ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

INGEGNERIA BIOMEDICA, ELETTRICA E DEI SISTEMI

Ciclo 38

Settore Concorsuale: 09/G1 - AUTOMATICA

Settore Scientifico Disciplinare: ING-INF/04 - AUTOMATICA

CONTROL-THEORETIC METHODS FOR REINFORCEMENT LEARNING

Presentata da: Simone Baroncini

Coordinatore Dottorato

Michele Monaci

Supervisore

Giuseppe Notarstefano

Co-supervisore

Alessandro Macchelli

Esame finale anno 2026

Control-Theoretic Methods for Reinforcement Learning

Simone Baroncini

2025

Abstract

Reinforcement Learning (RL) has become a cornerstone of modern artificial intelligence and control theory, enabling complex decision-making tasks in a wide variety of domains such as robotics, autonomous driving and strategic games. While deep neural networks allow RL methods to address large-scale problems, exploiting the mathematical structures underlying these problems remains essential for designing efficient and theoretically grounded algorithms. This thesis bridges two complementary perspectives, mathematical analysis of reinforcement learning problems and control-theoretic principles, to develop a rigorous foundation for the design and the analysis of novel reinforcement learning algorithms. The first part of this work investigates a class of learning algorithms that generalize the classical RL paradigm of concurrent policy evaluation and policy improvement. Extending the classical two-timescale framework, a third concurrent process is introduced to incorporate system measurements. Using tools based on singular perturbations theory and LaSalle's invariance principle, sufficient conditions for the convergence of this class of methods are derived. Two novel actor-critic variants are then proposed to address two scenarios with different available measurements from the system. The second part focuses on reward-balancing methods, a recently proposed family of algorithms that approach RL from a dual viewpoint. Unlike standard methods that iteratively improve the policy while keeping the reward fixed, these methods fix the policy to be greedy and iteratively adjust the reward function to make that policy optimal, without altering the underlying problem. This thesis provides a detailed characterization of the transformation group underlying this process and introduces a control-theoretic reformulation of the approach. The resulting framework enables principled inclusion of constraints and penalty terms to improve robustness under model uncertainty. Overall, by exploiting the inherent generality of systems theory, this thesis takes a step toward a unifying framework for the design and the analysis of RL algorithms that leverage their underlying structural properties.

Contents

Notation

We denote by $\mathrm{Map}(A, B)$ the collection of all the morphisms between two objects A and B in a given category. For instance, if A and B are topological spaces, then $\mathrm{Map}(A, B)$ denotes the set of all continuous maps from A to B. The notation $A \simeq B$, where A and B belong to the same category, indicates that A and B are isomorphic. If $f \in \mathrm{Map}(A, B)$, then we write: $f : A \twoheadrightarrow B$ if we want to highlight that the map is surjective, $f : A \hookrightarrow B$ if we want to highlight that the map is injective, and $f : A \to B$ otherwise. The symbol id_A denotes the identity map on a set A. We will use the symbol $\leq$ to denote both the standard order of $\mathbb{R}$ and the pointwise (partial) order between real-valued functions. Specifically, if $f, g \in \mathrm{Map}(A, \mathbb{R})$, for some set A, we write $f \leq g$ to indicate that $f(p) \leq g(p)$ for all $p \in A$. The same convention holds for $\geq$. Given an n-dimensional vector space V with an ordered basis $\mathfrak{B} = \{e_1, \cdots, e_n\}$ and a vector $v \in V$, we will use v^i to denote the i-th component of v with respect to the basis $\mathfrak{B}$. For vectors, the symbols $\leq$ and $\geq$ are also used to denote the component-wise (partial) order, whenever a basis is fixed. If V is a normed space and $v \in V$, we denote by $\|v\|_V$ the norm of v. For Euclidean spaces we use the symbol $\|\cdot\|_p$ to denote the p-norm. If we do not refer to a specific norm we instead use the general symbol $\|\cdot\|$. If M is a matrix, M^i_j denotes the element in the i-th row and j-th column. The identity matrix of dimension n is denoted by I_n, while $\mathbb{1}_n \in \mathbb{R}^n$ denotes the vector of all ones. The notation $\mathrm{diag}\{M_1, \cdots, M_n\}$ refers to the block-diagonal matrix whose i-th block is M_i. If X is a random variable and Y is a set of conditions on X, $\mathbb{E}[X|Y]$ denotes the expectation of X given Y. Finally, the symbol δ^κ denotes the Kronecker delta.

Introduction

Motivations and Challenges

Since the rise of deep neural networks, Reinforcement Learning (RL) has emerged as a central paradigm in modern artificial intelligence, enabling remarkable advances in complex decision-making tasks, ranging from enabling autonomous agents in robotics [1, 2, 3, 4] and autonomous driving [5, 6], to mastering strategic games like Go and chess [7, 8], and even contributing to the training of large language models [9]. However, to fully leverage the approximation power of neural networks, it is essential to exploit the mathematical structures of the learning problem of interest. Typically, at the core of reinforcement learning lies the framework of Markov Decision Processes (MDPs), which model the interaction between an agent and its environment in terms of states, transition probabilities between them and actions encoding how the agent can influence these transitions. While MDPs are often introduced from a probabilistic standpoint, they also possess topological and algebraic properties that offer deeper insights into their behavior and provide a higher theoretical understanding of RL algorithms. Furthermore, the structural properties of the specific MDP of interest can be studied to identify more effective solution strategies for reinforcement learning problems, and to design methods that explicitly exploit these features to improve performance. For example, studying the group associated with the symmetries of an MDP can help reduce its size by clustering states and actions in equivalence classes, thus lowering the algorithm complexity and improving the overall performance. For this reason, relying on neural network designed to preserve those symmetries could be a recommended architectural choice [10].
On the other hand, approaching reinforcement learning from a control-theoretic viewpoint further enhances our ability to exploit the structural properties of the underlying MDPs, while also providing a clear and rigorous framework for the design of learning algorithms. For instance, from a systems-theoretic standpoint, one of the most prominent algorithms in modern RL, the actor-critic, can be viewed as a feedback-interconnected system and analyzed using a wide range of tools, including singular perturbations techniques. After all, reinforcement learning itself can be regarded as the stochastic counterpart of optimal control, sharing many of its fundamental results [11].
The motivation, and the central challenge, of this thesis lies precisely in combining these two perspectives: characterizing the mathematical structure of different problems of interest, and grounding the design and analysis of reinforcement learning methods within a rigorous control-theoretic framework.

Summary of Main Contributions

This thesis contributes to the field of reinforcement learning from a twofold perspective: (i) by proposing novel approaches to the classical reinforcement learning setup, through the design of actor-critic variants based on extensions of singular-perturbations theory, and (ii) by providing deeper insight into a new class of methods, which approach the problem from a "dual" perspective compared to standard methods.
In particular, the first part of this dissertation focuses on the design and analysis of a general class of reinforcement learning methods that aim to maximize a linear functional of the value function in the discounted return case. This class generalizes the classical two-step approaches, which concurrently perform policy evaluation and policy improvement, by introducing a third concurrent step that processes measurements from the system. Starting from a full characterization of this class, the analysis exploits the topological properties of the underlying system to provide sufficient conditions for its convergence in a LaSalle sense, using timescale separation arguments. The discussion then turns to the design of two variants of the actor-critic paradigm, based on two different types of measurement units: the first assumes access to a model of the system, whereas the second operates in a model-free setting, relying solely on experimental data. While the first variant aligns with the aforementioned class of systems, the second one can be interpreted as a stochastic perturbation of a nominal system within this class.
On the other hand, the second part focuses on the novel reward-balancing methods, first introduced in [12], which tackle the discounted-return reinforcement learning problem from a dual perspective. In fact, while standard methods fix the reward function and progressively improve the policy, these methods fix the policy to be greedy and aim to modify the reward function so as to preserve the original problem. As a result, they generate a sequence of equivalent problems, eventually converging to one whose optimal policies are all greedy. The main contribution of this thesis in this framework is a thorough characterization of the algebraic properties of the involved group of transformations, followed by a novel control-theoretic reformulation of these methods. The proposed framework covers both the case of exact model knowledge and that of approximate models. The reformulation of the problem as an optimal control problem eventually provides a rigorous framework to handle additional constraints and penalty terms, which can improve performance in the presence of model uncertainties.

Organization and Chapter Contributions

The thesis is organized following the structure outlined in the previous section.
Chapter 1 begins with a brief introduction to reinforcement learning and presents the Markov decision processes (MDP) framework, along with the discounted-return reinforcement learning problem, which form the common thread of the entire thesis.
Chapter 2 introduces a class of dynamical systems that include the reinforcement learning framework and are composed of three interconnected subsystems: a policy improvement routine, a learning/policy evaluation routine and a measurement-processing mechanism. The main contribution of the chapter consists of a sufficient condition provided for LaSalle-type convergence of the policy to the set of equilibria of the policy improvement routine. Throughout the chapter, each subsystem is analyzed individually, following the approach tipically used

in timescale separation theory, and progressively establishing the properties they are required to satisfy. Finally, it is shown how these individual properties determine the convergence behavior of the interconnected system.
Chapter 3 presents two variants of the actor-critic paradigm, extended to include a third update representing a measurement mechanism. The first part of the chapter introduces the actor-critic setting common to both the algorithms and provides the technical assumptions used in the subsequent analysis. The remainder of the chapter is divided into two main sections, each dedicated to a different type of measurement unit. The first unit assumes a model of the MDP to be accessible and enables the design of a deterministic algorithm, which falls within the class of systems studied in the previous chapter. The second unit considers a model-free measurement unit, leading to the design of a stochastic, though less computationally demanding, algorithm, whose averaged dynamics also fit within the aforementioned class.
Chapter 4 provides a thorough mathematical analysis and a system-theoretic reformulation of the so-called reward-balancing methods, which are novel techniques that address the discounter-return reinforcement learning problem from a completely different perspective from standard methods. These methods fix the policy to be greedy and focus on carefully transforming the reward function in order for the fixed policy to be asymptotically optimal for the original problem. The chapter begins by analyzing the transformations involved, characterizing them as an abelian group related to the invariance of the advantage functions and the reward space as a principal bundle. It then identifies a subset of these transformations that are particularly useful for improving the reward function. Next, the chapter provides a control-theoretic reformulation of the problem, interpreting existing methods as feedback control laws, and introduces an optimal control problem that could serve as the basis for designing new methods, providing sufficient conditions for feasibility and existence of solutions. Eventually, it investigates the impact of model inaccuracies and presents a numerical example showing how the optimal control formulation can improve performance under such conditions.

Chapter 1

Background on Markov Decision Processes

Reinforcement learning (RL) is a machine-learning paradigm for decision-making and control where an agent learns a strategy to act in a given environment by interacting with it and receiving feedback in the form of numerical rewards. The agent's objective is to find a policy, that is, a strategy for selecting actions based on its current state, that maximizes the expected cumulative reward over time. Unlike supervised learning, where the agent has access to a set of correct input-output pairs, reinforcement learning relies only on trial-and-error and direct feedback from the environment. This makes it particularly suitable for sequential decision problems, where each action influences future states and outcomes. To intuitively explain the core components of RL, such as states, actions and rewards, consider the game of chess. Chess is a strategy game played on a 8×8 board, where two players control 16 pieces each. Every piece has a specific set of allowed movements (e.g. a pawn can only move forward and capture diagonally), and players take turns moving one piece at a time. A move can involve repositioning a piece to an empty square or capturing an opponent's piece and replacing it. In this context:

- ➢ A state can be identified with the arrangement of all the pieces on the board at a given time.
- ➢ An action corresponds to a sequence consisting of selecting a piece and moving it to a different square based on the rules of the game.
- ➢ A reward is the feedback received after taking an action, and it could be positive (e.g., capturing a piece), negative (e.g., exposing a piece to capture), or neutral.

However, while the movements rule are fixed, the set of available actions heavily depends on the current state. For instance, the action *move the queen forward one square* is not possible if the queen has already been captured. To handle this, we can introduce an action pool of all possible actions defined a priori, and we can assign to each state a copy of the subset of legal actions from this pool. In this way, *move the queen up one square* becomes two different actions in two different configurations. This state-dependent view of actions is especially useful when considering rewards. For example, the same abstract action *move the queen forward one square* may yield different rewards depending on the situation: it could result in capturing an opponent's piece (positive reward), losing the queen (negative reward), or having no immediate consequence (zero reward). Therefore, it makes sense to see actions in their state-specific context, as the outcome of an action depends on the configuration in which it is taken. Ultimately, the objective of chess is not simply to capture pieces, but to

checkmate the opponent's king. This introduces the need for a policy to account for long-term consequences. A player who always chooses the move with the highest reward, such as capturing a valuable piece whenever possible, might fall into a trap that leads to a checkmate in a few moves. This illustrates that a greedy policy is not always optimal, and the player must learn to weigh short-term gains against long-term outcomes.
The mathematical framework commonly used to model reinforcement learning problems is that of Markov decision processes (MDPs). MDPs capture the probabilistic nature of the environment through state transitions, whose probabilities depend on the actions taken by the agent. The next section introduces MDPs more formally, providing a rigorous framework for the concepts introduced in the previous example. For additional introductory texts on MDPs and reinforcement learning, consider consulting [11, 13, 14, 15].

Markov Decision Processes

In this thesis we focus on classes of Markov decision processes sharing, up to a bijection, a finite state space $\mathcal{X}$ of cardinality n, and a finite action pool $\widetilde{\mathcal{U}}$ of cardinality $\widetilde{m}$. We define the action space $\mathcal{U}$ as the disjoint union

$$\mathcal{U} \coloneqq \bigsqcup_{x\in\mathcal{X}} \widetilde{\mathcal{U}}_x,$$

where $\widetilde{\mathcal{U}}_x \subseteq \widetilde{\mathcal{U}}$, with cardinality m_x, denotes the subset of the overall action pool $\widetilde{\mathcal{U}}$ of all those actions which can be taken from the state x. It follows that the action space has cardinality $m \coloneqq \sum_x m_x$. This construction admits a natural projection map $s : \mathcal{U} \twoheadrightarrow \mathcal{X}$, which maps each "disjoint" action to the state it belongs to. Henceforth, we will denote by $\mathcal{U}_x$ the fiber $s^{-1}(x)$ of the state x. In this framework, a deterministic stationary policy (DSP or, simply, policy) π can be conveniently defined as a section of the projection map, i.e. a map $\pi : \mathcal{X} \to \mathcal{U}$ such that $s \circ \pi = \mathrm{id}_{\mathcal{X}}$. We denote the set of all the deterministic stationary policies by $\Gamma(\mathcal{X},\mathcal{U})$. For convenience, we define a preliminary classification of MDPs based solely on the "bundle" construction introduced so far.

Definition 1. *We say that two MDPs are comparable iff there exists a bijective bundle map*[1] *between their action spaces.*

Within each class, every MDP is characterized by its reward function $r : \mathcal{U} \to \mathbb{R}$, defining the goodness of each action, and its probability transition function $f : \mathcal{X} \times \mathcal{U} \to [0,1]$, where $f(x,u)$ denotes the probability of transitioning to x after applying u (from the state $s(u)$, of course). A policy π naturally induces pullbacks of both functions onto the state space given by $r_\pi \coloneqq r \circ \pi$ and $f_\pi \coloneqq f \circ (\mathrm{id}_{\mathcal{X}} \times \pi)$, respectively. The objective of discounted-return reinforcement learning is to find a DSP maximizing the so-called value function $v_{\pi,r} : \mathcal{X} \to \mathbb{R}$ (or a functional thereof) which assigns to each state the expected long-run cumulative reward associated with the given policy, and is formally defined as

$$v_{\pi,r}(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_\pi(x_t) \,\middle|\, \begin{array}{l} x_0 = x \\ x_{t+1} \sim f_\pi(\cdot, x_t) \end{array}\right], \tag{1.1}$$

[1] Given two projections $s_1 : \mathcal{U}_1 \twoheadrightarrow \mathcal{X}_1$ and $s_2 : \mathcal{U}_2 \twoheadrightarrow \mathcal{X}_2$, a map $F : \mathcal{U}_1 \to \mathcal{U}_2$ is said to be a *bundle map* iff it is fiber-preserving, i.e. there exists a map $G : \mathcal{X}_1 \to \mathcal{X}_2$ such that $s_2 \circ F = G \circ s_1$.

for some given discount factor $\gamma \in [0,1)$. Specifically, we aim to find $\pi^* \in \Gamma(\mathcal{X},\mathcal{U})$ such that:

$$v_{\pi^*,r}(x) \geq v_{\pi,r}(x), \quad \forall x \in \mathcal{X}, \forall \pi \in \Gamma(\mathcal{X},\mathcal{U}).$$

An optimal DSP is always guaranteed to exist in this framework (see [13, Theorem 7.1.9]),

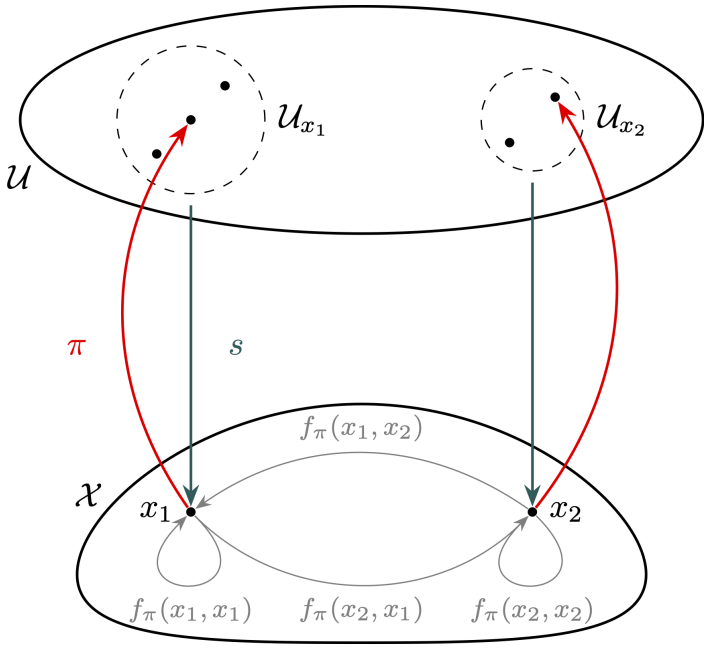


Figure 1.1: Representation of an MDP with two states x_1, x_2 and five actions, three available in x_1 and two available in x_2.

