^{\top}\mathbf{z}\|_p, \tag{6}$$

where the maximum ranges over $\mathbf{z} \in \mathbb{R}^n$.

The reader is referred to the original work (Ipsen and Selee, 2011, Theorem 5.1) for the most relevant properties of this object. Here, only the following useful results are restated.

Theorem 4.1 (Ipsen and Selee 2011, Theorem 6.21)**.** *Let $\mathbf{S} \in \mathbb{S}_n$ be an $n \times n$ row-stochastic irreducible matrix with eigenvalues $\Lambda(\mathbf{S})$ and right eigenvector $\mathbf{1}$ so that $\mathbf{S}\mathbf{1} = \mathbf{1}$, and $\lambda_1 = 1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Then, the ergodicity coefficient associated with the eigenvector $\mathbf{1}$ satisfying:*

$$|\lambda_i| \leq \tau_p(\mathbf{1}, \mathbf{S}), \quad \forall i \in \{2, \ldots, n\}. \tag{7}$$

Moreover, from (Ipsen and Selee, 2011, Theorem 6.23), defining $\mathbf{D}_{\boldsymbol{\eta}} \coloneqq \mathrm{diag}(\boldsymbol{\eta})$, with $\mathbf{S}^{\top}\boldsymbol{\eta} = \boldsymbol{\eta}$, it holds:

$$|\lambda_i| \leq \tau_p(\mathbf{1}, \mathbf{D}_{\boldsymbol{\eta}}^{-1}\mathbf{S}\mathbf{D}_{\boldsymbol{\eta}}), \quad \forall i \in \{2, \ldots, n\}. \tag{8}$$

Importantly, as stated in (Ipsen and Selee, 2011), Equation (8) for stochastic matrices typically offers tighter bounds than Equation (7).

Finally, it is possible to provide inclusion intervals containing general p-norm ergodicity coefficients.

Corollary 4.1 (Ipsen and Selee 2011, Corollary 6.25)**.** *If $\mathbf{S} \in \mathbb{S}_n$ is an irreducible row-stochastic matrix, and $\boldsymbol{\eta}^{\top}\mathbf{S} = \boldsymbol{\eta}^{\top}$, then:*

$$|\lambda_i| \leq \tau_p(\mathbf{1}, \mathbf{S}) \leq \|(\mathbf{S} - \mathbf{1}\boldsymbol{\eta}^{\top})\|_p, \quad \forall i \in \{2, \ldots, n\}. \tag{9}$$

[4] We refer to a "reversible" policy as one that induces a reversible transition matrix $\mathbf{P}^{\boldsymbol{\pi}}$.

As shown in (Ipsen and Selee, 2011), the result of Corollary 4.1 implies the following inclusion interval for $\tau_p(\mathbf{1}, \mathbf{S})$ in terms of the matrix $\mathbf{S}$ deflated by its dominant spectral projector:

$$\rho(\mathbf{S} - \mathbf{1}\boldsymbol{\eta}^\top) \leq \tau_p(\mathbf{1}, \mathbf{S}) \leq \|\mathbf{S} - \mathbf{1}\boldsymbol{\eta}^\top\|_p, \tag{10}$$

where $\rho(\mathbf{S}) = \max_{i \in \{1,\dots,n\}} |\lambda_i|$ is the spectral radius of matrix $\mathbf{S}$. From the above statements, we deduce the usefulness of ergodicity coefficients in bounding the SLEM of the irreducible MC.

Explicit Expressions for $p \in \{1, 2\}$. In the specific instances when $p = 1$ or $p = 2$, we can retrieve the following convenient explicit expressions of the ergodicity coefficients (Ipsen and Selee, 2011):

$$\tau_1(\mathbf{1}, \mathbf{S}) = \min_{i,j \in \{1,\dots,n\}} \|S(i,\cdot) - S(j,\cdot)\|_{\text{TV}} = 1 - \min_{i,j \in \{1,\dots,n\}} \sum_{k=1}^{n} \min\{S(i,k), S(j,k)\}, \tag{11}$$

$$\tau_2(\mathbf{1}, \mathbf{S}) = \left\| \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \right) \mathbf{S} \right\|_2. \tag{12}$$

For the 2-norm coefficient, we also restate the following result.