making the problem well-posed. Moreover, the value function $v_{\pi,r}$ can be characterized as the unique solution to the renowned Bellman equations (see, e.g., [13, Theorem 6.1.1]):

$$v(x) = r_\pi(x) + \gamma \sum_{z \in \mathcal{X}} f_\pi(z,x) v(z), \quad \forall x \in \mathcal{X}. \tag{1.2}$$

This condition can be also expressed in terms of action-level comparisons by introducing the advantage function $\mathrm{adv}_{\pi,r} : \mathcal{U} \to \mathbb{R}$ corresponding to the policy π, defined by

$$\mathrm{adv}_{\pi,r}(u) \coloneqq r(u) + \gamma \sum_{z \in \mathcal{X}} f(z,u) v_{\pi,r}(z) - v_{\pi,r}(s(u)),$$

for all $u \in \mathcal{U}$. As the name suggests, this function quantifies the benefit of deviating from the current policy by selecting u in place of the action $\pi(x)$ taken by the policy in state $x = s(u)$. Intuitively, if $\mathrm{adv}_{\pi,r}(u) > 0$ then replacing $\pi(x)$ with u improves the expected return, whereas if $\mathrm{adv}_{\pi,r}(u) < 0$ then keeping the current policy is preferable. With this formulation, the fact that $v_{\pi,r}$ satisfies the equations (1.2) can be compactly written as

$$\mathrm{adv}_{\pi,r} \circ \pi = 0. \tag{1.3}$$

The advantage function also allows us to introduce the action-value function $q_{\pi,r} : \mathcal{U} \to \mathbb{R}$, commonly referred to as Q-function, defined as

$$q_{\pi,r} \coloneqq \mathrm{adv}_{\pi,r} + v_{\pi,r} \circ s, \tag{1.4}$$

which assigns to each action a long-run expected return just as the value function does to states. By (1.3), it follows that the value function is given by the pullback, via the policy π, of the Q-function:

$$v_{\pi,r} = q_{\pi,r} \circ \pi.$$

Stochastic Stationary Policies

Although an optimal DSP always exists, in practice it is often convenient to work with stochastic policies. This is either because they enable exploration of the MDP or because the policy improvement method employed benefits from smooth policy parametrizations. Formally a stochastic stationary policy (SSP) π is a law assigning to each state x the probability mass function $\pi(\cdot, x)$ on its fiber $\mathcal{U}_x$, defining the probability with which each action of the fiber should be taken from that state. We denote by $\Gamma_s(\mathcal{X},\mathcal{U})$ the set of all stationary policies of the given MDP. Note that $\Gamma(\mathcal{X},\mathcal{U})$ can be viewed as a subset of $\Gamma_s(\mathcal{X},\mathcal{U})$ by identifying each deterministic policy $\pi : x \mapsto \pi(x)$ with the stochastic policy $\pi_s : x \mapsto \pi_s(\cdot, x)$, where $\pi_s(u, x) = 1$ if $u = \pi(x)$ and $\pi_s(u, x) = 0$ otherwise. This identification allows us to use the symbol π to represent both throughout the text. The context will make it clear whether we are focusing solely on DSPs or also allowing for SSPs. Similar to the deterministic case, once a stochastic policy π is specified, the MDP becomes an autonomous Markov process on $\mathcal{X}$, with a transition function $f_\pi : \mathcal{X} \times \mathcal{X} \to [0, 1]$ and a reward function $r_\pi : \mathcal{X} \to \mathbb{R}$. These functions are related to those of the MDP through the following relations

$$f_\pi(\cdot, x) = \mathbb{E}\left[f(\cdot, u)\middle|\ u \sim \pi(\cdot, x)\right] \tag{1.5a}$$

$$r_\pi(x) = \mathbb{E}\left[r(u)\middle|\ u \sim \pi(\cdot, x)\right]. \tag{1.5b}$$

In this framework, the value function, still defined by (1.1), remains the unique solution of (1.2), where r_π and f_π are replaced with their stochastic counterparts (1.5). Moreover, its relation with the Q-function is modified as follows:

$$v_{\pi,r}(x) = \mathbb{E}\left[q_{\pi,r}(u)\middle|\ u \sim \pi(\cdot, x)\right], \quad \forall x \in \mathcal{X}.$$

The Q-function is in turn the unique solution of an action-level version of (1.2):

$$q_{\pi,r}(u) = r(u) + \gamma \sum_{z \in \mathcal{X}} f(z, u) \sum_{v \in \mathcal{U}_z} \pi(v, z) q_{\pi,r}(v), \quad \forall u \in \mathcal{U}. \tag{1.6}$$

Finite-Dimensional Vector Representation

Often, it is convenient to identify the functions introduced so far with proper vectors and matrices. In order to do so, we index both the state space $\mathcal{X}$, by identifying it with the set of the first n positive integers, and, accordingly, the action space $\mathcal{U}$, with the set of the first m positive integers, by first enumerating the actions in the fiber of the first state, then those

in the fiber of the second state, and so on. Let us denote by $\mathcal{I}_{\mathcal{X}} : \mathcal{X} \to \{1, \cdots, n\}$ the first index map and by $\mathcal{I}_{\mathcal{U}} : \mathcal{U} \to \{1, \cdots, m\}$ the second one. With this indexing in mind, in order not to burden the notation any further we will use in the following m_i in place of m_x and $\mathcal{U}_i$ in place of $\mathcal{I}_{\mathcal{U}}(\mathcal{U}_x)$ whenever $i = \mathcal{I}_{\mathcal{X}}(x)$. We can then identify the reward function r with a vector $R \in \mathbb{R}^m$ whose components satisfy

$$R^i = \left(r \circ \mathcal{I}_{\mathcal{U}}^{-1}\right)(i)$$

for all $i \in \{1, \cdots, m\}$, and the probability transition function f with a row-stochastic matrix $F \in [0,1]^{m \times n}$, whose entries satisfy

$$F^i_j = \left[f \circ (\mathcal{I}_{\mathcal{X}} \times \mathcal{I}_{\mathcal{U}})^{-1}\right](j, i)$$

for all $i \in \{1, \cdots, m\}$ and $j \in \{1, \cdots, n\}$. Eventually, a stochastic stationary policy π can be represented by a row-stochastic, block-diagonal matrix $\Pi \in [0,1]^{n \times m}$ whose entries satisfy

$$\Pi^i_j = \begin{cases} \pi\big(\mathcal{I}_{\mathcal{U}}^{-1}(j),\ \mathcal{I}_{\mathcal{X}}^{-1}(i)\big), & \text{if } j \in \mathcal{U}_i \\ 0, & \text{otherwise,} \end{cases}$$

and we use this identification to equip $\Gamma_s(\mathcal{X}, \mathcal{U})$ with the subspace topology inherited by $\mathbb{R}^{n \times m}$. In particular, when the policy is deterministic, the matrix Π satisfies $\Pi^i_j = 1$ if $j = (\mathcal{I}_{\mathcal{U}} \circ \pi \circ \mathcal{I}_{\mathcal{X}}^{-1})(i)$ and $\Pi^i_j = 0$ otherwise. In this way, the "pullbacks" representations $R_\pi \in \mathbb{R}^n$ and $F_\pi \in \mathbb{R}^{n \times n}$ can be easily computed as

$$R_\pi = \Pi R \tag{1.7}$$

$$F_\pi = \Pi F. \tag{1.8}$$

Eventually, as done for the other functions, we can represent both value functions with two vectors $V_{\pi,r} \in \mathbb{R}^n$ and $Q_{\pi,r} \in \mathbb{R}^m$. In this representation, we can exploit the Bellman equations (1.2)-(1.6) to express the value vectors as functions of the reward vectors as follows

$$V_{\pi,r} = (I_n - \gamma F_\pi)^{-1} R_\pi, \tag{1.9}$$

$$Q_{\pi,r} = (I_m - \gamma F \Pi)^{-1} R. \tag{1.10}$$

for all $\pi \in \Gamma_s(\mathcal{X}, \mathcal{U})$. These two representations are naturally related by $V_{\pi,r} = \Pi Q_{\pi,r}$, as proved by the following identities:

$$\begin{aligned} \Pi Q_{\pi,r} &= \Pi (I_m - \gamma F \Pi)^{-1} R \\ &= \sum_{t=0}^{\infty} \gamma^t \Pi (F \Pi)^t R \\ &= \sum_{t=0}^{\infty} \gamma^t (\Pi F)^t \Pi R \\ &= (I_n - \gamma F_\pi)^{-1} R_\pi \\ &= V_{\pi,r} \end{aligned}$$

MDPs Relations

We eventually introduce two meaningful relations that enable us to further partition a class of comparable MDPs, which will be useful in Chapter 4.

Definition 2. *We say that two comparable MDPs are:*

- *reward-coupled iff their reward functions coincide (up to a bijection of action spaces).*
- *model-coupled iff they have the same discount factor and their transition functions coincide (up to a bijection of state and action spaces).*

Since in this thesis we will mostly consider comparable MDPs we will denote an MDP with reward r, transition function f and discount γ as $\mathfrak{M}(r, \gamma f)$. Note that the knowledge of γf is enough to retrieve both γ and f.

Group-Invariant Markov Decision Processes

In many cases, the MDP of interest exhibits some form of symmetry, meaning that certain states and actions are interchangeable. For example, in a game like tic-tac-toe, an empty board with a cross in the top-right corner is equivalent, from the point of view of the game dynamics, to an empty board with the cross in any other corner, provided that the available actions are rotated accordingly. More generally, a symmetry in an MDP implies that there exists a group of transformations acting simultaneously on the state and action spaces, under which the transition and reward functions remain invariant. We formalize this notion in the following definition (see, e.g., [10]).

Definition 3. *Let $\mathfrak{M}(r, \gamma f)$ be an MDP and let G be a group acting on both $\mathcal{X}$ and $\mathcal{U}$ through actions ρ^{x} and ρ^{u}, respectively. We say that the tuple $\mathfrak{M}_G(r, \gamma f) \coloneqq (\mathfrak{M}(r, \gamma f), G)$ is a G-invariant MDP iff the following properties hold:*

$$\textit{(Reward Invariance):}\ r\left(\rho_g^{\mathrm{u}}(u)\right) = r(u)$$
$$\textit{(Transition Invariance):}\ f\left(\rho_g^{\mathrm{x}}(x),\ \rho_g^{\mathrm{u}}(u)\right) = f(x, u)$$

for all $x \in \mathcal{X}$, $u \in \mathcal{U}$ and $g \in G$.

The group actions ρ^{x} and ρ^{u} can be viewed as homomorphisms from the group G to the symmetric groups $\mathrm{Sym}(\mathcal{X}) \simeq S_n$ and $\mathrm{Sym}(\mathcal{U}) \simeq S_m$, respectively. Therefore, in the finite-dimensional vector representation introduced at the beginning of the chapter, ρ_g^{x} and ρ_g^{u} are represented by two permutation matrices $A_g^{\mathrm{x}} \in \{0,1\}^{n \times n}$ and $A_g^{\mathrm{u}} \in \{0,1\}^{m \times m}$. Hence, in this representation, the group invariant conditions of Definition 3 read as:

$$A_g^{\mathrm{u}} R = R \tag{1.11a}$$
$$A_g^{\mathrm{u}} F \left(A_g^{\mathrm{x}}\right)^{\top} = F. \tag{1.11b}$$

To illustrate the benefit of exploiting MDP symmetries, consider a simple example: a "cyclic corridor" with three states $\{x_1, x_2, x_3\}$ and two actions a-priori: u_l to move left (mod 3) and

u_r to move right (mod 3). The action pool is then $\tilde{\mathcal{U}} \coloneqq \{u_l, u_r\}$, and the action space is made of three copies of these two actions. We will denote by u_l^i and u_r^i the left and right actions taken from state x_i. Hence, the action space $\mathcal{U}$ has cardinality 6. The transition function can be represented as follows:

f	x_1	x_2	x_3
u_l^1	0	0	1
u_r^1	0	1	0
u_l^2	1	0	0
u_r^2	0	0	1
u_l^3	0	1	0
u_r^3	1	0	0

Suppose that our goal is to make the agent frequently visit the state x_2 (i.e. it should keep returning to x_2). To this end, we may define the following reward map:

r	u_l	u_r
x_1	-1	0
x_2	-1	-1
x_3	0	-1

Now, consider the group $S_2 \coloneqq \{e, \sigma\}$ of permutations of 2 elements acting on this MDP. The group acts on $\tilde{\mathcal{U}}$ by either doing nothing ($\rho_e^{\tilde{\mathrm{u}}}$) or by swapping the two actions ($\rho_\sigma^{\tilde{\mathrm{u}}}$), and it acts on the state space by either doing nothing (ρ_e^{x}) or swapping x_1 and x_3 (ρ_σ^{x}). Specifically, the action of G on $\mathcal{X}$ defines an injective group homomorphism from G to the symmetric group of degree 3, S_3, whose range is the subgroup generated by the 2-cycle $\sigma_{13} \coloneqq (13)$. The resulting non-trivial action on $\mathcal{U}$ consists of applying $\rho_\sigma^{\tilde{\mathrm{u}}}$ to each fiber, followed by the permutation of the fibers induced by ρ_σ^{x}. With respect to the actions of the group generator σ, the MDP exhibits transition invariance

f	x_1	x_2	x_3
u_l^1	0	0	1
u_r^1	0	1	0
u_l^2	1	0	0
u_r^2	0	0	1
u_l^3	0	1	0
u_r^3	1	0	0

$\xmapsto{\rho_\sigma^{\tilde{\mathrm{u}}}}$

f	x_1	x_2	x_3
u_r^1	0	1	0
u_l^1	0	0	1
u_r^2	0	0	1
u_l^2	1	0	0
u_r^3	1	0	0
u_l^3	0	1	0

$\xmapsto{\rho_\sigma^{\mathrm{x}}}$

f	x_3	x_2	x_1
u_r^3	0	0	1
u_l^3	0	1	0
u_r^2	1	0	0
u_l^2	0	0	1
u_r^1	0	1	0
u_l^1	1	0	0

and reward invariance

r	u_l	u_r
x_1	-1	0
x_2	-1	-1
x_3	0	-1

$\xmapsto{\rho_\sigma^{\tilde{\mathrm{u}}}}$

r	u_r	u_l
x_1	0	-1
x_2	-1	-1
x_3	-1	0

$\xmapsto{\rho_\sigma^{\mathrm{x}}}$

r	u_r	u_l
x_3	-1	0
x_2	-1	-1
x_1	0	-1

This implies that the considered MDP is S_2-invariant. This allows us to identify the original MDP with a "reduced" MDP having:

➢ a reduced state space:

$$\mathcal{X}' \coloneqq \{[x_1], x_2\}, \quad [x_1] \coloneqq \{x_1, x_3\},$$

➢ a reduced action space:

$$\mathcal{U}' \coloneqq \{[u_l^1], [u_r^1], [u_l^2]\},$$

where

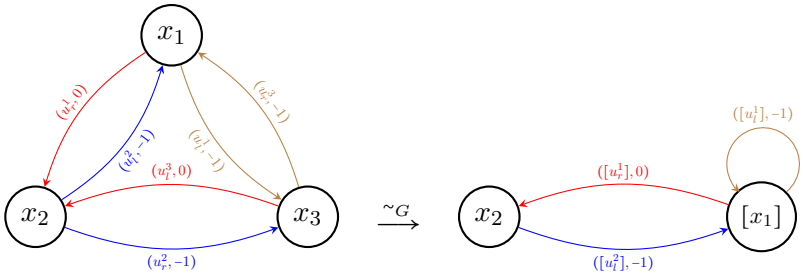
$$[u_l^1] = \{u_l^1, u_r^3\}, \quad [u_r^1] = \{u_r^1, u_l^3\}, \quad [u_l^2] = \{u_l^2, u_r^2\},$$

➢ a reduced transition map:

f'	$[x_1]$	x_2
$[u_l^1]$	1	0
$[u_r^1]$	0	1
$[u_l^2]$	1	0

➢ a reduced reward map:

r	u_l	u_r
$[x_1]$	-1	0
x_2	-1	



This example shows how exploiting the symmetries of the underlying MDP can help reduce its size, and thus the complexity of the reinforcement learning methods employed. The concepts of G-invariant MDPs will be recalled in Chapter 4, to characterize when a reward transformation preserves these symmetries.

Chapter 2

Timescale Separation for Optimization, Learning and Control Systems

In this chapter, we analyze a family of dynamical systems composed of three interconnected subsystems, designed to abstract the core components of many reinforcement learning algorithms: a policy improvement routine, a learning (or policy evaluation) mechanism, and a measurement unit. Each subsystem is characterized individually, following a timescale separation approach, with the goal of establishing properties that each subsystem should satisfy to guarantee desirable behavior of the overall system. In particular, the main result of the chapter provides a sufficient condition for LaSalle-type convergence of the policy to the set of equilibria associated with the policy improvement routine.

Literature Review

In this chapter we make extensive use of tools from control theory, with particular emphasis on timescale separation methods and Lyapunov theory. In the continuous-time setting, the classical reference for Lyapunov theory is [16], while the discrete-time counterpart is well covered in [17]. Timescale separation methods were first introduced in control theory, where they are commonly referred to as *singular perturbations* methods, by Kokotović and collaborators [18, 19], and has since been developed in a wide variety of works over the years. Among these, [20] is a comprehensive introduction to the topic. The discrete-time counterpart, while less extensively treated, has also been explored in significant works such as [21, 22]. The idea of timescale separation approaches in reinforcement learning gained popularity since the introduction of actor-critic methods, and was first rigorously investigated by Borkar and collaborators in [23], based on they earlier results on stochastic approximation with two timescales [24]. More recent developments on actor-critic methods based on timescale separation can be found in [25, 26, 27, 28].

A General Class of Learning Problems

Policy optimization mechanism

We are interested in optimizing a control policy parametrized by a vector $\theta \in \mathbb{R}^{n_\theta}$, through a topological embedding $\wp : \mathbb{R}^{n_\theta} \hookrightarrow \hat{\mathbb{P}}$, where $\hat{\mathbb{P}}$ is a compact metric space (e.g. a compact

N-manifold). We then formulate an optimization problem of the form

$$\sup_{\theta \in \mathbb{R}^{n_\theta}} J_{\wp}(\theta), \tag{2.1}$$

where $J_{\wp} \coloneqq J \circ \wp$ is the pullback through $\wp$ of a continuous objective function $J : \hat{\mathbb{P}} \to \mathbb{R}$, typically encoding performance and/or robustness criteria of the controlled system. Our objective is to iteratively find a solution to problem (2.1). To this end, given the iteration index $k \in \mathbb{N}$, we maintain an estimate $\theta_k \in \mathbb{R}^{n_\theta}$ of a solution to problem (2.1), and update it using a chosen optimization algorithm, namely

$$\theta_{k+1} = f^{\alpha}(\theta_k), \tag{2.2}$$

where $f^{\alpha} : \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_\theta}$ encodes the specific update rule (e.g., the gradient method if $J_{\wp}$ is continuously differentiable), and $\alpha \in [0, 1]$ denotes an algorithm parameter (such as a step size). The embedding $\wp$ induces a corresponding dynamics $\bar{\varphi}^{\alpha} : \wp(\mathbb{R}^{n_\theta}) \to \wp(\mathbb{R}^{n_\theta})$ on the image of the parametrization map, defined as

$$\bar{\varphi}^{\alpha} = \wp \circ f^{\alpha} \circ \wp^{-1}$$

which will be central to the analysis in the following sections. Since $\hat{\mathbb{P}}$ is compact (and thus complete), under suitable uniform continuity assumptions, this dynamics can be continuously extended to the closure $\mathbb{P} \coloneqq \kappa(\wp(\mathbb{R}^{n_\theta}))$ (see [29, Th.26, p.195]). Thus, in what follows, we will restrict our attention to the compact metric subspace $\mathbb{P}$, and we will consider $\pi \in \mathbb{P}$ as the optimization variable, iteratively updated by:

$$\pi_{k+1} = \bar{\varphi}^{\alpha}(\pi_k). \tag{2.3}$$

We now focus on the case where the objective function $J|_{\mathbb{P}}$ can be expressed as the restriction of a real-valued function $(\pi, \omega) \mapsto J'(\pi, \omega)$ to the graph of a map $\bar{\omega} : \mathbb{P} \to \Omega \simeq \mathbb{R}^{n_\omega}$, namely

$$J(\pi) \coloneqq J'(\pi, \bar{\omega}(\pi)).$$

Thus, we rewrite problem (2.1) as

$$\max_{\pi \in \mathbb{P}} J'(\pi, \bar{\omega}(\pi)). \tag{2.4}$$

The vector $\bar{\omega}(\pi)$ captures the value of certain components of the objective function that are typically not directly accessible, such as the value function. Since the optimization method (2.2) is usually built upon the given objective (as in the case of an ascent direction method), we assume it depends on $\bar{\omega}(\pi)$ as well. In practical scenarios, this ideal quantity is replaced by an estimate, which is continuously updated using data collected from a measurement unit. Furthermore, the policy update function itself is often not fully known and relies on data obtained from the underlying system. To model these effects, we introduce the following quantities:

1. a vector $\omega \in \Omega \simeq \mathbb{R}^{n_\omega}$, representing the current data-based estimate of $\bar{\omega}(\pi)$, updated by a learning mechanism;

2. a vector $\zeta \in Z \simeq \mathbb{R}^{n_\zeta}$, capturing the contribution of the underlying measurement mechanism, which provides data for both the learning and the optimization mechanisms.

Accordingly, we define $v \coloneqq (\omega, \zeta)$. Therefore, in the ideal scenario of perfect learning and measurement, we rewrite the update rule in (2.3) as

$$\bar{\varphi}^{\alpha}(\pi) = \varphi^{\alpha}\left(\pi, \left(\bar{\omega}(\pi), \hat{\zeta}(\pi, \bar{\omega}(\pi))\right)\right), \tag{2.5}$$

for some map $\varphi^{\alpha} : \mathbb{P} \times (\Omega \times Z) \to \mathbb{P}$, where $\hat{\zeta} : \mathbb{P} \times \Omega \to Z \simeq \mathbb{R}^{n_\zeta}$ is the map representing the ideal measurement mechanism, which we generally assume to be affected by both the policy and the estimate. In the next two subsections we briefly discuss the cases in which these two mechanisms are no longer ideal. In such setting, the optimization mechanism at iteration k must rely on two estimates: ω_k for $\bar{\omega}(\pi_k)$, and ζ_k for $\hat{\zeta}(\pi_k, \bar{\omega}(\pi_k))$. This leads to the following online update rule:

$$\pi_{k+1} = \varphi^{\alpha}(\pi_k, (\omega_k, \zeta_k)). \tag{2.6}$$

Learning mechanism

As anticipated in the previous discussion, the second component of the methods in the considered class is then a learning mechanism, designed to steer the estimate ω_k toward the exact value $\bar{\omega}(\pi_k)$, while π_k evolves according to (2.6). We generally define the learning mechanism under the assumption of ideal offline measurements $\hat{\zeta}(\pi_k, \omega_k)$ as

$$\omega_{k+1} = g_k^{\beta}\left(\pi_k, \left(\omega_k, \hat{\zeta}(\pi_k, \omega_k)\right)\right), \tag{2.7}$$

where $g_k^{\beta} : \mathbb{P} \times (\Omega \times Z) \to \Omega$ represents a given learning update, and $\beta \in [0, 1]$ is a tuning parameter. Here, as in the policy block, we are interested in online strategies where ω_k is iteratively updated while gathering information ζ_k from the actuated system. Thus, the offline update (2.7) is replaced by

$$\omega_{k+1} = g_k^{\beta}(\pi_k, (\omega_k, \zeta_k)). \tag{2.8}$$

Measurements unit

Finally, we describe the update of ζ_k in the general form

$$\zeta_{k+1} = h_k(\pi_k, (\omega_k, \zeta_k)), \tag{2.9}$$

where $h_k : \mathbb{P} \times (\Omega \times Z) \to Z$ describes the information-gathering phase and typically depends on the dynamics of the plant under the current policy π_k.

The Overall System Structure

By collecting the optimization part (2.6), the learning part (2.8), and the measurements part (2.9), the overall closed-loop system is described by the following three interconnected dynamics

$$\pi_{k+1} = \varphi^{\alpha}(\pi_k, (\omega_k, \zeta_k)) \tag{2.10a}$$
$$\omega_{k+1} = g_k^{\beta}(\pi_k, (\omega_k, \zeta_k)) \tag{2.10b}$$
$$\zeta_{k+1} = h_k(\pi_k, (\omega_k, \zeta_k)). \tag{2.10c}$$

In particular, the state space of interest is given by a product bundle $\mathcal{B} \coloneqq \mathbb{P} \times V$ consisting of a compact metric space $\mathbb{P}$, with metric $d_{\mathbb{P}}$ and a direct sum $V \coloneqq \Omega \oplus Z$ of two normed spaces, Ω, with norm $\|\cdot\|_{\Omega}$, and Z, with norm $\|\cdot\|_{Z}$, equipped with the norm

$$\|(\omega, \zeta)\|_{V} \coloneqq \|\omega\|_{\Omega} + \|\zeta\|_{Z}, \quad \omega \in \Omega, \zeta \in Z.$$

The resulting bundle $\mathcal{B}$ is then equipped with the metric

$$d_{\mathcal{B}}((\pi_1, v_1), (\pi_2, v_2)) \coloneqq d_{\mathbb{P}}(\pi_1, \pi_2) + \|v_1 - v_2\|_{V},$$

where $\pi_1, \pi_2 \in \mathbb{P}, v_1, v_2 \in V$. In the next sections, we will detail the assumptions by separately characterizing each block of (2.10). Before doing so, we first provide the next assumption, which characterizes the equilibria of the learning part (2.10b) and the measurements part (2.10c).

Assumption 1. *The functions $\bar{\omega} : \mathbb{P} \to \Omega$ and $\hat{\zeta} : \mathbb{P} \times \Omega \to Z$ satisfy*

$$\begin{aligned} \bar{\omega}(\pi) &= g_k^{\beta}\left(\pi, \left(\bar{\omega}(\pi), \hat{\zeta}(\pi, \bar{\omega}(\pi))\right)\right) \\ \hat{\zeta}(\pi, \omega) &= h_k\left(\pi, \left(\omega, \hat{\zeta}(\pi, \omega)\right)\right), \end{aligned}$$

for all $\pi \in \mathbb{P}, \omega \in \Omega$, $k \in \mathbb{N}$, and $\beta \in [0, 1]$. Moreover, both functions are uniformly Lipschitz continuous in each argument, namely there exist $L_{\bar{\omega}}, L_{\hat{\zeta},1}, L_{\hat{\zeta},2} > 0$ such that

$$\begin{aligned} \|\bar{\omega}(\pi_1) - \bar{\omega}(\pi_2)\|_{\Omega} &\leq L_{\bar{\omega}} d_{\mathbb{P}}(\pi_1, \pi_2) \\ \left\|\hat{\zeta}(\pi_1, \omega_1) - \hat{\zeta}(\pi_2, \omega_2)\right\|_{Z} &\leq L_{\hat{\zeta},1} d_{\mathbb{P}}(\pi_1, \pi_2) + L_{\hat{\zeta},2} \|\omega_1 - \omega_2\|_{\Omega}, \end{aligned}$$

for all $\pi_1, \pi_2 \in \mathbb{P}$ and $\omega_1, \omega_2 \in \Omega$.

Assumption 1 ensures that the learning dynamics (2.10b) has equilibria parametrized in the policy state π through the function $\bar{\omega}$ and that the measurements' dynamics (2.10c) has equilibria parametrized in the policy and learning states π and ω through the function $\hat{\zeta}$. Introducing the shorthand $\bar{\zeta}(\pi) \coloneqq \hat{\zeta}(\pi, \bar{\omega}(\pi))$, Assumption 1 implies that

$$\bar{\omega}(\pi) = g_k^{\beta}\left(\pi, \left(\bar{\omega}(\pi), \bar{\zeta}(\pi)\right)\right) \tag{2.11a}$$

$$\bar{\zeta}(\pi) = h_k\left(\pi, \left(\bar{\omega}(\pi), \bar{\zeta}(\pi)\right)\right), \tag{2.11b}$$

for all $\pi \in \mathbb{P}, k \in \mathbb{N}$, and $\beta \in [0, 1]$.

Properties of Individual Subsystems

In this section, following the standard practice of singular perturbations theory, we provide a technical characterization of each of the three building blocks introduced in the previous section. In particular, we characterize the behavior of each individual subsystem, in the ideal case where the other subsystems behave perfectly.

The Policy Optimization Dynamics

We characterize the policy optimization mechanism as a continuous one-parameter family of maps $\{\varphi^\alpha : \mathcal{B} \to \mathbb{P}\}_{\alpha\in[0,1]}$ such that

$$\varphi^0(\pi, v) = \pi, \quad \forall \pi \in \mathbb{P}, v \in V. \tag{2.12}$$

That is, φ^0 is the canonical projection map $\mathcal{P}_{\mathcal{B}\to\mathbb{P}}$ and thus, for $\alpha = 0$, π_k does not vary in time. Furthermore, we also make stronger continuity assumptions on this family of maps.

Assumption 2. *The slow dynamics φ^α satisfy*

$$d_{\mathbb{P}}(\varphi^\alpha(\pi_1, v), \varphi^\alpha(\pi_2, v)) \le L^\alpha_{\varphi,1}\, d_{\mathbb{P}}(\pi_1, \pi_2)$$
$$d_{\mathbb{P}}(\varphi^\alpha(\pi, v_1), \varphi^\alpha(\pi, v_2)) \le L^\alpha_{\varphi,2} \|v_1 - v_2\|_V$$

for all $\pi, \pi_1, \pi_2 \in \mathbb{P}$ and for all $v, v_1, v_2 \in V$, where $L^\alpha_{\varphi,1} > 0$ and $L^\alpha_{\varphi,2} \ge 0$ are continuous functions of α, satisfying

$$L^0_{\varphi,1} = 1 \tag{2.13a}$$
$$L^\alpha_{\varphi,2} > 0, \ \forall \alpha \in (0,1], \quad L^\alpha_{\varphi,2} \sim_A l\alpha \ \ as \ \ \alpha \to 0, \tag{2.13b}$$

for some $l \ge 0$, where $\sim_A$ denotes asymptotic equivalence[1]

Assumption 2 ensures that the parameter α can be used to arbitrarily reduce the speed of variation of π_k. In light of this property and the fact that the equilibria of (2.10b)-(2.10c) are parametrized in π_k (cf. Assumption 1), we will refer to the optimization part (2.10a) as *slow* system. To complete the characterization, we now examine the system under the assumption that $\omega_k = \bar{\omega}(\pi_k)$ and $\zeta_k = \bar{\zeta}(\pi_k)$ for all $k \in \mathbb{N}$. That is, we impose that both ω_k and ζ_k lie on the corresponding equilibria manifolds parametrized by the slow state π_k (see Assumption 1). Under this condition, the system reduces to

$$\pi_{k+1} = \bar{\varphi}^\alpha(\pi_k). \tag{2.14}$$

where $\bar{\varphi}^\alpha : \mathbb{P} \to \mathbb{P}$ is defined as in (2.5). As is typical in the context of singular perturbations, we refer to this dynamics as the *reduced system*. We further assume that the chosen optimization method is reliable. That is, we assume that the set Θ^α of equilibria of (2.14) is nonempty and that it contains the (compact) set M^* of maximizers of (2.4). Finally, we introduce a nonnegative real-valued function $\Delta^\alpha : \mathcal{B} \to \mathbb{R}_{\ge 0}$, defined as

$$\Delta^\alpha(\pi, v) \coloneqq \begin{cases} \frac{1}{\alpha} d_{\mathbb{P}}\big(\varphi^\alpha(\pi, v), \pi\big) & \text{if } \alpha > 0 \\ 0 & \text{if } \alpha = 0 \end{cases} \tag{2.15}$$

together with a continuous residual function $D : \mathbb{P} \to \mathbb{R}_{\ge 0}$ satisfying:

$$D(\pi) \ge \Delta^\alpha(\pi, \bar{v}(\pi)), \quad \forall \pi \in \mathbb{P}. \tag{2.16}$$

We are now ready to characterize the convergence properties of the reduced system (2.14).

[1] Two functions $f, g : I \subset \mathbb{R} \to \mathbb{R}$, defined on a interval $I \ni 0$, are asymptotically equivalent as $x \to 0$ if $\lim_{x\to 0} f(x)/g(x) = 1$.

Assumption 3. *There exist a continuous function* $\mathcal{V}^{\mathbb{P}} : \mathbb{P} \to \mathbb{R}$ *and a constant* $\bar{\alpha}_1 > 0$ *such that, for all* $\alpha \in (0, \bar{\alpha}_1)$*, we have*

$$\mathcal{V}^{\mathbb{P}}\left(\bar{\varphi}^{\alpha}(\pi)\right) - \mathcal{V}^{\mathbb{P}}(\pi) \leq -c_1^{\alpha} D(\pi)^2 \tag{2.17a}$$

$$\mathcal{V}^{\mathbb{P}}\left(\varphi^{\alpha}(\pi, v_1)\right) - \mathcal{V}^{\mathbb{P}}\left(\varphi^{\alpha}(\pi, v_2)\right) \leq c_2^{\alpha} D(\pi) \|v_1 - v_2\|_V + (c_3^{\alpha})^2 \left(\Delta^{\alpha}(\pi, v_1)^2 + \Delta^{\alpha}(\pi, v_2)^2\right), \tag{2.17b}$$

for all $\pi \in \mathbb{P}$ *and* $v_1, v_2 \in V$*, where, for all* $i \in \{1, 2, 3\}$*, it holds*

$$c_i^{\alpha} \sim_A \alpha \bar{c}_i \quad as \quad \alpha \to 0,$$

for some $\bar{c}_i > 0$*.*

Assumption 3 characterizes the convergence properties of the reduced system (2.14) under the lens of LaSalle's Invariance Principle. In particular, it provides a descent-like function $\mathcal{V}^{\mathbb{P}}$ ensuring that, with sufficiently small α, the trajectories of the reduced system (2.14) approach the zero set of the function D as $k \to \infty$. In turn, this means that, if $\omega_k = \bar{\omega}(\pi_k)$, $\zeta_k = \hat{\zeta}(\pi_k, \bar{\omega}(\pi_k))$ at each iteration k, and with sufficiently small α, the optimization mechanism (2.10a) asymptotically converges to the zero set of D.