Corollary 4.2 (Ipsen and Selee 2011, Corollary 6.20)**.** *Let $\mathbf{S} \in \mathbb{S}_n$ be an irreducible row-stochastic matrix with singular values $\sigma_1(\mathbf{S}) \geq \sigma_2(\mathbf{S}) \geq \cdots \geq \sigma_n(\mathbf{S})$ and dominant right and left singular vectors $\mathbf{v}$ and $\mathbf{u}$ respectively, that is:*

$$\mathbf{S}\mathbf{v} = \sigma_1(\mathbf{S})\mathbf{u} \quad \mathbf{S}^\top\mathbf{u} = \sigma_1(\mathbf{S})\mathbf{v}, \tag{13}$$

also assume that the euclidean norm satisfies $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$. Then:

$$\tau_2(\mathbf{u}, \mathbf{S}) = \tau_2(\mathbf{v}, \mathbf{S}^\top) = \sigma_2(\mathbf{S}). \tag{14}$$

From Equation (14) and the *Courant-Fischer theorem* (see Horn and Johnson, 2012, Theorem 4.2.6), applied to the matrix $\mathbf{S}^\top\mathbf{S}$, we deduce that $\tau_2(\mathbf{1}, \mathbf{S}) \geq \sigma_2(\mathbf{S})$. Meaning that, from Equation (10), the following inequalities hold: $\mu(\mathbf{S}) = \rho(\mathbf{S} - \mathbf{1}\boldsymbol{\eta}^\top) \leq \sigma_2(\mathbf{S}) \leq \tau_2(\mathbf{1}, \mathbf{S}) \leq \|\mathbf{S} - \mathbf{1}\boldsymbol{\eta}^\top\|_2$. In general, we have that $\mu(\mathbf{S}) = \|\mathbf{S} - \mathbf{1}\boldsymbol{\eta}^\top\|_2$ only when the pair $(\mathbf{1}, \boldsymbol{\eta})$ coincides with the top singular pair, namely when matrix $\mathbf{S}$ is normal. One such case is when the matrix $\mathbf{S}$ is the transition matrix of a reversible MC.

4.1 Explicit Optimization Problem in the Reversible Case: Bilinearity Issues

For the sake of simplicity and exposition, Problem (4) is addressed first by restricting the search to the class of reversible Markovian policies Π^{MRR}. In the literature (Boyd et al., 2004), the SLEM minimization problem has usually been posed starting from a known *target* stationary distribution $\boldsymbol{\eta}$, and either the policy or the MC converging as fast as possible to this target distribution has been searched for (Boyd et al., 2004; Tarbouriech and Lazaric, 2019; Mutti and Restelli, 2020). In this work, instead, the target distribution is not given, but it is considered as a decision variable. In fact, it is known that, in a given MDP, restricting to a target $\boldsymbol{\eta}$ beforehand might inevitably result in arbitrarily slow mixing (Tarbouriech and Lazaric, 2019) for any policy. The explicit version of Problem (4) is then:

$$\begin{aligned} \underset{\boldsymbol{\Pi} \in \mathbb{S}_{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}, \boldsymbol{\eta} \in \Delta(\mathcal{S})}{\text{maximize}} \quad & \boldsymbol{\eta}^\top (\boldsymbol{\Pi}\mathbf{r}) - \left\| \mathbf{D}_{\boldsymbol{\eta}}^{1/2} (\boldsymbol{\Pi}\mathbf{P}) \mathbf{D}_{\boldsymbol{\eta}}^{-1/2} - \sqrt{\boldsymbol{\eta}}\sqrt{\boldsymbol{\eta}}^\top \right\|_2 \\ \text{subject to} \quad & \boldsymbol{\eta} = (\boldsymbol{\Pi}\mathbf{P})^\top \boldsymbol{\eta}, \\ & \mathbf{D}_{\boldsymbol{\eta}} (\boldsymbol{\Pi}\mathbf{P}) = (\boldsymbol{\Pi}\mathbf{P})^\top \mathbf{D}_{\boldsymbol{\eta}} \end{aligned} \tag{15}$$

The major issue of the above problem is the presence of *bilinear dependencies* between the optimization variables in both the objective function and constraints. Specifically, despite the problem being convex in $\boldsymbol{\Pi}$ once $\boldsymbol{\eta}$ is fixed, and vice versa, it is not jointly convex. Moreover, fixing $\boldsymbol{\pi}$ implies fixing $\mathbf{P}^{\boldsymbol{\pi}}$ and thus $\boldsymbol{\eta}$ because of the stationary constraints, ruling out any possibility of using alternate minimization frameworks, as observed in (Tarbouriech and Lazaric, 2019). In the following section, Problem (15) is modified to obtain a surrogate showing convexity and other amenable properties to most common convex optimization methods.