The Learning Dynamics

We characterize the learning dynamics as a sequence $k \mapsto g_k^{\beta}$ of continuous families of maps $\{g_k^{\beta} : \mathcal{B} \to \Omega\}_{\beta \in [0,1]}$ such that

$$g_k^0(\pi, (\omega, \zeta)) = \omega, \quad \forall \pi \in \mathbb{P}, \omega \in \Omega, \zeta \in Z,$$

namely, g_k^0 is the projection map $\mathcal{P}_{\mathcal{B} \twoheadrightarrow \Omega}$ and thus, for $\beta = 0$, ω_k does not vary in time. Moreover, we make the following continuity assumption.

Assumption 4. *The learning dynamics* g_k^{β} *satisfy*

$$\begin{aligned}
\left\|g_k^{\beta}(\pi_1, v) - g_k^{\beta}(\pi_2, v)\right\|_{\Omega} &\leq L_{g,1}^{\beta} d_{\mathbb{P}}(\pi_1, \pi_2) \\
\left\|g_k^{\beta}(\pi, (\omega_1, \zeta)) - g_k^{\beta}(\pi, (\omega_2, \zeta))\right\|_{\Omega} &\leq L_{g,\omega}^{\beta} \|\omega_1 - \omega_2\|_{\Omega} \\
\left\|g_k^{\beta}(\pi, (\omega, \zeta_1)) - g_k^{\beta}(\pi, (\omega, \zeta_2))\right\|_{\Omega} &\leq L_{g,\zeta}^{\beta} \|\zeta_1 - \zeta_2\|_{Z},
\end{aligned}$$

for all $\pi, \pi_1, \pi_2 \in \mathbb{P}$*,* $v \in V$*,* $\omega_1, \omega_2 \in \Omega$*,* $\zeta_1, \zeta_2 \in Z$*, and* $k \in \mathbb{N}$*, where* $L_{g,1}^{\beta}, L_{g,\omega}^{\beta}, L_{g,\zeta}^{\beta} \geq 0$ *are continuous functions of* β *that are strictly positive for* $\beta > 0$ *and satisfy*

$$\begin{aligned}
&L_{g,1}^{\beta} \sim_A l_1 \beta \quad as \quad \beta \to 0 \\
&L_{g,\omega}^{0} = 1 \\
&L_{g,\zeta}^{\beta} \sim_A l_2 \beta \quad as \quad \beta \to 0,
\end{aligned}$$

for some $l_1, l_2 \geq 0$*.*

Assumption 4 ensures that the parameter β can be used to arbitrarily reduce the speed of variation of ω_k. In order to further characterize the learning dynamics (2.10b), we analyze it under the assumption that the slow state is frozen in time, while the fast state lies on its

equilibrium manifold. That is, we assume $\pi_k = \pi$ and $\zeta_k = \hat{\zeta}(\pi, \omega_k)$ for any $\pi \in \mathbb{P}$ and all $k \in \mathbb{N}$. Under this assumption, the dynamics reduce to

$$\omega_{k+1} = \bar{g}_k^{\beta}\big|_{\pi}(\omega_k), \tag{2.19}$$

where $\bar{g}_k^{\beta}\big|_{\pi} : \Omega \to \Omega$ is defined by

$$\bar{g}_k^{\beta}\big|_{\pi}(\omega_k) \coloneqq g_k^{\beta}\left(\pi, \left(\omega_k, \hat{\zeta}(\pi, \omega_k)\right)\right). \tag{2.20}$$

We refer to (2.20) as *mid auxiliary system*, and we characterize its stability properties as follows.

Assumption 5. *There exists a sequence of continuous functions* $\{\mathcal{V}_k^{\Omega} : \Omega \to \mathbb{R}\}_{k\in\mathbb{N}}$ *and* $\bar{\beta}_1 > 0$ *such that, for all* $\beta \in (0, \bar{\beta}_1)$*, it holds*

$$b_1\|\omega - \bar{\omega}(\pi)\|_{\Omega}^2 \le \mathcal{V}_k^{\Omega}(\omega) \le b_2\|\omega - \bar{\omega}(\pi)\|_{\Omega}^2 \tag{2.21a}$$

$$\mathcal{V}_{k+1}^{\Omega}\left(\bar{g}_k^{\beta}\big|_{\pi}(\omega)\right) - \mathcal{V}_k^{\Omega}(\omega) \le -\beta b_3\|\omega - \bar{\omega}(\pi)\|_{\Omega}^2 \tag{2.21b}$$

$$|\mathcal{V}_k^{\Omega}(\omega_1) - \mathcal{V}_k^{\Omega}(\omega_2)| \le b_4\|\omega_1 - \omega_2\|_{\Omega}\left(\|\omega_1 - \bar{\omega}(\pi)\|_{\Omega} + \|\omega_2 - \bar{\omega}(\pi)\|_{\Omega}\right), \tag{2.21c}$$

for all $k \in \mathbb{N}$, $\pi \in \mathbb{P}$, $\omega, \omega_1, \omega_2 \in \Omega$*, and some positive constants* b_1, b_2, b_3, b_4*.*

Assumption 5 provides a Lyapunov function $\mathcal{V}_k^{\Omega}$ ensuring that $\bar{\omega}(\pi)$ is a globally exponentially stable equilibrium point of the mid dynamics (2.19). Therefore, if $\pi \in \mathbb{P}$ were fixed and $\zeta_k = \hat{\zeta}(\pi, \omega_k)$ for all $k \in \mathbb{N}$, then $\bar{\omega}(\pi)$ would be a globally exponentially stable equilibrium for the learning dynamics (2.10b), uniformly in π.

The Measurements' Dynamics

We finally characterize the measurements' dynamics as a sequence $k \mapsto h_k$ of continuous map $h_k : \mathcal{B} \to Z$ satisfying the following condition.

Assumption 6. *There exist* $L_{h,1}, L_{h,2} > 0$ *such that*

$$\|h_k(\pi_1, v) - h_k(\pi_2, v)\|_Z \le L_{h,1} d_{\mathbb{P}}(\pi_1, \pi_2)$$
$$\|h_k(\pi, v_1) - h_k(\pi, v_2)\|_Z \le L_{h,2}\|v_1 - v_2\|_V$$

for all $\pi, \pi_1, \pi_2 \in \mathbb{P}$ *and* $v, v_1, v_2 \in V$*.*

As done for the optimization and learning dynamics, we characterize the stability properties of the measurements' dynamics (2.10c) individually, by analyzing an associated auxiliary system. In particular, we focus on the so-called *boundary-layer system* obtained by rewriting the measurement dynamics (2.10c) with arbitrarily fixed $\pi_k = \pi \in \mathbb{P}$ and $\omega_k = \omega \in \Omega$ for all $k \in \mathbb{N}$. Then, such a boundary-layer system reads as

$$\zeta_{k+1} = \bar{h}_k\big|_{(\pi,\omega)}(\zeta_k), \tag{2.22}$$

where, for all fixed $\pi \in \mathbb{P}, \omega \in \Omega$, $\bar{h}_k\big|_{(\pi,\omega)} : Z \to Z$ is defined by

$$\bar{h}_k\big|_{(\pi,\omega)}(\zeta_k) \coloneqq h_k(\pi, (\omega, \zeta_k)). \tag{2.23}$$

The stability properties of the boundary-layer system (2.22) are outlined in the following assumption.

Assumption 7. *There exists a sequence of continuous functions* $\{\mathcal{V}_k^Z : Z \to \mathbb{R}\}_{k\in\mathbb{N}}$ *such that*

$$d_1\left\|\zeta - \hat{\zeta}(\pi,\omega)\right\|_Z^2 \le \mathcal{V}_k^Z(\zeta) \le d_2\left\|\zeta - \hat{\zeta}(\pi,\omega)\right\|_Z^2 \tag{2.24a}$$

$$\mathcal{V}_{k+1}^Z\left(\bar{h}_k\big|_{(\pi,\omega)}(\zeta)\right) - \mathcal{V}_k^Z(\zeta) \le -d_3\left\|\zeta - \hat{\zeta}(\pi,\omega)\right\|_Z^2 \tag{2.24b}$$

$$|\mathcal{V}_k^Z(\zeta_1) - \mathcal{V}_k^Z(\zeta_2)| \le d_4\|\zeta_1 - \zeta_2\|_Z\left(\left\|\zeta_1 - \hat{\zeta}(\pi,\omega)\right\|_Z + \left\|\zeta_2 - \hat{\zeta}(\pi,\omega)\right\|_Z\right), \tag{2.24c}$$

for all $k \in \mathbb{N}, \pi \in \mathbb{P}, \omega \in \Omega$, $\zeta, \zeta_1, \zeta_2 \in Z$, *and some positive constants* d_1, d_2, d_3, d_4.

Assumption 6 provides a Lyapunov function $\mathcal{V}_k^Z$ ensuring that $\hat{\zeta}(\pi,\omega)$ is a globally exponentially stable equilibrium point for the boundary-layer system (2.22). Hence, if π and ω were fixed, $\hat{\zeta}(\pi,\omega)$ would be a globally exponentially stable equilibrium for the measurements' dynamics (2.10c), uniformly in π and ω.

Convergence Result

With the individual blocks of (2.10) characterized, we are now ready to determine the convergence properties of the interconnected three-timescale system.

Theorem 1. *Consider system* (2.10) *and let Assumptions 1-7 hold. Then, there exist* $\bar{\alpha}, \bar{\beta} \in (0,1]$ *such that, for all* $\alpha \in (0,\bar{\alpha})$ *and* $\beta \in (0,\bar{\beta})$, *the trajectories* $\{(\pi_k,(\omega_k,\zeta_k))\}_{k\in\mathbb{N}}$ *of* (2.10) *satisfy*

$$\lim_{k\to\infty} \min_{\pi\in\mathcal{S}} d_{\mathbb{P}}(\pi,\pi_k) = 0$$
$$\lim_{k\to\infty} \left\|\left(\omega_k - \bar{\omega}(\pi_k), \zeta_k - \hat{\zeta}(\pi_k,\omega_k)\right)\right\|_V = 0,$$

for all $(\pi_0,\omega_0,\zeta_0) \in \mathcal{B}$, *where*

$$\mathcal{S} \coloneqq \{\pi \in \mathbb{P} : D(\pi) = 0\} \subseteq \Theta^\alpha,$$

with Θ^α *being the equilibrium set of the reduced system* (2.14).

Theorem 1 establishes that, for sufficiently small values of α and β, the closed-loop system (2.10) inherits the same convergence properties of the idealized subsystems (2.14), (2.19), and (2.22). In the next section we present a few intermediate lemmas that will be used in the proof of Theorem 1 at the end of the chapter. In particular, we view the optimization part (2.10a) as the slowest system, the measurements part (2.10c) as the fastest system, and the learning one (2.10b) as the intermediate one.

Intuition behind the Convergence Behavior of the Three-Timescale System

Before proceeding with the proof of Theorem 1, we briefly provide a toy example showing the timescale-separation behavior of system (2.10). Let us consider the following three-dimensional system:

$$\pi_{k+1} = \mathcal{P}_{\mathbb{R}\to[-\pi,\pi]}\left(\pi_k - \alpha\left((\pi_k - \bar{\pi}) + (\omega_k - \sin(\pi_k)) + \left(\zeta_k - \sin(\pi_k)e^{-\omega_k/2}\right)\right)\right) \tag{2.25a}$$

$$\omega_{k+1} = \omega_k - \beta(\omega_k - \sin(\pi_k)) \tag{2.25b}$$

$$\zeta_{k+1} = \zeta_k - \eta\left(\zeta_k - \sin(\pi_k)e^{-\omega_k/2}\right) \tag{2.25c}$$

where π is the mathematical constant pi and $\bar{\pi}$ is a constant in the interval $(-\pi, \pi)$. In this case, the parametrized equilibria are

$$\bar{\omega}(\pi) = \sin(\pi)$$
$$\hat{\zeta}(\pi, \omega) = \sin(\pi)e^{-\omega/2}$$

and, for $\beta, \eta \in (0, 1)$, the mid and fast auxiliary systems satisfy the exponential convergence properties required by Assumptions 5-7. Furthermore, for $\omega_k = \sin(\pi_k)$ and $\zeta_k = \sin(\pi_k)e^{-\omega_k/2}$, the slow system becomes a projected gradient method for the function $1/2\|\pi - \bar{\pi}\|_2^2$. Figure 2.1 shows the convergence behavior of the three-timescale system (2.25) under proper selection of the parameters α and β. The surface represents the graph of the parametrized equilibrium $\hat{\zeta}$, while the blue curve is the section corresponding to the equilibrium manifold $(\pi, \bar{\omega}(\pi))$ of the mid-level system. The black dot marks the point $(\bar{\pi}, \bar{\omega}(\bar{\pi}), \bar{\zeta}(\bar{\pi}))$. The sequence of red dots shows the trajectory of the interconnected system: it first approaches the surface, as ζ evolves faster toward its equilibrium manifold, then it approaches the section of the surface relative to the equilibrium of the mid-scale system, and finally it approaches the black dot.

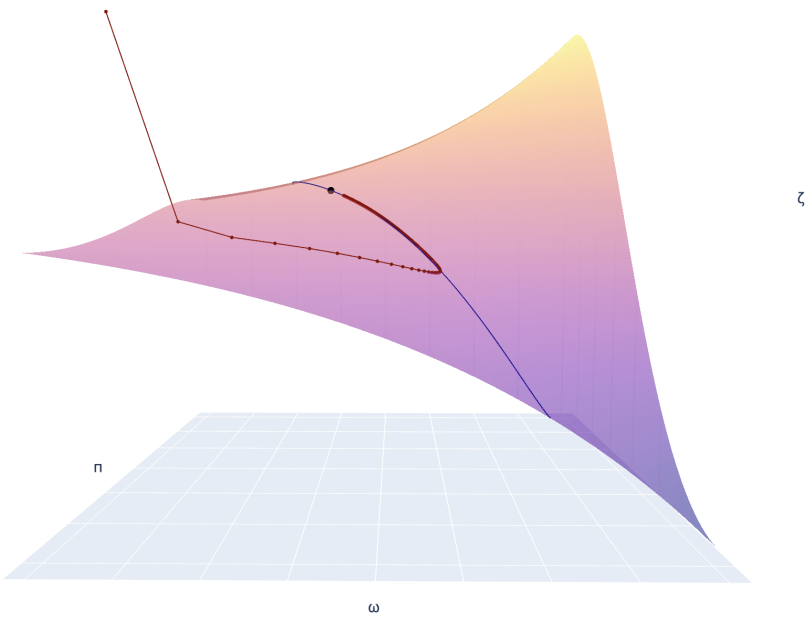


Figure 2.1: Convergence behavior of the three-timescale system (2.25).

Intermediate Lemmas and Considerations

In this section, we provide some intermediate results on system (2.10) that support the proof of Theorem 1, provided at the end of the chapter. The analysis is organized around three main points. First, we further characterize the equilibrium set of the reduced system (2.14). Then we rewrite the overall interconnected system (2.10) in error coordinates with respect to the parametrized equilibria $\bar{\omega}(\pi_k)$ and $\hat{\zeta}(\pi_k, \omega_k)$, and we bound the increment of the Lyapunov

function $\mathcal{V}^{\mathbb{P}}$ along the trajectories of the resulting system. Finally, we bound the increment of the Lyapunov functions $\mathcal{V}_k^{\Omega}$ and $\mathcal{V}_k^{Z}$ along the trajectories of their respective subsystems in the case where π is fixed.

On the Set of Equilibrium-Points of the Optimization Dynamics

In the previous section, we needed to assume the nonemptiness of the set of equilibria of (2.14) because, in general, a Lipschitz continuous map from a compact space to itself does not necessarily admit fixed points. For example, the antipodal map $A : S^N \to S^N, p \mapsto -p$ on the N-sphere is a Lipschitz continuous map (in fact, an isometry), yet it has no fixed points. However, many sufficient conditions for the existence of fixed points are known, depending on the properties of the map. Clearly, by (2.12), for $\alpha = 0$ every point in $\mathbb{P}$ is a fixed point for (2.14). Hence, the reduced dynamics can fail to have equilibria only when $\alpha > 0$. The following lemma shows that, in our case, this assumption can, in fact, be removed, at least for α sufficiently small, since the existence of a special Lyapunov-like function as in Assumption 3 implies the existence of at least one fixed-point.

Lemma 1. *If the reduced system* (2.14) *admits a function* $\mathcal{V}^{\mathbb{P}}$ *satisfying* (2.17a) *for* α *in some interval* $(0, \alpha_1)$ *and for some function* D *satisfying* (2.16)*, then:*

$$\Theta^{\alpha} \neq \varnothing, \quad \forall \alpha \in (0, \alpha_1).$$

In other words, the reduced-system dynamics has at least an equilibrium point in $\mathbb{P}$*.*

Proof. Fix $\alpha \in (0, \alpha_1)$ and assume by contradiction that $\Theta^{\alpha} = \varnothing$. This implies that there exists $\bar{\varepsilon} > 0$ such that

$$\min_{\pi \in \mathbb{P}} D(\pi) = \bar{\varepsilon}.$$

Hence, given a trajectory $\{\pi_k\}_{k \in \mathbb{N}}$ of the reduced system (2.14), consider the sequence $\{\mathcal{V}_k^{\mathbb{P}} \coloneqq \mathcal{V}^{\mathbb{P}}(\pi_k)\}_{k \in \mathbb{N}}$. By (2.17a), it holds

$$\mathcal{V}_{k+1}^{\mathbb{P}} \leq \mathcal{V}_k^{\mathbb{P}} - c_1^{\alpha} D(\pi)^2 \leq \mathcal{V}_k^{\mathbb{P}},$$

which ensures that $\{\mathcal{V}_k^{\mathbb{P}}\}_{k \in \mathbb{N}}$ is bounded from below by $\mathbb{V}_{\min}^{\mathcal{P}} \coloneqq \min_{\pi \in \mathbb{P}} \mathcal{V}^{\mathbb{P}}(\pi)$ and monotonically nonincreasing. Hence, there exists $\mathcal{V}_{\lim}^{\mathbb{P}} \geq \mathcal{V}_{\min}^{\mathbb{P}}$ such that $\mathcal{V}_k^{\mathbb{P}} \to \mathcal{V}_{\lim}^{\mathbb{P}}$ as $k \to \infty$. The fact that $\{\mathcal{V}_k^{\mathbb{P}}\}_{k \in \mathbb{N}}$ is convergent implies that it is a fundamental sequence, i.e., for all $\varepsilon > 0$ there exists $N_{\varepsilon} \in \mathbb{N}$ such that, for all $k \geq N_{\varepsilon}$ and $m \in \mathbb{N}$:

$$0 < \mathcal{V}_k^{\mathbb{P}} - \mathcal{V}_{k+m}^{\mathbb{P}} < \varepsilon.$$

In particular, by choosing $\varepsilon = c_1^{\alpha} \bar{\varepsilon}^2 / 2$ and $m = 1$, then, for all $k > N_{\varepsilon}$, it holds

$$\frac{c_1^{\alpha} \bar{\varepsilon}^2}{2} > \mathcal{V}_k^{\mathbb{P}} - \mathcal{V}_{k+1}^{\mathbb{P}} \geq c_1^{\alpha} D(\pi_k)^2 \geq c_1^{\alpha} \bar{\varepsilon}^2,$$

which is a contradiction. Hence, D has at least one zero in $\mathbb{P}$, which is also a fixed point of $\bar{\varphi}^{\alpha}$ by (2.16). □

The Dynamics in Error Coordinates

In order to better study the stability properties of the mid and slow dynamics, we define the following global trivialization

$$\begin{aligned}\psi : \mathcal{B} &\longrightarrow \mathbb{P} \times \mathbb{R}^{n_v} \\ (\pi, (\omega, \zeta)) &\longmapsto \big(\pi, \psi_V\big(\omega - \bar{\omega}(\pi), \zeta - \hat{\zeta}(\pi, \omega)\big)\big),\end{aligned} \tag{2.26}$$

where $\psi_V : V \to \mathbb{R}^{n_v}$ is an isometric isomorphism. Since ψ is a fiber-preserving homeomorphism with respect to the projection $\mathcal{P}_{\mathcal{B}\twoheadrightarrow\mathbb{P}}$, its restriction to the fiber of any fixed $\pi \in \mathbb{P}$ induces a π-dependent chart $\Gamma_\pi : V \to \mathbb{R}^{n_v}$ such that

$$\begin{aligned} g_k^\beta\left(\psi^{-1}\big(\pi, (\tilde{\omega}, \tilde{\zeta})\big)\right) &= g_k^\beta\left(\pi, \Gamma_\pi^{-1}(\tilde{\omega}, \tilde{\zeta})\right) \\ h_k\left(\psi^{-1}\big(\pi, (\tilde{\omega}, \tilde{\zeta})\big)\right) &= h_k\left(\pi, \Gamma_\pi^{-1}(\tilde{\omega}, \tilde{\zeta})\right). \end{aligned}$$

Hence, we can write the dynamics (2.10) in the so-called error coordinates:

$$\pi_{k+1} = \tilde{\varphi}^\alpha\left(\pi_k, (\tilde{\omega}_k, \tilde{\zeta}_k)\right) \tag{2.27a}$$

$$\tilde{\omega}_{k+1} = \tilde{g}_k^{\alpha\beta}\left(\pi_k, (\tilde{\omega}_k, \tilde{\zeta}_k)\right) \tag{2.27b}$$

$$\tilde{\zeta}_{k+1} = \tilde{h}_k^{\alpha\beta}\left(\pi_k, (\tilde{\omega}_k, \tilde{\zeta}_k)\right), \tag{2.27c}$$

where

$$\begin{aligned} \tilde{\varphi}^\alpha\left(\pi_k, (\tilde{\omega}_k, \tilde{\zeta}_k)\right) &\coloneqq \varphi^\alpha\left(\pi_k, \Gamma_{\pi_k}^{-1}(\tilde{\omega}_k, \tilde{\zeta}_k)\right) \\ \underbrace{\left(\tilde{g}_k^{\alpha\beta}(\pi_k, \cdot), \tilde{h}_k^{\alpha\beta}(\pi_k, \cdot)\right)}_{\coloneqq \tilde{F}_k^{\alpha\beta}(\pi_k, \cdot)} &\coloneqq \Gamma_{\pi_{k+1}^\alpha} \circ \left(g_k^\beta(\pi_k, \cdot), h_k(\pi_k, \cdot)\right) \circ \Gamma_{\pi_k}^{-1}, \end{aligned}$$

in which we made explicit the dependence of π_{k+1} on α in the second definition. Then, we can compactly rewrite the error dynamics (2.27) as

$$\pi_{k+1} = \tilde{\varphi}^\alpha(\pi_k, \tilde{v}_k) \tag{2.28a}$$

$$\tilde{v}_{k+1} = \tilde{F}_k^{\alpha\beta}(\pi_k, \tilde{v}_k). \tag{2.28b}$$

With this notation at hand, we characterize the increment $\mathcal{V}^{\mathbb{P}}(\pi_{k+1}) - \mathcal{V}^{\mathbb{P}}(\pi_k)$ along the trajectories of (2.28a), rather than along those of the reduced system (2.14), with the following lemma.

Lemma 2. *Consider the Lyapunov function $\mathcal{V}^{\mathbb{P}}$ introduced in Assumption 3. Then, for all $\alpha \in (0, \bar{\alpha}_1)$, the trajectories of* (2.28a) *satisfy*

$$\mathcal{V}^{\mathbb{P}}(\pi_{k+1}) - \mathcal{V}^{\mathbb{P}}(\pi_k) \leq -\left(a_{11}(\alpha) D(\pi_k)^2 + a_{22}(\alpha)\|\tilde{v}_k\|_V^2 + 2a_{12}(\alpha) D(\pi_k) \cdot \|\tilde{v}_k\|_V\right)$$

where

$$\begin{aligned} a_{11}(\alpha) &= c_1^\alpha - 2(c_3^\alpha)^2 \\ a_{12}(\alpha) &= -\frac{c_2^\alpha}{2} - (c_3^\alpha)^2 \frac{L_{\tilde{\varphi},2}^\alpha}{\alpha} \\ a_{22}(\alpha) &= -\left(\frac{c_3^\alpha L_{\tilde{\varphi},2}^\alpha}{\alpha}\right)^2. \end{aligned}$$

Proof. First, observe that the error chart (2.26) does not alter the Lipschitz properties of the slow dynamics. In particular, we have

$$\begin{aligned} d_{\mathbb{P}}\left(\tilde{\varphi}^{\alpha}\left(\pi,\tilde{v}_1\right),\tilde{\varphi}^{\alpha}\left(\pi,\tilde{v}_2\right)\right) &\overset{(a)}{\leq} L^{\alpha}_{\varphi,2}\|\Gamma_{\pi}^{-1}(\tilde{v}_1)-\Gamma_{\pi}^{-1}(\tilde{v}_2)\|_{V} \\ &\overset{(b)}{\leq} \underbrace{L^{\alpha}_{\varphi,2}(1+L_{\hat{\zeta},2})}_{=:L^{\alpha}_{\tilde{\varphi},2}}\|\tilde{v}_1-\tilde{v}_2\| \end{aligned} \tag{2.29}$$

where (a) follows directly from Assumption 2 and (b) follows from the definition of Γ and Assumption 1. Then, we evaluate the increment $\mathcal{V}^{\mathbb{P}}(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}))-\mathcal{V}^{\mathbb{P}}(\pi)$ by adding and subtracting $\mathcal{V}^{\mathbb{P}}(\bar{\varphi}^{\alpha}(\pi))$ thus obtaining, for all $\pi\in\mathbb{P},\tilde{v}\in\mathbb{R}^{n_v}$

$$\begin{aligned} \mathcal{V}^{\mathbb{P}}(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}))-\mathcal{V}^{\mathbb{P}}(\pi) &= \mathcal{V}^{\mathbb{P}}(\bar{\varphi}^{\alpha}(\pi))-\mathcal{V}^{\mathbb{P}}(\pi)+\mathcal{V}^{\mathbb{P}}(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}))-\mathcal{V}^{\mathbb{P}}(\bar{\varphi}^{\alpha}(\pi)) \\ &\overset{(a)}{=} -c_1^{\alpha}D(\pi)^2+\mathcal{V}^{\mathbb{P}}(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}))-\mathcal{V}^{\mathbb{P}}(\bar{\varphi}^{\alpha}(\pi)) \\ &\overset{(b)}{\leq} -c_1^{\alpha}D(\pi)^2+c_2^{\alpha}D(\pi)\|\tilde{v}\|+(c_3^{\alpha})^2\big(\tilde{\Delta}^{\alpha}(\pi,\tilde{v})^2+\tilde{\Delta}^{\alpha}(\pi,0)^2\big) \\ &\overset{(c)}{\leq} -c_1^{\alpha}D(\pi)^2+c_2^{\alpha}D(\pi)\|\tilde{v}\|+(c_3^{\alpha})^2\tilde{\Delta}^{\alpha}(\pi,\tilde{v})^2+(c_3^{\alpha})^2D(\pi)^2, \end{aligned} \tag{2.30}$$

where in (a) we use (2.17a) (cf. Assumption 3), in (b) we leveraged $\bar{\varphi}^{\alpha}(\pi)=\tilde{\varphi}^{\alpha}(\pi,0)$ to use (2.17b) (cf. Assumption 3), while (c) follows from (2.16). Now, observe that, for $\alpha>0$, the triangle inequality allows us to write

$$\begin{aligned} \tilde{\Delta}^{\alpha}(\pi,\tilde{v}) &\leq \frac{1}{\alpha}d_{\mathbb{P}}\big(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}),\bar{\varphi}^{\alpha}(\pi)\big)+\frac{1}{\alpha}d_{\mathbb{P}}\big(\bar{\varphi}^{\alpha}(\pi),\pi\big) \\ &\overset{(a)}{\leq} \frac{L^{\alpha}_{\tilde{\varphi},2}}{\alpha}\|\tilde{v}\|_{V}+\frac{d_{\mathbb{P}}\big(\bar{\varphi}^{\alpha}(\pi),\pi\big)}{\alpha} \\ &\overset{(b)}{\leq} \frac{L^{\alpha}_{\tilde{\varphi},2}}{\alpha}\|\tilde{v}\|+D(\pi) \end{aligned}$$

where in (a) we used (2.29), together with the identity $\bar{\varphi}^{\alpha}(\pi)=\tilde{\varphi}^{\alpha}(\pi,0),\forall\pi\in\mathbb{P}$, while in (b) we used (2.16). The statement follows by using this inequality and developing the square. □

The Two-Scale Boundary-Layer System

Consider the auxiliary system obtained by setting $\alpha=0$ in (2.28), i.e., by considering π fixed in time and by letting the faster states evolve freely. Hence, the auxiliary dynamics reads as

$$\begin{aligned} \tilde{\omega}_{k+1} &= \tilde{g}_k^{0\beta}\left(\pi,\left(\tilde{\omega}_k,\tilde{\zeta}_k\right)\right) \\ \tilde{\zeta}_{k+1} &= \tilde{h}_k^{0\beta}\left(\pi,\left(\tilde{\omega}_k,\tilde{\zeta}_k\right)\right), \end{aligned}$$

where

$$\underbrace{\left(\tilde{g}_k^{0\beta}(\pi,\cdot),\tilde{h}_k^{0\beta}(\pi,\cdot)\right)}_{=:\tilde{F}_k^{0\beta}(\pi,\cdot)} = \Gamma_{\pi}\circ\left(g_k^{\beta}(\pi,\cdot),h_k(\pi,\cdot)\right)\circ\Gamma_{\pi}^{-1}$$

Compactly:

$$\tilde{v}_{k+1}=\tilde{F}_k^{0\beta}(\pi,\tilde{v}_k)$$

Remark 1. *Note that, in this chart, the identities* (2.11a)-(2.11b) *read as*

$$\tilde{g}_k^{0\beta}(\pi, 0) = 0 \tag{2.31a}$$
$$\tilde{h}_k^{0\beta}(\pi, (\tilde{\omega}, 0)) = 0, \tag{2.31b}$$

for all $\tilde{\omega} \in \mathbb{R}^{n_\omega}, \pi \in \mathbb{P}, k \in \mathbb{N}$, *and* $\beta \in [0,1]$. *In particular, by choosing* $\tilde{\omega} = 0$, *we also have*

$$\tilde{F}_k^{0\beta}(\pi, 0) = 0, \tag{2.32}$$

for all $\pi \in \mathbb{P}, k \in \mathbb{N}$, *and* $\beta \in [0,1]$.

Then, the auxiliary dynamics and the exact one are related by the transition function:

$$T^{\pi}_{\tilde{\varphi}^\alpha(\pi,\tilde{v})} \coloneqq \Gamma_{\tilde{\varphi}^\alpha(\pi,\tilde{v})} \circ \Gamma_\pi^{-1}$$

leading to:

$$\tilde{F}_k^{\alpha\beta}(\pi, \tilde{v}) = T^{\pi}_{\tilde{\varphi}^\alpha(\pi,\tilde{v})} \circ \tilde{F}_k^{0\beta}(\pi, \tilde{v})$$

Note that, in general, $T^{\pi_1}_{\pi_2}$ is a homeomorphism of $\mathbb{R}^{n_v}$ onto itself (i.e. a transition function between charts of V), induced by a homeomorphism from the fiber of π_1 to the fiber of π_2. We can explicitly write the transition function as

$$T^{\pi_1}_{\pi_2} : (\tilde{\omega}, \tilde{\zeta}) \mapsto \big(\tilde{\omega} + e^{\Omega}_{\pi_1,\pi_2},\ \tilde{\zeta} + e^{Z}_{\pi_1,\pi_2}(\tilde{\omega})\big),$$

where we introduced

$$\begin{aligned} e^{\Omega}_{\pi_1,\pi_2} &\coloneqq \bar{\omega}(\pi_1) - \bar{\omega}(\pi_2) \\ e^{Z}_{\pi_1,\pi_2}(\tilde{\omega}) &\coloneqq \hat{\zeta}(\pi_1, \tilde{\omega} + \bar{\omega}(\pi_1)) - \hat{\zeta}(\pi_2, \tilde{\omega} + \bar{\omega}(\pi_1)), \end{aligned}$$

which will be referred to as *coordinate errors*. It follows from the Lipschitz continuity of $\bar{\omega}$ and $\hat{\zeta}$ (Assumption 1) that the errors are bounded as follows

$$\begin{aligned} \big\| e^{\Omega}_{\pi_1,\pi_2} \big\|_\Omega &\le L_{\bar{\omega}}\, d_{\mathbb{P}}(\pi_2, \pi_1) \\ \big\| e^{Z}_{\pi_1,\pi_2}(\tilde{\omega}) \big\|_Z &\le L_{\hat{\zeta},1}\, d_{\mathbb{P}}(\pi_2, \pi_1), \end{aligned}$$

from which we have, for all $\tilde{v} \in \mathbb{R}^{n_v}$ and for all $\pi_1, \pi_2 \in \mathbb{P}$

$$\big\| T^{\pi_1}_{\pi_2}(\tilde{v}) - \tilde{v} \big\|_V \le \underbrace{\max\Big\{ L_{\bar{\omega}}, L_{\hat{\zeta}} \Big\}}_{=: L_e} d_{\mathbb{P}}(\pi_2, \pi_1). \tag{2.33}$$

We then describe the true dynamics in terms of the auxiliary one, in error coordinates, as

$$\begin{aligned} \tilde{F}_k^{\alpha\beta}(\pi, \tilde{v}) &= T^{\pi}_{\tilde{\varphi}^\alpha(\pi,\tilde{v})}\big(\tilde{F}_k^{0\beta}(\pi, \tilde{v})\big) \\ &= T^{\pi}_{\tilde{\varphi}^\alpha(\pi,\tilde{v})}\big(\tilde{g}_k^{0\beta}(\pi, \tilde{v}), \tilde{h}_k^{0\beta}(\pi, \tilde{v})\big) \\ &= \big(\tilde{g}_k^{0\beta}(\pi, \tilde{v}) + e^{\Omega}_{\pi,\pi^+_\alpha}, \tilde{h}_k^{0\beta}(\pi, \tilde{v}) + e^{Z}_{\pi,\pi^+_\alpha}\big(\tilde{g}_k^{0\beta}(\pi, \tilde{v})\big)\big)\big|_{\pi^+_\alpha = \tilde{\varphi}^\alpha(\pi,\tilde{v})}. \end{aligned} \tag{2.34}$$

Eventually, the following lemma holds for the two-scale boundary-layer system.