5 Bypassing Bilinearity: A Surrogate Objective in the $\mathcal{S} \times \mathcal{A}$ Space

Drawing inspiration from the approach used in (Tarbouriech and Lazaric, 2019), one might try to address the optimization Problem (15) by treating the transition matrix $\mathbf{P}^{\boldsymbol{\pi}} = \boldsymbol{\Pi}\mathbf{P}$ as the transition matrix $\mathbf{M}$ of a generic MC and remove the bilinearity of the reversibility constraints by introducing the auxiliary variable $\mathbf{X} = \mathbf{D}_{\boldsymbol{\eta}}\mathbf{M}$. This allows reducing the problem to a setting closer to that exposed in (Boyd et al., 2004). Unfortunately, this strategy is not enough in the setting studied in this paper, as the presence of the average expected reward term, namely $\boldsymbol{\eta}^\top(\boldsymbol{\Pi}\mathbf{r})$, makes this simple change of variable ineffective because of the variable $\boldsymbol{\Pi}$.[5] A viable alternative leveraged in this work consists in framing Problem (15) in terms of the *joint state-action* MC, $(S_t, A_t)_{t\geq 0}$.

The Bivariate Process over $\mathcal{S} \times \mathcal{A}$. The most straightforward method to resolve the issue generated by the average reward term $\boldsymbol{\eta}^\top\boldsymbol{\Pi}\mathbf{r}$ is to start from the classical dual program formulation of the average reward criterion (Puterman, 1994). The problem is thus "lifted" from distributions over $\mathcal{S}$ to the distributions over the product space $\mathcal{S} \times \mathcal{A}$. In doing so, let $Z_t = (S_t, A_t)$ denote a random vector defined over $\mathcal{S} \times \mathcal{A}$. The sequence $(Z_t)_{t\geq 0}$ is the MC representing the sequences of state and action pairs generated by policy π executed on the MDP $\mathcal{M}$. We denote with $\mathbf{T} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ the transition matrix of this MC, with components $\mathbf{T}((s,a),(s',a')) = \mathbb{P}(S_{t+1} = s', A_{t+1} = a' \mid S_t = s, A_t = a) = \pi(a' \mid s')p(s' \mid s, a)$. Denote with $\mathbf{x} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ the stationary distribution of the chain, whose components are given by $x(s,a)$. Importantly, $\mathbf{x} = \boldsymbol{\eta}^\top\boldsymbol{\Pi}$, thus allowing to remove the bilinear term associated with the expected average reward term. Also, the following relationship holds: $\mathbf{T} = \mathbf{P}\boldsymbol{\Pi}$.

Relationships between the spectrum of $\mathbf{P}^{\boldsymbol{\pi}}$ and $\mathbf{T}$. In translating Problem (15) in terms of the transition matrix $\mathbf{T}$, it is helpful to leverage the result from (Horn and Johnson, 2012, Theorem 1.3.22). Such a theorem allows to relate the spectra $\Lambda(\mathbf{P}\boldsymbol{\Pi})$ and $\Lambda(\boldsymbol{\Pi}\mathbf{P})$. Specifically it allows to conclude that the matrix $\mathbf{T}$ has $|\mathcal{S}||\mathcal{A}|$ eigenvalues, $|\mathcal{S}|$ of which coincide with the eigenvalues of $\mathbf{P}^{\boldsymbol{\pi}} = \boldsymbol{\Pi}\mathbf{P}$, while the other $|\mathcal{S}||\mathcal{A}| - |\mathcal{S}|$ are zero-valued. More formally, $\Lambda(\mathbf{T}) = \Lambda(\mathbf{P}^{\boldsymbol{\pi}}) \cup \{0\}^{|\mathcal{S}||\mathcal{A}|-|\mathcal{S}|}$. Importantly, this allows us to conclude that $\mu(\mathbf{T}) = \mu(\mathbf{P}^{\boldsymbol{\pi}})$, and optimizing the SLEM of $\mathbf{T}$ is equivalent to optimizing the SLEM of $\mathbf{P}^{\boldsymbol{\pi}}$.

A Heuristic Surrogate Optimization Problem. Wishful thinking would lead one to account for the mixing behavior of matrix $\mathbf{T}$ by replacing the term $\|\mathbf{D}_{\boldsymbol{\eta}}^{1/2}(\boldsymbol{\Pi}\mathbf{P})\mathbf{D}_{\boldsymbol{\eta}}^{-1/2} - \sqrt{\boldsymbol{\eta}}\sqrt{\boldsymbol{\eta}}^\top\|_2$ with the corresponding $\|\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2} - \sqrt{\mathbf{x}}\sqrt{\mathbf{x}}^\top\|_2$. Such a substitution is indeed sensible and, most importantly, coherent with the objectives of this paper after observing the following facts:

- If the matrix $\mathbf{T}$ is reversible, then from the equality $\mu(\mathbf{P}^{\boldsymbol{\pi}}) = \mu(\mathbf{T})$, the substitution follows immediately from the spectral norm characterization of the dominant eigenvalue. Unfortunately, requiring matrix $\mathbf{T}$ to be reversible is much more restrictive and unrealistic than the reversibility of $\mathbf{P}^{\boldsymbol{\pi}}$, as it would require a constraint of the type $\mathbf{D}_{\boldsymbol{\eta}}\mathbf{P}^a = (\mathbf{P}^{a'})^\top\mathbf{D}_{\boldsymbol{\eta}} \;\; \forall a, a' \in \mathcal{A}$.
- Being $\sqrt{\mathbf{x}}$ both a right and left dominant singular vector of matrix $\mathbf{T}$, from (Ipsen and Selee, 2011, Corollary 6.20), one can notice that $\tau_2(\sqrt{\mathbf{x}}, \mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2}) = \sigma_2(\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2})$. More explicitly, the spectral norm satisfies $\|\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2} - \sqrt{\mathbf{x}}\sqrt{\mathbf{x}}^\top\|_2 = \tau_2(\sqrt{\mathbf{x}}, \mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2}) = \sigma_2(\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2})$.
- Finally, one can directly relate the SLEM of matrix $\mathbf{D}_{\boldsymbol{\eta}}^{1/2}(\boldsymbol{\Pi}\mathbf{P})\mathbf{D}_{\boldsymbol{\eta}}^{-1/2}$ to $\sigma_2(\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2})$ through the following series of inequalities:

$$\begin{aligned} |\lambda_2(\mathbf{P}^{\boldsymbol{\pi}})| = |\lambda_2(\mathbf{T})| \overset{(a)}{=} |\lambda_2(\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2})| \overset{(b)}{=} |\lambda_1(\tilde{\mathbf{P}}(\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2})\tilde{\mathbf{P}})| \leq \\ \leq \sigma_1(\tilde{\mathbf{P}}(\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2})\tilde{\mathbf{P}}) = \sigma_2(\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2}). \end{aligned} \tag{16}$$

  Here, $\tilde{\mathbf{P}} = (\mathbf{I} - \sqrt{\mathbf{x}}\sqrt{\mathbf{x}}^\top)$ is the matrix projecting onto $\mathrm{range}(\sqrt{\mathbf{x}})^\perp$, equality (a) comes from the properties of similarity transformations, and (b) comes from (Ding and Zhou, 2007, Theorem 2.1).

It follows that $\mu(\mathbf{P}^{\boldsymbol{\pi}}) \leq \sigma_2(\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2})$, and the objective function in Problem (15) can be replaced by the lower bound:

$$\mathcal{L}(\mathbf{x}, \mathbf{T}) \coloneqq \mathbf{x}^\top\mathbf{r} - \sigma_2(\mathbf{D}_{\mathbf{x}}^{1/2}\mathbf{T}\mathbf{D}_{\mathbf{x}}^{-1/2}) \leq \boldsymbol{\eta}^\top\boldsymbol{\Pi}\mathbf{r} - \mu(\mathbf{P}^{\boldsymbol{\pi}}). \tag{17}$$

[5] In (Tarbouriech and Lazaric, 2019), the trick is possible since the objective does not depend explicitly on $\boldsymbol{\Pi}$.

Lifting the Constraints. The constraints of Problem (15) can be rewritten in terms of the variables $\mathbf{T}$ and $\mathbf{x}$ as well. First, stationarity is ensured by imposing $\mathbf{x} = \mathbf{T}^\top \mathbf{x}$. Notably, the surrogate objective represents a lower bound that does not depend on the particular explicit representation of $\mu(\mathbf{P}^{\boldsymbol{\pi}})$; consequently, it is possible to drop the reversibility constraints on $\mathbf{P}^{\boldsymbol{\pi}}$. This allows not only to remove a source of bilinearity of the problem, but also allows to generalize the approach to non-reversible MCs. Finally, the adjacency constraints can be "lifted" and expressed component-wise as $p(s' \mid s, a) = \sum_{a'} T(s', a' \mid s, a),\ \forall (s, a, s')$. Notably, all constraints remain linear in the decision variables after "lifting".

Solution. Even after formulating the optimization Problem (4) in terms of $\mathbf{T}$ and $\mathbf{x}$, the bilinear terms associated with the stationary constraint remain. Taking inspiration from (Tarbouriech and Lazaric, 2019), the auxiliary variable $\mathbf{X} = \mathbf{D}_{\mathbf{x}}\mathbf{T}$ is introduced. Problem (15) then becomes:

$$\begin{aligned} \underset{\mathbf{X} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}, \mathbf{x} \in \Delta(\mathcal{S} \times \mathcal{A})}{\text{maximize}} \quad & (\mathbf{X}\mathbf{1})^\top \mathbf{r} - \left\| \mathbf{D}_{\mathbf{x}}^{-1/2} \mathbf{X} \mathbf{D}_{\mathbf{x}}^{-1/2} - \sqrt{\mathbf{x}}\sqrt{\mathbf{x}}^\top \right\|_2 \\ \text{subject to} \quad & \mathbf{X}\mathbf{1} = \mathbf{x}, \\ & \mathbf{D}_{\mathbf{x}}\mathbf{P} = \mathbf{X}\mathbf{K}^\top, \\ & \mathbf{X} \geq 0 \end{aligned} \tag{18}$$