Lemma 3. *Under Assumptions 4-5-7, there exists $\bar{\beta}_2 > 0$ such that, for all $\beta \in (0, \bar{\beta}_2)$ and $\pi \in \mathbb{P}$, there exists a sequence of continuous functions $\{\tilde{\mathcal{V}}_k^{aux} : \mathbb{R}^{n_v} \to \mathbb{R}\}_{k\in\mathbb{N}}$ and $k_1, k_2, k_3, k_4 > 0$ such that*

$$k_1\|\tilde{v}\|^2 \le \tilde{\mathcal{V}}_k^{aux}(\tilde{v}) \le k_2\|\tilde{v}\|^2 \tag{2.35a}$$

$$\tilde{\mathcal{V}}_{k+1}^{aux}\big(\tilde{F}_k^{0\beta}(\pi, \tilde{v})\big) - \tilde{\mathcal{V}}_k^{aux}(\tilde{v}) \le -k_3\|\tilde{v}\|^2 \tag{2.35b}$$

$$\big|\tilde{\mathcal{V}}_k^{aux}(\tilde{v}_1) - \tilde{\mathcal{V}}_k^{aux}(\tilde{v}_2)\big| \le k_4\|\tilde{v}_1 - \tilde{v}_2\|\left(\|\tilde{v}_1\| + \|\tilde{v}_2\|\right), \tag{2.35c}$$

for all $\pi \in \mathbb{P}$, $\tilde{v}, \tilde{v}_1, \tilde{v}_2 \in \mathbb{R}^{n_v}$ and $k \in \mathbb{N}$.

Proof. Let $\tilde{\mathcal{V}}_k^{\Omega}$ and $\tilde{\mathcal{V}}_k^{Z}$ be the representations of $\mathcal{V}^{\Omega}$ and $\mathcal{V}^{Z}$, respectively, in error coordinates. Then, let us define

$$\tilde{\mathcal{V}}_k^{\text{aux}}(\tilde{\omega}, \tilde{\zeta}) \coloneqq \tilde{\mathcal{V}}_k^{\Omega}(\tilde{\omega}) + \tilde{\mathcal{V}}_k^{Z}(\tilde{\zeta}),$$

for all $k \in \mathbb{N}$ and let $\tilde{v} = (\tilde{\omega}, \tilde{\zeta})$. By recalling that $\|(\tilde{\omega}, \tilde{\zeta})\| = \|\tilde{\omega}\| + \|\tilde{\zeta}\|$, the conditions (2.35a) and (2.35c) directly follow from the corresponding inequalities (2.21a) (cf. Assumption 5) and (2.24a) (cf. Assumption 7) and (2.21c) (cf. Assumption 5) and (2.24c) (cf. Assumption 7), respectively, and by setting $k_1 \coloneqq \frac{1}{2}\min\{b_1, d_1\}$, $k_2 \coloneqq \max\{b_1, d_2\}$, and $k_4 \coloneqq \max\{b_4, d_4\}$. In order to check (2.35b), we apply (2.35c) to write the preparatory bound

$$\begin{aligned}
&\tilde{\mathcal{V}}_{k+1}^{\Omega}\big(\tilde{g}_k^{\beta 0}\big(\pi, (\tilde{\omega}, \tilde{\zeta})\big)\big) - \tilde{\mathcal{V}}_{k+1}^{\Omega}\big(\tilde{g}_k^{\beta 0}(\pi, (\tilde{\omega}, 0))\big) \\
&\quad \le b_4\big\|\tilde{g}_k^{\beta 0}\big(\pi, \big(\tilde{\omega}, \tilde{\zeta}\big)\big)\big\| \cdot \big\|\tilde{g}_k^{\beta 0}\big(\pi, \big(\tilde{\omega}, \tilde{\zeta}\big)\big)\big\| + b_4\big\|\tilde{g}_k^{\beta 0}\big(\pi, \big(\tilde{\omega}, \tilde{\zeta}\big)\big)\big\| \cdot \big\|\tilde{g}_k^{\beta 0}(\pi, (\tilde{\omega}, 0))\big\| \\
&\quad \overset{(a)}{\le} b_4 L_{g,\zeta}^{\beta}\big\|\tilde{\zeta}\big\| \cdot \big(2L_{g,\omega}^{\beta}\|\tilde{\omega}\| + L_{g,\zeta}^{\beta}\big\|\tilde{\zeta}\big\|\big) \\
&\quad = 2b_4 L_{g,\omega}^{\beta} L_{g,\zeta}^{\beta}\|\tilde{\omega}\|\big\|\tilde{\zeta}\big\| + b_4\left(L_{g,\zeta}^{\beta}\right)^2\big\|\tilde{\zeta}\big\|^2,
\end{aligned} \tag{2.36}$$

where in (a) we use the bounds in Assumption 4 together with the fact that $\tilde{g}_k^{\beta 0}(\pi, (0,0)) = 0$. With this bound at hand, we now add and subtract the term $\tilde{\mathcal{V}}_{k+1}^{\Omega}\big(\tilde{g}_k^{\beta 0}(\pi, (\tilde{\omega}, 0))\big)$ to the increment $\tilde{\mathcal{V}}_{k+1}^{\text{aux}}\big(\tilde{F}_k^{0\beta}\big(\pi, \big(\tilde{\omega}, \tilde{\zeta}\big)\big)\big) - \tilde{\mathcal{V}}_k^{\text{aux}}\big(\big(\tilde{\omega}, \tilde{\zeta}\big)\big)$, thus obtaining

$$\begin{aligned}
\tilde{\mathcal{V}}_{k+1}^{\text{aux}}\big(\tilde{F}_k^{0\beta}\big(\pi, \big(\tilde{\omega}, \tilde{\zeta}\big)\big)\big) - \tilde{\mathcal{V}}_k^{\text{aux}}\big(\big(\tilde{\omega}, \tilde{\zeta}\big)\big) &= \tilde{\mathcal{V}}_{k+1}^{\Omega}\big(\tilde{g}_k^{\beta 0}\big(\pi, \big(\tilde{\omega}, \tilde{\zeta}\big)\big)\big) - \tilde{\mathcal{V}}_{k+1}^{\Omega}\big(\tilde{g}_k^{\beta 0}(\pi, (\tilde{\omega}, 0))\big) \\
&\quad + \tilde{\mathcal{V}}_{k+1}^{\Omega}\big(\tilde{g}_k^{\beta 0}(\pi, (\tilde{\omega}, 0))\big) - \tilde{\mathcal{V}}_k^{\Omega}(\tilde{\omega}) \\
&\quad + \tilde{\mathcal{V}}_{k+1}^{Z}\big(\tilde{h}_k^{\beta 0}\big(\pi, \big(\tilde{\omega}, \tilde{\zeta}\big)\big)\big) - \tilde{\mathcal{V}}_k^{Z}\big(\tilde{\zeta}\big) \\
&\overset{(a)}{\le} 2b_4 L_{g,\omega}^{\beta} L_{g,\zeta}^{\beta}\|\tilde{\omega}\|\big\|\tilde{\zeta}\big\| + b_4\left(L_{g,\zeta}^{\beta}\right)^2\big\|\tilde{\zeta}\big\|^2 \\
&\quad - \beta b_3\|\tilde{\omega}\|^2 - d_3\big\|\tilde{\zeta}\big\|^2 \\
&\overset{(b)}{=} -q^{\beta}\big(\|\tilde{\omega}\|, \big\|\tilde{\zeta}\big\|\big),
\end{aligned}$$

where in (a) we use (2.36) to bound the first two terms, while in (b) we introduce the quadratic form $q^{\beta} : \mathbb{R}^2 \to \mathbb{R}$, identified by the matrix

$$Q^{\beta} \coloneqq \begin{bmatrix} \beta b_3 & -b_4 L_{g,\omega}^{\beta} L_{g,\zeta}^{\beta} \\ -b_4 L_{g,\omega}^{\beta} L_{g,\zeta}^{\beta} & d_3 - b_4\left(L_{g,\zeta}^{\beta}\right)^2 \end{bmatrix}.$$

By the asymptotic properties of $L^{\beta}_{g,\omega}$ and $L^{\beta}_{g,\zeta}$ as $\beta \to 0$ (see Assumption 4), there exists $\bar{\beta}_2 > 0$ such that, for all $\beta \in (0, \bar{\beta}_2)$, it holds

$$\begin{aligned}\beta b_3 d_3 - b_3 b_4 \beta \left(L^{\beta}_{g,\zeta}\right)^2 - b_4^2 \left(L^{\beta}_{g,\omega}\right)^2 \left(L^{\beta}_{g,\zeta}\right)^2 &= \beta(b_3 d_3 + o(\beta)) \\ &> 0,\end{aligned}$$

which, together with $\beta b_3 > 0$, implies, by the Sylvester criterion, the positive definiteness of the matrix Q^{β} and, thus, of the quadratic form q^{β}. In turn, this implies that there exists a real number $\tilde{k}_3 > 0$ such that

$$\begin{aligned}\tilde{\mathcal{V}}^{\text{aux}}_{k+1}\big(\tilde{F}^{0\beta}_k\left(\pi, (\tilde{\omega}, \tilde{\zeta})\right)\big) - \tilde{\mathcal{V}}^{\text{aux}}_k\big((\tilde{\omega}, \tilde{\zeta})\big) &\leq -\tilde{k}_3\big(\|\tilde{\omega}\|^2 + \|\tilde{\zeta}\|^2\big) \\ &\leq -\frac{\tilde{k}_3}{2}\big(\|\tilde{\omega}\| + \|\tilde{\zeta}\|\big)^2 \\ &= -\frac{\tilde{k}_3}{2}\|(\tilde{\omega}, \tilde{\zeta})\|^2,\end{aligned}$$

where we used the inequality $(a+b)^2 \leq 2(a^2+b^2), \forall a, b \in \mathbb{R}$. The proof concludes by setting $k_3 = \tilde{k}_3/2$. □

We also provide the following technical lemma:

Lemma 4. *If $\{(\pi_k, \tilde{v}_k)\}_{k\in\mathbb{N}}$ is a trajectory for system (2.28) and $\{\tilde{\mathcal{V}}^{aux}_k : \mathbb{R}^{n_v} \to \mathbb{R}\}_{k\in\mathbb{N}}$ is the sequence of Lyapunov functions on $\mathbb{R}^{n_v}$ as in Lemma 3, then:*

$$\tilde{\mathcal{V}}^{aux}_{k+1}(\tilde{v}_{k+1}) - \tilde{\mathcal{V}}^{aux}_k(\tilde{v}_k) \leq b_{11}(\alpha) D(\pi_k)^2 + b_{22}(\alpha)\|\tilde{v}_k\|^2 + 2b_{12}(\alpha) D(\pi_k) \cdot \|\tilde{v}_k\|,$$

where

$$\begin{aligned}b_{11}(\alpha) &= \alpha^2 k_4 L_e^2 \\ b_{12}(\alpha) &= \alpha k_4 L_e\big(L^{\beta}_{\tilde{F}} + L_e L^{\alpha}_{\tilde{\varphi},2}\big) \\ b_{22}(\alpha) &= k_4 L_e L^{\alpha}_{\tilde{\varphi},2}\big(2L^{\beta}_{\tilde{F}} + L_e L^{\alpha}_{\tilde{\varphi},2}\big) - k_3.\end{aligned}$$

Proof. First, we note that, by (2.32) and Assumptions 4 and 6, it holds

$$\begin{aligned}\|\tilde{F}^{0\beta}_k(\pi, 0)\| &= \|\tilde{F}^{0\beta}_k(\pi, \tilde{v}) - \tilde{F}^{0\beta}_k(\pi, 0)\| \\ &\leq L^{\beta}_{\tilde{F}}\|\tilde{v}\|,\end{aligned} \tag{2.37}$$

with

$$L^{\beta}_{\tilde{F}} = \max\left\{L^{\beta}_{g,\omega}, L^{\beta}_{g,\zeta}, L_{h,2}\right\}(1 + L_{\hat{\zeta},2}). \tag{2.38}$$

Now, we evaluate the increment of $\tilde{\mathcal{V}}^{\text{aux}}_k$ along the trajectories of system (2.28). By adding

and subtracting $\tilde{\mathcal{V}}_{k+1}^{\text{aux}}\big(\tilde{F}_k^{\alpha\beta}(\pi,\tilde{v})\big)$, we get

$$\begin{aligned}
\tilde{\mathcal{V}}_{k+1}^{\text{aux}}\big(\tilde{F}_k^{\alpha\beta}(\pi,\tilde{v})\big) - \tilde{\mathcal{V}}_k^{\text{aux}}(\tilde{v}) &= \tilde{\mathcal{V}}_{k+1}^{\text{aux}}\big(\tilde{F}_k^{\alpha\beta}(\pi,\tilde{v})\big) - \tilde{\mathcal{V}}_{k+1}^{\text{aux}}\big(\tilde{F}_k^{0\beta}(\pi,\tilde{v})\big) + \tilde{\mathcal{V}}_{k+1}^{\text{aux}}\big(\tilde{F}_k^{0\beta}(\pi,\tilde{v})\big) - \tilde{\mathcal{V}}_k^{\text{aux}}(\tilde{v}) \\
&\overset{(a)}{\leq} \tilde{\mathcal{V}}_{k+1}^{\text{aux}}\big(\tilde{F}_k^{\alpha\beta}(\pi,\tilde{v})\big) - \tilde{\mathcal{V}}_{k+1}^{\text{aux}}\big(\tilde{F}_k^{0\beta}(\pi,\tilde{v})\big) - k_3\|\tilde{v}\|^2 \\
&\overset{(b)}{\leq} k_4 L_e d_{\mathbb{P}}\big(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}),\pi\big)\cdot\big\|\tilde{F}_k^{\alpha\beta}(\pi,\tilde{v})\big\| \\
&\quad + k_4 L_e d_{\mathbb{P}}\big(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}),\pi\big)\cdot\big\|\tilde{F}_k^{0\beta}(\pi,\tilde{v})\big\| - k_3\|\tilde{v}\|^2 \\
&\overset{(c)}{\leq} 2k_4 L_e d_{\mathbb{P}}\big(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}),\pi\big)\cdot\big\|\tilde{F}_k^{0\beta}(\pi,\tilde{v})\big\| \\
&\quad + k_4 L_e^2 d_{\mathbb{P}}\big(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}),\pi\big)^2 - k_3\|\tilde{v}\|^2 \\
&\overset{(d)}{\leq} 2k_4 L_e L_{\tilde{F}}^{\beta} d_{\mathbb{P}}\big(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}),\pi\big)\cdot\|\tilde{v}\| \\
&\quad + k_4 L_e^2 d_{\mathbb{P}}\big(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}),\pi\big)^2 - k_3\|\tilde{v}\|^2,
\end{aligned}$$

where (a) follows from Lemma 3, (b) follows from of Lemma 3 and (2.33), (c) follows from (2.33)-(2.34) and the triangle inequality, while in (d) we use (2.37). Now, the statement follows by considering that:

$$\begin{aligned}
d_{\mathbb{P}}\big(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}),\pi\big) &\leq d_{\mathbb{P}}\big(\tilde{\varphi}^{\alpha}(\pi,\tilde{v}),\bar{\varphi}^{\alpha}(\pi)\big) + d_{\mathbb{P}}\big(\bar{\varphi}^{\alpha}(\pi),\pi\big) \\
&\overset{(a)}{\leq} L_{\tilde{\varphi},2}^{\alpha}\|\tilde{v}\|_V + d_{\mathbb{P}}\big(\bar{\varphi}^{\alpha}(\pi),\pi\big),
\end{aligned}$$

where in (a) we use (2.29), and the fact that, by (2.16), it holds

$$d_{\mathbb{P}}\big(\bar{\varphi}^{\alpha}(\pi_k),\pi_k\big) \leq \alpha D(\pi_k).$$

□

Proof of Main Theorem

First, if we take $\alpha \in (0,\alpha_1)$ (with α_1 as in Assumption 3), by Lemma 1 the set Θ^{α} is nonempty. Now, consider as a candidate time-varying Lyapunov function on $\mathcal{B}$ the sequence of maps $\{\tilde{\mathcal{V}}_k\}_{k\in\mathbb{N}}$ each defined as

$$\begin{aligned}
\tilde{\mathcal{V}}_k : \mathbb{P}\times\mathbb{R}^{n_v} &\to \mathbb{R} \\
(\pi,\tilde{v}) &\mapsto \mathcal{V}^{\mathbb{P}}(\pi) + \tilde{\mathcal{V}}_k^{\text{aux}}(\tilde{v}),
\end{aligned}$$

where $\mathcal{V}^{\mathbb{P}}$ is the Lyapunov function of the reduced system introduced in Assumption 3 and $\{\tilde{\mathcal{V}}_k^{\text{aux}}\}_{k\in\mathbb{N}}$ is the sequence of Lyapunov functions resulting from Lemma 3 by choosing β sufficiently close to 0. By evaluating its variation along a trajectory $\{(\pi_k,\tilde{v}_k)\}_{k\in\mathbb{N}}$ of the system in the error coordinates 2.26, we have, by combining Lemma 2 and Lemma 4:

$$\tilde{\mathcal{V}}_{k+1}(\pi_{k+1},\tilde{v}_{k+1}) - \tilde{\mathcal{V}}_k(\pi_k,\tilde{v}_k) \leq -q^{\alpha}\big(D(\pi_k),\|\tilde{v}_k\|\big), \tag{2.39}$$

where $q^{\alpha}:\mathbb{R}^2\to\mathbb{R}$ is the quadratic function defined as

$$q^{\alpha}(a,b) \coloneqq h_{11}(\alpha)a^2 + h_{22}^{\beta}(\alpha)b^2 + 2h_{12}^{\beta}(\alpha)ab, \tag{2.40}$$

with coefficients defined by

$$
\begin{aligned}
h_{11}(\alpha) &= c_1^\alpha - 2(c_3^\alpha)^2 - \alpha^2 k_4 L_e^2 \\
h_{12}^\beta(\alpha) &= -\frac{c_2^\alpha}{2} - (c_3^\alpha)^2 \frac{L_{\tilde{\varphi},2}^\alpha}{\alpha} - \alpha k_4 L_e \big(L_{\tilde{F}}^\beta + L_e L_{\tilde{\varphi},2}^\alpha\big) \\
h_{22}^\beta(\alpha) &= k_3 - k_4 L_e L_{\tilde{\varphi},2}^\alpha \big(2L_{\tilde{F}}^\beta + L_e L_{\tilde{\varphi},2}^\alpha\big) - \left(\frac{c_3^\alpha L_{\tilde{\varphi},2}^\alpha}{\alpha}\right)^2.
\end{aligned}
$$

By the Sylvester criterion, the quadratic function q^α (cf. (2.40)) is positive definite if and only if

$$h_{11}(\alpha) > 0 \tag{2.41}$$

$$h_{11}(\alpha) h_{22}^\beta(\alpha) - \big(h_{12}^\beta(\alpha)\big)^2 > 0. \tag{2.42}$$

Now, observe that, as $\alpha \to 0$:

$$
\begin{aligned}
h_{11}(\alpha) &\sim_{\mathrm{A}} \alpha(\bar{c}_1 + o(\alpha)) \\
h_{12}^\beta(\alpha) &\sim_{\mathrm{A}} -\frac{\alpha}{2}(\bar{c}_2 + 2k_4 L_e L_{\tilde{F}}^\beta + o(\alpha)) \\
h_{22}^\beta(\alpha) &= k_3 + o(\alpha),
\end{aligned}
$$

and

$$h_{11}(\alpha) h_{22}^\beta(\alpha) - \big(h_{12}^\beta(\alpha)\big)^2 \sim_{\mathrm{A}} k_3 \bar{c}_1 \alpha + o(\alpha^2),$$

from which it follows that there exists $\bar{\alpha} \in (0, \bar{\alpha}_1)$ such that both (2.41)-(2.42) are verified. Hence, by choosing $\alpha \in (0, \bar{\alpha})$, we use (2.39) to claim

$$\tilde{\mathcal{V}}_{k+1}(\pi_{k+1}, \tilde{v}_{k+1}) \leq \tilde{\mathcal{V}}_k(\pi_k, \tilde{v}_k). \tag{2.43}$$

Moreover, in light of the first condition of Lemma 3, we have

$$k_1 \|\tilde{v}\|^2 + \mathcal{V}^{\mathbb{P}}(\pi) \leq \tilde{\mathcal{V}}_k(\pi, \tilde{v}) \leq k_2 \|\tilde{v}\|^2 + \mathcal{V}^{\mathbb{P}}(\pi), \tag{2.44}$$

for all $\pi \in \mathbb{P}, \tilde{v} \in \mathbb{R}^{n_v}, k \in \mathbb{N}$. Now, for any fixed $c > 0$, let Ω^c be the set

$$\Omega^c = \bigcap_{k \in \mathbb{N}} \Omega_k^c,$$

where

$$\Omega_k^c \coloneqq \Big\{(\pi, \tilde{v}) \in \mathbb{P} \times \mathbb{R}^{n_v} : \tilde{\mathcal{V}}_k(\pi, \tilde{v}) \leq c + \max_{x \in \mathbb{P}} \mathcal{V}^{\mathbb{P}}(x)\Big\}.$$

For fixed k, observe first that Ω_k^c is closed, as it is the preimage of a closed interval via a continuous map. Then, by the left-hand side inequality of 2.44, we have:

$$k_1 \|\tilde{v}\|_V^2 + \min_{x_1 \in \mathbb{P}} \mathcal{V}^{\mathbb{P}}(x_1) \leq \mathcal{V}_k(\pi, \tilde{v}) \leq c + \max_{x_2 \in \mathbb{P}} \mathcal{V}^{\mathbb{P}}(x_2),$$

for all $(\pi, \tilde{v}) \in \Omega_k^c$. Thus, by defining

$$\bar{c} \coloneqq c + \max_{x_2 \in \mathbb{P}} \mathcal{V}^{\mathbb{P}}(x_2) - \min_{x_1 \in \mathbb{P}} \mathcal{V}^{\mathbb{P}}(x_1),$$

we have

$$\Omega_k^c \subseteq \mathbb{P} \times \left\{ \tilde{v} \in \mathbb{R}^{n_v} : k_1 \|\tilde{v}\|^2 \leq \bar{c} \right\},$$

where the right-hand side is a compact subset of $\mathcal{B}$. Hence, as a closed subset of a compact set, the set Ω_k^c is compact. Then, also Ω^c is compact since it is the intersection of compact subsets of a metric space. Now, in light of (2.43), the set Ω^c is positively invariant with respect to the three-scale dynamics for all fixed $c > 0$. If $\{(\pi_k, \tilde{v}_k)\}_{k\in\mathbb{N}}$ is a trajectory of the system starting in Ω^c, the induced sequence $\{\tilde{\mathcal{V}}_k(\pi_k, \tilde{v}_k)\}_{k\in\mathbb{N}}$ is bounded from below, since Ω^c is compact, and monotonically nonincreasing by (2.43), and thus admits a finite limit. Note also that, by (2.39), it holds

$$\begin{aligned}
\sum_{\tau=0}^{k-1} q^\alpha(D(\pi_k), \|\tilde{v}_k\|) &\leq -\sum_{\tau=0}^{k-1} \left(\tilde{\mathcal{V}}_{\tau+1}(\pi_{\tau+1}, \tilde{v}_{\tau+1}) - \tilde{\mathcal{V}}_\tau(\pi_\tau, \tilde{v}_\tau) \right) \\
&= \tilde{\mathcal{V}}_0(\pi_0, \tilde{v}_0) - \tilde{\mathcal{V}}_k(\pi_k, \tilde{v}_k) \\
&\leq \tilde{\mathcal{V}}_0(\pi_0, \tilde{v}_0) - \lim_{k\to\infty} \tilde{\mathcal{V}}_k(\pi_k, \tilde{v}_k).
\end{aligned}$$

Since the summation on the left-hand side is monotonically nondecreasing in k (recall that q^α is positive definite for the chosen α) and bounded from above, it is convergent as $k \to \infty$, and this implies that its summand converges to zero, i.e.:

$$\lim_{k\to\infty} q^\alpha(D(\pi_k), \|\tilde{v}_k\|) = 0.$$

Eventually, the sign-definiteness of q^α, together with the previous limit, implies that:

$$\begin{aligned}
\lim_{k\to\infty} \|\tilde{v}_k\| &= 0 \\
\lim_{k\to\infty} D(\pi_k) &= 0,
\end{aligned}$$

where the last limit implies that:

$$\lim_{k\to\infty} \min_{\pi\in\Theta^\alpha} d_{\mathbb{P}}\left(\pi, \pi_k\right) = 0,$$

for any fixed $\alpha > 0$, and this concludes the proof.

Chapter 3

Analysis and Design of Actor-Critic Methods using Timescale Separation

In this chapter, we introduce two extensions of the actor-critic paradigm that incorporate an additional update step representing a measurement mechanism. The two methods differ in the type of measurement unit employed, and both can be tracked back to the class of problems analyzed in Chapter 2. We begin by presenting the common actor-critic framework and providing the necessary assumptions for the subsequent analyses. Then we introduce the first variant, based on a measurement unit with access to the MDP model, and we show that it satisfies all the assumptions required to apply the results of the previous chapter. Next, we introduce the second variant, which relies on a model-free measurement unit and whose averaged dynamics also belong to the aforementioned class. Finally, we provide simulation results on a common benchmark.
The first part of this work was presented at the 63rd Conference on Decision and Control (CDC) in 2024 and published as a conference paper [30].

Literature Review

Reinforcement learning has proven to be a powerful framework for sequential decision-making, especially in the presence of high dimensional Markov decision processes. Due to the curse of dimensionality associated with such problems (see, e.g., [31]), effective RL methods commonly rely on function approximation techniques to represent key quantities, such as policy and value functions, using lower-dimensional parametrizations. This approach is applied in both finite state spaces [32, 33] and infinite state spaces [34]. Many RL algorithms follow the classical policy evaluation and improvement paradigm, leading to complex nonlinear coupled dynamics, particularly when function approximation is involved. Notable examples are policy gradients and actor-critic methods [35, 36]. The work [37] provides a comprehensive overview of policy gradient methods, covering various parametrization classes. Early works such as [38] analyzed the properties of temporal-difference (TD) learning methods with linear function approximators, providing the theoretical foundation for value estimation in actor-critic methods. On the other hand, other studies focused on critic-only approaches without the presence of an actor mechanism, such as TD(λ) [39] and Q-learning [40].

Framework Description

In this framework, our goal is to find a policy π solving the maximization problem

$$\max_{\pi \in \Gamma_s(\mathcal{X},\mathcal{U})} J_\pi^w, \tag{3.1}$$

where $J_\pi^w : \Gamma_s(\mathcal{X},\mathcal{U}) \to \mathbb{R}$ is a performance index defined as

$$J_\pi^w \coloneqq \mathbb{E}\left[v_{\pi,r}(x_0) \middle|\ x_0 \sim w \right], \tag{3.2}$$

with w being a given initial state distribution and $v_{\pi,r} : \mathcal{X} \to \mathbb{R}$ being the value function. Now, due to the high dimensionality of the MDPs which are usually encountered in reinforcement learning applications, it is customary to introduce parametrization maps to manipulate the needed quantities by using a considerably smaller amount of variables. In particular, we restrict the policies to the image of a parametrization map

$$\wp : \mathbb{R}^{n_\theta} \hookrightarrow \Gamma_s(\mathcal{X},\mathcal{U}) \tag{3.3}$$

$$\theta \mapsto \pi_\theta \tag{3.4}$$

which we assume to satisfy the following property.

Assumption 8. *The policy parametrization $\wp$ is a smooth embedding, $L_\wp$-Lipschitz and its partial derivatives satisfy*

$$\frac{\partial \wp}{\partial \theta^i} = \psi_i \circ \wp, \quad \forall i \in \{1, \cdots, n_\theta\}, \tag{3.5}$$

for some smooth maps $\psi_i : \Gamma_s(\mathcal{X},\mathcal{U}) \to Map(\mathcal{U},\mathbb{R}) \simeq \mathbb{R}^m$. Furthermore, it satisfies the following inequality:

$$dJ^w|_{\pi_\theta}\big(\wp(\theta_1) - \wp(\theta_2)\big) \leq c \left\| \nabla_\theta J_\wp^w \big|_\theta \right\| \cdot \|\theta_1 - \theta_2\|, \tag{3.6}$$

for all $\theta, \theta_1, \theta_2 \in \mathbb{R}^{n_\theta}$ and some $c > 0$.

Moreover, instead of manipulating every entry of the Q-function to evaluate a given policy, we model it as a linear function:

$$\hat{q}_\omega = \sum_{i=1}^{n_\omega} \omega^i \phi_i$$

where $\omega \in \mathbb{R}^{n_\omega}$ a low-dimensional vector of tunable parameters, and $\phi_i : \mathcal{U} \to \mathbb{R}, i \in \{1, \cdots, n_\omega\}$, are fixed basis functions. For analysis purposes we also make the following assumption.

Assumption 9. *There exists a continuous function $\bar{\omega} : \kappa(\wp(\mathbb{R}^{n_\theta})) \to \mathbb{R}^{n_\omega}$ such that*

$$q_{\pi,r} = \hat{q}_{\bar{\omega}(\pi)}, \quad \forall \pi \in \Gamma_s(\mathcal{X},\mathcal{U}).$$

Furthermore, it holds

$$\|\bar{\omega}(\pi) - \bar{\omega}(\pi')\| \leq L_\omega d(\pi, \pi'), \quad \forall \pi, \pi' \in \kappa(\wp(\mathbb{R}^{n_\theta})) \tag{3.7}$$

for some $L_\omega > 0$, where d denotes the metric inherited[1] *by $\Gamma_s(\mathcal{X},\mathcal{U})$ from $\mathbb{R}^{n \times m}$.*

[1] The metric d is defined as $d(\pi, \pi') = \|\Pi - \Pi'\|$. For notational convenience, in the remainder of the chapter we will write $\|\pi - \pi'\|$ with slight abuse of notation.

Ideal Policy Improvement

The parametrization (3.3) allows us to apply the renowned policy gradient theorem [32] to compute each partial derivative of $J^w_\wp \coloneqq J^w \circ \wp$ with respect to the policy parameters as

$$\begin{aligned}\left.\frac{\partial J^w_\wp}{\partial \theta^i}\right|_\theta &= \sum_{t=0}^{\infty} \sum_{x\in\mathcal{X}} \gamma^t p_{t,\pi_\theta}(x) \sum_{u\in\mathcal{U}_x} q_{\pi_\theta,r}(u)\big[\psi_i \circ \pi_\theta\big](u) \\ &= \sum_{u\in\mathcal{U}} \left(\sum_{t=0}^{\infty} \gamma^t \big[p_{t,\pi_\theta} \circ s\big](u)\right) q_{\pi_\theta,r}(u)\big[\psi_i \circ \pi_\theta\big](u) \end{aligned} \tag{3.8}$$

for all $i \in \{1,\dots,n_\theta\}$, where p_{t,π_θ} is a time-parametric probability distribution on $\mathcal{X}$ assigning to each state x the probability that the MDP reaches x at time t after following the policy π_θ. If $W \in \mathbb{R}^n$ represents the initial state distribution, then the distribution p_{t,π_θ} is represented by the probability vector

$$\big(F^\top_{\pi_\theta}\big)^t W.$$

Furthermore, we make the following assumption.

Assumption 10. *The gradient of $J^w_\wp$ is L_V-Lipschitz continuous in θ, for some $L_V > 0$.*

Hence, assuming that the model of the MDP is available and that we can compute the Q-function for every policy π_{θ_k}, we can update the policy through a policy gradient method:

$$\theta_{k+1} = \theta_k + \alpha_1 \nabla_\theta J^w_\wp\big|_{\theta_k} \tag{3.9}$$

where each component of the gradient $\nabla_\theta J^w_\wp\big|_{\theta_k}$ is defined by (3.8), evaluated at $\theta = \theta_k$.

Ideal Policy Evaluation

The Q-function, which is required for computing the gradient, can be obtained by solving the Bellman equations (1.6). In particular, the uniqueness of their solution enables us to quantify the discrepancy between any ω and the unique $\bar{\omega}(\pi)$ solving the equations, through the Bellman error function $e(\cdot|\omega,\pi) : \mathcal{U} \to \mathbb{R}$ defined as

$$e(u|\omega,\pi) \coloneqq r(u) + \gamma \sum_{z\in\mathcal{X}} f(z,u) \sum_{v\in\mathcal{U}_z} \pi(v,z)\hat{q}_\omega(v) - \hat{q}_\omega(u). \tag{3.10}$$

Hence, we could update each parameter ω^i with a linear functional of the Bellman error function, with the corresponding basis function acting as the coefficient function. Specifically:

$$\omega^i_{k+1} = \omega^i_k + \alpha_2 \sum_{u\in\mathcal{U}} e(u|\omega_k,\pi)\phi_i(u), \quad \forall i \in \{1,\cdots,n_\omega\}. \tag{3.11}$$

By combining the updates in (3.11), the overall update for ω can be expressed in a more compact form as

$$\omega_{k+1} = \big(I_{n_\omega} - \alpha_2 A(\pi_\theta)\big)\omega_k + \alpha_2 b, \tag{3.12}$$

where, we have introduced

$$A(\pi_\theta) \coloneqq \Phi^\top\big(I_m - \gamma F \Pi_\theta\big)\Phi, \quad b \coloneqq \Phi^\top R,$$

with $\Phi \in \mathbb{R}^{m\times n_\omega}$ being a matrix whose i-th column represents the i-th basis function. Furthermore, we make the following technical assumption on $A(\pi_\theta)$.

Assumption 11. *There exist constants $\mu, L > 0$ such that*

$$\langle \omega,\ A(\pi_\theta)\,\omega \rangle \geq \mu \|\omega\|_2^2$$
$$\|A(\pi_\theta)\,\omega\|_2^2 \leq L\|\omega\|_2^2,$$

for all $\theta \in \mathbb{R}^{n_\theta}$ and $\omega \in \mathbb{R}^{n_\omega}$.