Here, matrix $\mathbf{K} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|}$ represent an "action-aggregation" matrix, defined as:

$$K(s, (s', a)) \begin{cases} 1 & \text{if } s' = s \\ 0 & \text{otherwise} \end{cases}. \tag{19}$$

Despite circumventing bilinearity in the objective and in the constraints, the decision variables $\mathbf{X}, \mathbf{x}$ remain tightly coupled by the equality constraints of Problem (18). Specifically, these constraints do not allow to optimize independently with respect to each variable. Such an issue can be resolved heuristically by fixing $\mathbf{x} = \mathbf{x}^\star$, namely the solution of the classical average reward problem. This fix requires to relax the steady state constraint, which is thus replaced with $\|\mathbf{X}\mathbf{1} - \mathbf{x}^\star\|_2^2 \leq \delta^2$, where the slack variable δ allows the stationary distribution associated with X, namely $\mathbf{X}\mathbf{1} = \mathbf{x}$ to be close to $\mathbf{x}^\star$ but not fixed to satisfy $\mathbf{X}\mathbf{1} = \mathbf{x}^\star$. Finally, the adjacency constraints, $\mathbf{D}_{\mathbf{x}}\mathbf{P} = \mathbf{X}\mathbf{K}^\top$, are rewritten to ensure that the solution $\mathbf{X}^\dagger$ to the resulting problem satisfies the underlying transition kernel $\mathbf{P}$. For these, the following convenient matrix form is available, $(\mathbf{I}_{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|} \odot (\mathbf{X}\mathbf{1})\mathbf{1}^\top)\mathbf{P} = \mathbf{X}\mathbf{K}^\top$. The whole problem then becomes an optimization problem with respect to $\mathbf{X}$, and can be rewritten as follows:

$$\begin{aligned} \underset{\mathbf{X} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}}{\text{maximize}} \quad & (\mathbf{X}\mathbf{1})^\top \mathbf{r} - \left\| \mathbf{D}_{\mathbf{x}^\star}^{-1/2} \mathbf{X} \mathbf{D}_{\mathbf{x}^\star}^{-1/2} - \sqrt{\mathbf{x}^\star}\sqrt{\mathbf{x}^\star}^\top \right\|_2 \\ \text{subject to} \quad & \mathbf{1}^\top \mathbf{X} \mathbf{1} = 1, \\ & (\mathbf{I} \odot (\mathbf{X}\mathbf{1})\mathbf{1}^\top)\mathbf{P} = \mathbf{X}\mathbf{K}, \\ & \|\mathbf{X}\mathbf{1} - \mathbf{x}^\star\|_2^2 \leq \delta^2, \\ & \mathbf{X} \geq 0 \end{aligned} \tag{20}$$

which is concave in $\mathbf{X}$.[6] Solving the problem above yields an optimal $\mathbf{X}^\dagger$ such that, $\mathbf{x}^\dagger = \mathbf{X}^\dagger \mathbf{1}$ is the stationary distribution of the MC with transition matrix $T^\dagger(s', a' \mid s, a) = \frac{X^\dagger((s,a),(s',a'))}{(\mathbf{X}^\dagger \mathbf{1})(s,a)}$. One can thus retrieve the optimal policy with the usual formula (Puterman, 1994, Theorem 8.8.2), $\pi^{\mathbf{x}^\dagger}(a \mid s) \coloneqq \frac{x^\dagger(s,a)}{\sum_{a \in \mathcal{A}} x^\dagger(s,a)}$.

Theorem 5.1. *Let $\boldsymbol{\pi}^\dagger \in \Pi^{MR}$ be a solution of the Problem* (20)*. Then, it holds:*

$$\langle \mathbf{x}^\star - \mathbf{x}^\dagger; \mathbf{r} \rangle \leq \delta R_{\max} \sqrt{|\mathcal{S}||\mathcal{A}|}, \tag{21}$$

and:

$$\sigma_2(\mathbf{D}_{\mathbf{x}^\star}^{1/2} \mathbf{P} \boldsymbol{\Pi}^\dagger \mathbf{D}_{\mathbf{x}^\star}^{-1/2}) \leq \sigma_2(\mathbf{D}_{\mathbf{x}^\star}^{1/2} \mathbf{P} \boldsymbol{\Pi}^\star \mathbf{D}_{\mathbf{x}^\star}^{-1/2}). \tag{22}$$

[6] From the point of view of the implementation, to avoid $\mathbf{D}_{\mathbf{x}^\star}^{-1/2}$ being ill-conditioned, we might use the state-action distribution $\mathbf{x}_\epsilon = (1 - \epsilon)\mathbf{x}^\star + \frac{\epsilon}{|\mathcal{S}||\mathcal{A}|}\mathbf{1}$, in place of $\mathbf{x}^\star$, with ϵ being a hyperparameter.

Proof. Since we are guaranteed that $\|\boldsymbol{x}^\dagger - \boldsymbol{x}^\star\|_2 \leq \delta$, by applying Cauchy-Schwarz, we have: $\left|\langle \mathbf{x}^\star - \mathbf{x}^\dagger; \mathbf{r}\rangle\right| \leq \|\mathbf{x}^\star - \mathbf{x}^\dagger\|_1 \|\mathbf{r}\|_\infty \leq \sqrt{|\mathcal{S}||\mathcal{A}|}\|\mathbf{x}^\star - \mathbf{x}^\dagger\|_2 R_{\max} \leq \delta\sqrt{|\mathcal{S}||\mathcal{A}|}R_{\max}$. Equation (22) follows from the average-reward optimality of $\mathbf{x}^\star$ and observing that $\sigma_2(\mathbf{D}_{\mathbf{x}^\star}^{1/2}\mathbf{P}\boldsymbol{\Pi}^\dagger\mathbf{D}_{\mathbf{x}^\star}^{-1/2}) \leq \sigma_2(\mathbf{D}_{\mathbf{x}^\star}^{1/2}\mathbf{P}\boldsymbol{\Pi}^\star\mathbf{D}_{\mathbf{x}^\star}^{-1/2}) + \langle \mathbf{x}^\dagger - \mathbf{x}^\star; \mathbf{r}\rangle \leq \sigma_2(\mathbf{D}_{\mathbf{x}^\star}^{1/2}\mathbf{P}\boldsymbol{\Pi}^\star\mathbf{D}_{\mathbf{x}^\star}^{-1/2})$. □

Despite being a simple workaround to resolve the difficulties associated with the bilinear terms and the tight dependence between the policy and the stationary distribution, Problem (20) has the major drawback of not guaranteeing $\sigma_2(\mathbf{D}_{\mathbf{x}^\dagger}^{1/2}\mathbf{T}^\dagger\mathbf{D}_{\mathbf{x}^\dagger}^{-1/2}) \leq \sigma_2(\mathbf{D}_{\mathbf{x}^\star}^{1/2}\mathbf{T}^\star\mathbf{D}_{\mathbf{x}^\star}^{-1/2})$, namely an improvement on the upper-bound of the SLEM of the resulting MC.

6 Conclusions and Future Works

In this work, we have taken an initial step toward exploring a notion of stability within RL, drawing inspiration from the control-theoretic interpretation of stability. Specifically, we interpret a policy as "more" stable if it converges more rapidly to its steady-state behavior. Within the framework of finite MDPs, we introduced an optimization problem aimed at identifying a Markovian policy capable of explicitly and controllably balancing gain optimality and convergence rate toward steady-state behavior. Crucially, the optimization problem we propose can be efficiently solved using standard convex optimization algorithms. Nonetheless, this study is preliminary and opens several avenues for further research. Firstly, a formulation of the objective provably guaranteeing an improvement of the upper bound of the SLEM is needed. Finally, for our approach to be truly impactful, it must first extend to the classical RL setting, where the model and reward structure are presumed unknown. More crucially, it will be necessary to address continuous state and action spaces, where stability concerns associated with MCs (Meyn and Tweedie, 2012) become considerably more complex, raising fundamental questions not just about convergence rates but also about the existence of equilibrium distributions.

References

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994.

Richard S. Sutton. The reward hypothesis. `http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html`, 2004.

Lucian Buşoniu, Tim de Bruin, Domagoj Tolić, Jens Kober, and Ivana Palunko. Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control*, 46:8–28, 2018.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 89–96. Curran Associates, Inc., 2008.

Alexander Strehl, Lihong Li, and Michael Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

Jean-Jacques E. Slotine and Weiping Li. *Applied Nonlinear Control*. Prentice-Hall, 1991.

Frank L. Lewis, Draguna Vrabie, and Kyriakos G. Vamvoudakis. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems Magazine*, 32(6):76–105, 2012.

Frank L. Lewis and Derong Liu. *Reinforcement learning and approximate dynamic programming for feedback control*. John Wiley & Sons, 2013.

Romain Postoyan, Lucian Buşoniu, Dragan Nešić, and Jamal Daafouz. Stability analysis of discrete-time infinite-horizon optimal control with discounted cost. *IEEE Transactions on Automatic Control*, 62(6), 2017.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022.

Teodor M. Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Eitan Altman, Said Boularouk, and Didier Josselin. Constrained markov decision processes with total expected cost criteria. In *Proceedings of the International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*, pages 191–192. ACM, 2019.

Alessandro Montenegro, Marco Mussi, Matteo Papini, and Alberto M. Metelli. Last-iterate global convergence of policy gradients for constrained reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 126363–126416. Curran Associates, Inc., 2024.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems (NIPS)*, pages 908–918, 2017.

Spencer M. Richards, Felix Berkenkamp, and Andreas Krause. The lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems. In *Conference on robot learning*, pages 466–476. PMLR, 2018.

Hassan K. Khalil. *Nonlinear Control*. Pearson, 2014.

Matthew Ellis, Helen Durand, and Panagiotis D. Christofides. A tutorial review of economic model predictive control methods. *Journal of Process Control*, 24(8):1156–1178, 2014.

James B. Rawlings, David Angeli, and Cuyler N. Bates. Fundamentals of economic model predictive control. In *IEEE Conference on Decision and Control (CDC)*, pages 3851–3861, 2012.

Sebastien Gros and Mario Zanon. Economic mpc of markov decision processes: Dissipativity in undiscounted infinite-horizon optimal control. *Automatica*, 146:110602, 2022.

Sean P. Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Soc., 2017.

Eddie Seabrook and Laurenz Wiskott. A tutorial on the spectral theory of markov chains. *Neural Computation*, 35(11):1713–1796, 2023.

Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra, Second Edition*. Society for Industrial and Applied Mathematics, 2023.

Sean Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, 2022.

Stephen P. Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM Rev.*, 46(4):667–689, 2004.

Jean Tarbouriech and Alessandro Lazaric. Active exploration in markov decision processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 974–982. PMLR, 2019.

Mirco Mutti and Marcello Restelli. An intrinsically-motivated approach for learning highly exploring and fast mixing policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5232–5239, 2020.

Peter D. Lax. *Linear algebra and its applications*. Wiley, 2014.

Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

Ilse C. F. Ipsen and Teresa M. Selee. Ergodicity coefficients defined by vector norms. *SIAM Journal on Matrix Analysis and Applications*, 32(1):153–200, 2011.

Jiu Ding and Aihui Zhou. Eigenvalues of rank-one updated matrices with some applications. *Applied Mathematics Letters*, 20(12):1223–1226, 2007.

Sheldon Axler. *Measure, Integration & Real Analysis*. Springer International Publishing, 2019.

Roger Nussbaum. Notes on the second eigenvalue of the google matrix, 2003.

Martin Kruzik. Bauer's maximum principle and hulls of sets. *Calculus of Variations*, 11:321–332, 2000.

A Heuristics: ϵ-greedy

In the context of finite MDPs, a heuristic approach to achieve a faster convergence rate to a stationary distribution might be to deploy an ϵ-greedy policy, which in general selects actions according to:

$$\pi_\epsilon(a|s) \coloneqq \begin{cases} 1-\epsilon+\frac{\epsilon}{|\mathcal{A}|} & \text{if } a = \pi^\star(s) \\ \frac{\epsilon}{|\mathcal{A}|} & \text{otherwise} \end{cases} \tag{23}$$

Where $\boldsymbol{\pi}^\star : \mathcal{S} \to \mathcal{A}$ denotes a deterministic optimal policy for the average reward objective. In what follows, we characterize first the performance degradation in terms of average expected reward that an agent suffers when sticking to the ϵ-greedy policy, then we characterize the spectral properties induced by such a naive approach, identifying the conditions under which it provides an improvement to the convergence rate.

Theorem A.1 (ϵ-greedy performance loss)**.** *The difference between the average reward of the ϵ-greedy approach $\boldsymbol{\pi}_\epsilon \in \Pi^{MR}$ and the average reward achieved by an average reward optimal policy $\boldsymbol{\pi}^\star$ can be quantified as:*

$$|\langle \mathbf{x}^\star - \mathbf{x}_\epsilon; \mathbf{r}\rangle| \leq \frac{2R_{\max}\epsilon}{1-\tau_1(\mathbf{P}\boldsymbol{\Pi}^\star)}\frac{|\mathcal{A}|-1}{|\mathcal{A}|} \quad \epsilon \in (0,1], \tag{24}$$

where $\mathbf{x}_\epsilon$ denotes the state-action stationary distribution induced by the ϵ-greedy policy.

Proof. Introduce the bivariate MC transition matrices $\mathbf{T}^\star, \mathbf{T}^\epsilon$ induced by the optimal and ϵ-greedy policies, respectively. Because of ergodicity (Assumption 2.1), these chains are ergodic too (Meyn, 2022, Chapter 9).

From Holder's inequality (Axler, 2019, Definition 7.9) we obtain:

$$|\langle \mathbf{x}^\star - \mathbf{x}_\epsilon; \mathbf{r}\rangle| \leq \|\mathbf{x}^\star - \mathbf{x}_\epsilon\|_1 \|\mathbf{r}\|_\infty \leq \|\mathbf{x}^\star - \mathbf{x}_\epsilon\|_1 R_{\max} \tag{25}$$

The one-norm between the stationary distributions can then be characterized as:

$$\begin{aligned} \|\mathbf{x}^\star - \mathbf{x}_\epsilon\|_1 &\leq \frac{1}{1-\tau_1(\mathbf{P}\boldsymbol{\Pi}^\star)}\|\mathbf{T}^\star - \mathbf{T}^\epsilon\|_\infty && (26)\\ &\leq \frac{1}{1-\tau_1(\mathbf{P}\boldsymbol{\Pi}^\star)}\|\mathbf{P}\|_\infty\|\boldsymbol{\Pi}^\star - \boldsymbol{\Pi}^\epsilon\|_\infty && (27)\\ &\leq \frac{1}{1-\tau_1(\mathbf{P}\boldsymbol{\Pi}^\star)}\|\boldsymbol{\Pi}^\star - \boldsymbol{\Pi}^\epsilon\|_\infty && (28)\\ &\leq \frac{1}{1-\tau_1(\mathbf{P}\boldsymbol{\Pi}^\star)}\frac{2\epsilon(|\mathcal{A}|-1)}{|\mathcal{A}|}, && (29) \end{aligned}$$

where line (26) derives from the application of (Ipsen and Selee, 2011, Theorem 3.14), line (27) from the sub-multiplicative property of norms and line (28) from the definition of the optimal deterministic policy $\boldsymbol{\pi}^\star$ and the ϵ-greedy one $\boldsymbol{\pi}^\epsilon$. Importantly, being $\mathbf{T}^\star$ a Markov Matrix (Ipsen and Selee, 2011, Definition 3.11), we are guaranteed that $\tau_1(\mathbf{T}^\star) < 1$ (Ipsen and Selee, 2011, Corollary 3.9). □

Spectral Properties. To study the spectral properties induced by the ϵ-greedy policy, we first notice that the closed loop transition matrix can be written as:

$$\mathbf{P}^\epsilon = (1-\epsilon)\mathbf{P}^{\boldsymbol{\pi}^\star} + \epsilon\mathbf{P}^u, \tag{30}$$

where $\mathbf{P}^u$ denotes the transition matrix obtained as $\mathbf{P}^u = \frac{1}{|\mathcal{A}|^{|\mathcal{S}|}}\sum_{\boldsymbol{\pi}\in\Pi^{MD}}\mathbf{P}^{\boldsymbol{\pi}}$. Because the set $\mathbb{S}_{|\mathcal{S}|}$ is convex, then $\mathbf{P}^\epsilon \in \mathbb{S}_{|\mathcal{S}|}$. For general irreducible stochastic matrices, there is no way of easily relating the change in subdominant eigenvalues to the change in the matrix entries. In fact, despite the function $\lambda_i(\mathbf{A})$ being continuous with respect to the matrix entries $a_{i,j}$, apart from specific cases like symmetric matrices, these functions are in general not convex. To allow for the study of non-symmetric stochastic matrices, we leverage *ergodicity coefficients* (Ipsen and Selee, 2011). Moreover, we leverage (Nussbaum, 2003, Corollary 1) to write $|\lambda_2(\mathbf{P}^\epsilon)| \leq (1-\epsilon)\tau_1(\mathbf{P}^{\boldsymbol{\pi}^\star}) + \epsilon\tau_1(\mathbf{P}^u)$.

Notably, for a matrix $\mathbf{P}^{\boldsymbol{\pi}} \in \mathbb{S}_{|\mathcal{S}|}$ the function $\tau_1(\mathbf{P}^{\boldsymbol{\pi}})$, see Definition 4.1, is a convex function of the policy $\boldsymbol{\pi} \in \Pi^{MR}$. Being Π^{MR} a compact set, the extreme values are achieved at the extremum

points of the feasible set, meaning that the policy that maximizes τ_1, thus achieving the worst possible SLEM upper bound, belongs to Π^{MD} (Kruzik, 2000, Theorem 1). Denote such a policy as $\tilde{\boldsymbol{\pi}}$ and its ergodicity coefficient as $\tilde{\tau}_1$. We can conclude that in the worst case, in which $\boldsymbol{\pi}^\star = \tilde{\boldsymbol{\pi}}$, a necessary condition for $\boldsymbol{\pi}_\epsilon$ to improve the upper bound on the SLEM is the existence of at least one deterministic policy $\boldsymbol{\pi}^\dagger \in \Pi^{MD}$ achieving $\tau_1(\mathbf{P}^{\boldsymbol{\pi}^\dagger}) < \tilde{\tau}_1$. The same conclusions can be derived for the true eigenvalues under the more restrictive assumption that both $\mathbf{P}^{\boldsymbol{\pi}}$ and $\mathbf{P}^u$ are symmetric matrices.