Model-based Actor-Critic

Instead of running (3.11) until convergence for every intermediate policy π_{θ_k}, we can update both parameters concurrently, provided that the policy evaluation update makes progress at a faster rate than the policy improvement update. This gives rise to the following interconnected system

$$\theta_{k+1} = \theta_k + \alpha_1 \sum_{u \in \mathcal{U}} a\left(u \middle| \pi_{\theta_k}, \omega_k\right), \tag{3.13a}$$
$$\omega_{k+1} = \omega_k + \alpha_2 \sum_{u \in \mathcal{U}} c\left(u \middle| \omega_k, \pi_{\theta_k}\right), \tag{3.13b}$$

where, for notational convenience, we introduced the shorthand terms $a\left(u \middle| \pi_\theta, \omega\right) \in \mathbb{R}^{n_\theta}$ and $c\left(u \middle| \omega, \pi_\theta\right) \in \mathbb{R}^{n_\omega}$, whose components are defined as:

$$a^i\left(u \middle| \pi_\theta, \omega\right) \coloneqq \left(\sum_{t=0}^{\infty} \gamma^t \left[p_{t,\pi_\theta} \circ s\right](u)\right) \hat{q}_\omega(u)\left[\psi_i \circ \pi_\theta\right](u), \tag{3.14a}$$
$$c^i\left(u \middle| \omega, \pi_\theta\right) \coloneqq e(u|\omega, \pi)\phi_i(u). \tag{3.14b}$$

In fact, the theory of timescale separation guarantees that, under proper assumptions, for any fixed stepsize α_2 for which (3.11) converges with θ held fixed, there exists a sufficiently small stepsize $\alpha_1 > 0$ for which the interconnected system (3.13) converges to its equilibrium manifold. In particular, θ_k converges to the set of stationary points of J_φ^w. The first challenge with the updates in (3.13) lies in computing, for each action, the term inside the brackets in (3.14a) and the Bellman error function in (3.14b). However, even if an oracle provided both of these quantities, the updates in (3.13) would still be computationally expensive due to the two summations running over the entire action space. In the next sections, we first address the case where these terms can be computed or measured, so that the summations remain the only computational bottleneck. We then turn to the more challenging setting in which they are not directly accessible. As we will see, while the former scenario admits a deterministic algorithm, the latter requires resorting to a stochastic method. In other words, we pay computational efficiency with randomness. Before proceeding, we briefly discuss a suitable candidate for the policy parametrization.

Brief Discussion about the Softmax Function

In this subsection we discuss the softmax function, one of the most commonly used differentiable parametrizations in reinforcement learning and, more broadly, in machine learning, as it stands out as a promising candidate for designing a policy parametrization that satisfies the conditions required in this chapter.

The softmax in $\mathbb{R}^{N+1}$ is defined as

$$\sigma : \mathbb{R}^{N+1} \to \Delta^N \subset \mathbb{R}^{N+1} \tag{3.15}$$

$$p \mapsto \sigma(p) \tag{3.16}$$

where

$$\Delta^N = \left\{ q \in \mathbb{R}^{N+1} : \sum_{j=1}^{N+1} q^j = 1,\ q^i \geq 0\ \forall i \right\}$$

denotes the N-simplex and

$$[\sigma(p)]^i = \frac{e^{p^i}}{\sum_{j=1}^{N+1} e^{p^j}}. \tag{3.17}$$

It is clear from the definition that σ is smooth (in fact, analytic), since it is a composition of smooth (analytic) functions and its denominator never vanishes. Moreover, it holds:

$$\left.\frac{\partial \sigma^k}{\partial p^i}\right|_{\bar{p}} = \delta^\kappa(i,k) \frac{e^{\bar{p}^k}}{\sum_{j=1}^{N+1} e^{\bar{p}^j}} - \frac{e^{\bar{p}^k} e^{\bar{p}^i}}{\left(\sum_{j=1}^{N+1} e^{\bar{p}^j}\right)^2} \tag{3.18}$$

$$= \sigma(\bar{p})^k \left(\delta^\kappa(i,k) - \sigma(\bar{p})^i\right) \tag{3.19}$$

$$= \left[\left(\psi_i \circ \sigma\right)(\bar{p})\right]^k, \tag{3.20}$$

where

$$\psi_i \coloneqq \left(y^1\left(\delta^\kappa(i,1) - y^i\right), \cdots, y^{N+1}\left(\delta^\kappa(i,(N+1)) - y^i\right)\right),$$

with y^j being the j-th standard coordinate of $\mathbb{R}^{N+1}$. Thus, σ satisfies condition (3.5) of Assumption 8. Furthermore, it is Lipschitz (see, e.g., [41, Proposition 4]) and its image lies in the interior[2] of Δ^N since its entries are always positive. However, the softmax is not an embedding of $\mathbb{R}^{N+1}$, as it maps $\mathbb{R}^{N+1}$ to an N-submanifold (see, e.g., [42, Corollary 8.7] or [43, Theorem 17.26]). In particular, in the chart $q \mapsto (q^1, \cdots, q^N)$ for the interior of Δ^N, the Jacobian of σ is given by

$$J_\sigma(q) = \begin{pmatrix} q^1(1-q^1) & -q^1 q^2 & \cdots & -q^1 q^N & -q^1 q^{N+1} \\ -q^2 q^1 & q^2(1-q^2) & \cdots & -q^2 q^N & -q^2 q^{N+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -q^N q^1 & -q^N q^2 & \cdots & q^N(1-q^N) & -q^N q^{N+1} \end{pmatrix},$$

[2]Here by *interior* we intend the manifold interior of Δ^N viewed as a submanifold with corners of $\mathbb{R}^{N+1}$, which differs from its topological interior, which is empty in $\mathbb{R}^{N+1}$. For a formal definition, see [42, Chapter 22.2].

where $q = \sigma(p)$. Now, the principal $(N \times N)$-submatrix of $J_\sigma(q)$ is symmetric, has positive diagonal entries, and is strictly diagonally dominant since:

$$\begin{aligned}
[J_\sigma(q)]_i^i - \sum_{j\neq i} \left|[J_\sigma(q)]_j^i\right| &= q^i\left(1-q^i\right) - \sum_{\substack{j\neq i\\ 1\leq j\leq N}} q^i q^j \\
&= q^i\left(1-q^i\right) - \sum_{\substack{j\neq i\\ 1\leq j\leq N}} q^i q^j \pm q^i q^{N+1} \\
&= q^i\left(1-q^i\right) - \sum_{\substack{j\neq i\\ 1\leq j\leq N+1}} q^i q^j + q^i q^{N+1} \\
&\overset{(a)}{=} q^i\left(1-q^i\right) - q^i\left(1-q^i\right) + q^i q^{N+1} \\
&= q^i q^{N+1} \\
&\overset{(b)}{>} 0,
\end{aligned}$$

where in (a) we used the fact that $\sum_{j=1}^{N+1} q^j = 1$ for all $q \in \Delta^{N+1}$, while (b) holds since q lies in the interior of the simplex. Hence, by [44, Theorem 6.1.10] this submatrix is positive definite for all q in the interior of Δ^N and the softmax is therefore a (surjective) smooth submersion. In particular, its level sets are smooth 1-submanifolds of $\mathbb{R}^{N+1}$, by [42, Theorem 11.2]. Furthermore, the softmax is also invariant under uniform shift of its arguments:

$$\sigma(p + \alpha \mathbb{1}_{N+1}) = \sigma(p), \quad \forall \alpha \in \mathbb{R}, \tag{3.21}$$

and then its level sets are affine 1-dimensional subspaces, cosets of the vector subspace $V \coloneqq \operatorname{span}\{\mathbb{1}_{N+1}\}$. This shows that we can actually take the quotient $\mathbb{R}^{N+1}/V \simeq \mathbb{R}^N$, making σ a diffeomorphism onto its image (which is the interior of Δ^N), and thus an embedding.
For example, consider the case $N = 3$. We can restrict the softmax to the linear subspace $V^\perp$ orthogonal to $V = \operatorname{span}\{\mathbb{1}_3\}$. By choosing, for instance, the orthonormal basis

$$B = \left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{6}} \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \right\} \tag{3.22}$$

we can identify $V^\perp$ with $\mathbb{R}^2$ via the coordinate chart $\varphi_1 = (\alpha, \beta)$ associated with the basis B. On the other hand, we can define the chart $\varphi_2 : (a, b, c) \mapsto (a, b)$ on the interior of the 2-simplex Δ^2. In these coordinates, the softmax reads as:

$$\left(\varphi_2 \circ \sigma \circ \varphi_1^{-1}\right)(\alpha, \beta) = \frac{1}{e^{-\frac{\alpha}{\sqrt{2}}} + e^{\frac{\alpha}{\sqrt{2}}} + e^{\sqrt{3/2}\beta}} \left(e^{-\frac{\alpha}{\sqrt{2}}},\ e^{\frac{\alpha}{\sqrt{2}}}\right), \tag{3.23}$$

which has a smooth inverse given by:

$$(a, b) \mapsto \left(-\frac{g_2(a,b)}{\sqrt{2}},\ \sqrt{2/3}\, g_1(a,b) - \frac{g_2(a,b)}{\sqrt{6}}\right),$$

where

$$g_1(a, b) \coloneqq \ln\left(\frac{1-a-b}{b}\right) \tag{3.24}$$

$$g_2(a, b) \coloneqq \ln\left(\frac{a}{1-a}\left(1 + \frac{1-a-b}{b}\right)\right), \tag{3.25}$$

both defined whenever $a \in (0,1)$, $b \in (0,1)$ and $c = 1 - a - b \in (0,1)$. Figure 3.1 represents a homotopy between the identity map $\mathrm{id}_{\mathbb{R}^2}$ and the embedding (3.23) of $\mathbb{R}^2$ into the 2-simplex (thought of as a subset of $\mathbb{R}^2$).
This discussion suggests that a policy parametrization defined, for all $x \in \mathcal{X}$, as

$$\wp(\theta)|_{\mathcal{U}_x} = \sigma(\phi_x(\theta)), \tag{3.26}$$

for some embedding $\phi_x : \mathbb{R}^{n_\theta} \hookrightarrow \mathbb{R}^{m_x}$, is a promising candidate within the framework presented in this chapter, provided that it satisfies the relationship with the objective gradient given in (3.6).

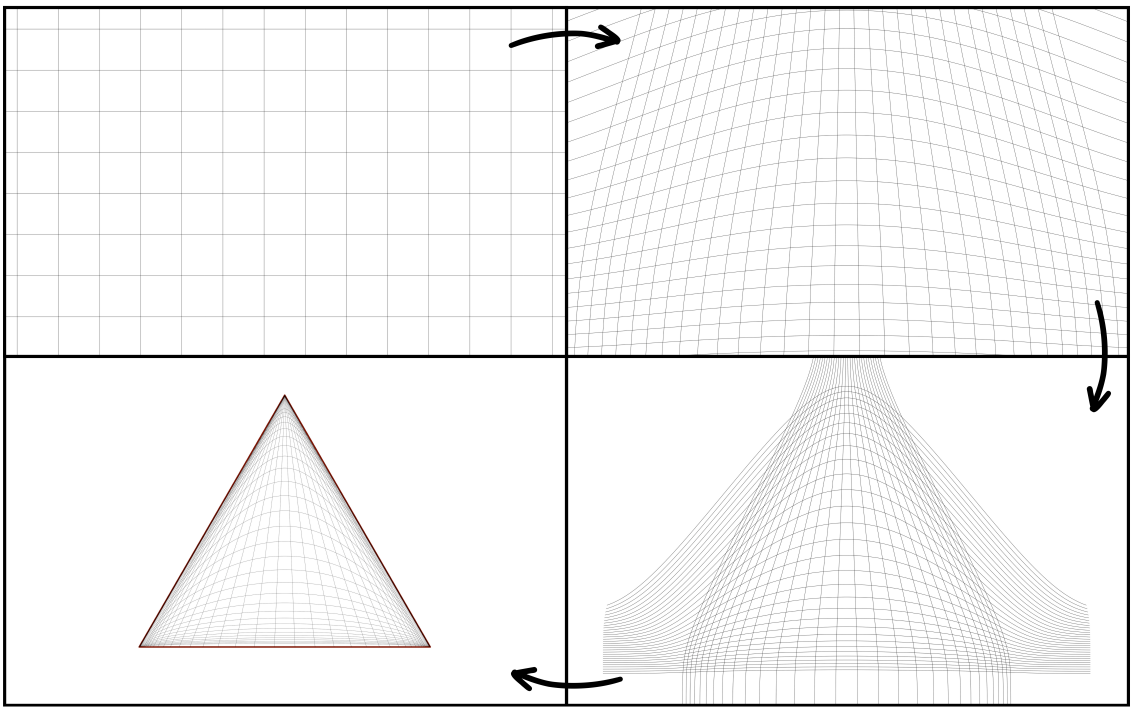


Figure 3.1: Representation of the embedding of $\mathbb{R}^2$ into the 2-simplex induced by the softmax function. Specifically, the figure illustrates four instances of a homotopy between the identity map $\mathrm{id}_{\mathbb{R}^2}$ and the softmax embedding, enabled by treating the 2-simplex as a (topological) subspace of $\mathbb{R}^2$.

Model-based Sampled-Average Actor-Critic

In this section we introduce Model-based Sampled-Average Actor-Critic, a model-based variant of the actor-critic algorithm designed to efficiently address problem (3.1) in cases where the terms a and c are either measurable or easily computable. The main difference with

standard methods stands in the introduction of two auxiliary variables for each disjoint action. Their role is to accumulate samples of (3.14a) and of (3.14b) as they become available, and they are used to update θ and ω respectively. This approach is inspired by works on incremental stochastic gradient methods such as [45, 46]. The proposed method is outlined in Algorithm 1 and will be derived step-by-step in the next subsections. The sequences $\sigma^z : \mathbb{N} \to \mathcal{U}, \sigma^y : \mathbb{N} \to \mathcal{U}$ are used to select the actions (see (3.27c)-(3.27d)) to be considered at each step and are characterized as follows.

Assumption 12. *There exists $T \in \mathbb{N}$ such that, $\sigma^z_{k_1} = u$ and $\sigma^y_{k_2} = u$ for some $k_1, k_2 \in \{k_0, \cdots, k_0 + T - 1\}$ for all $k_0 \in \mathbb{N}$ and $u \in \mathcal{U}$.*

In other words, we are asking for the sequences σ^z_k and σ^y_k to persistently meet every action with a frequency of at least $1/T$.

Algorithm 1 Model-based Sampled-Average Actor-Critic

for $k = 0, 1, 2 \ldots$ **do**

Policy Improvement Step

$$\theta_{k+1} = \theta_k + \alpha_1 \sum_{j=1}^{m} z_{j,k} \tag{3.27a}$$

Policy Evaluation Step

$$\omega_{k+1} = \omega_k + \alpha_2 \sum_{j=1}^{m} y_{j,k} \tag{3.27b}$$

Auxiliary Variables Update

for $j \in \{1, \cdots, m\}$ **do**:

$$z_{j,k+1} = \begin{cases} a\left(\sigma^z_k \middle| \pi_{\theta_k}, \omega_k\right) & \text{if } \mathcal{I}_{\mathcal{U}}(\sigma^z_k) = j \\ z_{j,k} & \text{otherwise} \end{cases} \tag{3.27c}$$

$$y_{j,k+1} = \begin{cases} c\left(\sigma^y_k \middle| \omega_k, \pi_{\theta_k}\right) & \text{if } \mathcal{I}_{\mathcal{U}}(\sigma^y_k) = j \\ y_{j,k} & \text{otherwise} \end{cases} \tag{3.27d}$$

A Timescale Separation Interpretation

We now interpret Algorithm 1 as a *singularly perturbed system* in the form (2.10). To this end, we first group the subschemes of 1 in three classes depending on their "speed":

- ➢ Slow Subsystem: made by the dynamics (3.27a)
- ➢ Mid Subsystem: made by the dynamics (3.27b)
- ➢ Fast Subsystem: made by (3.27c)-(3.27d).

This classification is motivated by the presence of the parameters α_1 and α_2, which allow us to arbitrarily reduce the variation rate of the dynamics (3.27a)-(3.27b), and the fact that (3.27b),

(3.27c), and (3.27d), have equilibria parametrized in the slower states. Indeed, $c(\omega_k, \pi_{\theta_k})$ and $a(\pi_{\theta_k}, \omega_k)$ are equilibria of (3.27d) and (3.27c), respectively, for all $(\omega_k, \pi_{\theta_k}) \in \mathbb{R}^{n_\omega} \times \wp(\mathbb{R}^{n_\theta})$. Furthermore, by setting $\omega_k = c(\omega_k, \pi_{\theta_k})$ for all k, we note that $\bar{\omega}(\pi_{\theta_k})$ is an equilibrium of (3.27b) for all $\theta_k \in \mathbb{R}^{n_\theta}$ (cf. Assumption 9). Along this perspective, we will separately study three auxiliary schemes related to the fast, the mid, and the slow subsystem, respectively.

Analysis of the Fast Subsystem

We now consider the fast system (3.27c)-(3.27d), ignoring the drift of the slower dynamics, i.e., by considering $(\theta_k, \omega_k) = (\theta, \omega)$ for all k and using the error coordinates

$$s_k^z \coloneqq \mathcal{I}_{\mathcal{U}}(\sigma_k^z), \;\; s_k^y \coloneqq \mathcal{I}_{\mathcal{U}}(\sigma_k^y) \tag{3.28a}$$

$$\tilde{z}_{j,k} \coloneqq z_{j,k} - \delta^{\kappa}(j, s_k^z)\, a\left(\sigma_k^z \middle| \pi_\theta, \omega\right), \tag{3.28b}$$

$$\tilde{y}_{j,k} \coloneqq y_{j,k} - \delta^{\kappa}(j, s_k^y)\, c\left(\sigma_k^y \middle| \omega, \pi_\theta\right), \tag{3.28c}$$

for all $j \in \{1, \cdots, m\}$. Thus, if $\tilde{z}_k$ and $\tilde{y}_k$ denote the stack of the auxiliary variables in error coordinates, we can rewrite the dynamics (3.27c)-(3.27d) as

$$\begin{pmatrix} \tilde{z}_{k+1} \\ \tilde{y}_{k+1} \end{pmatrix} = \begin{pmatrix} \tilde{z}_k \\ \tilde{y}_k \end{pmatrix} - \begin{pmatrix} S_k & 0 \\ 0 & V_k \end{pmatrix} \begin{pmatrix} \tilde{z}_k \\ \tilde{y}_k \end{pmatrix}, \tag{3.29}$$

where we introduced two diagonal selector matrices:

$$S_k \coloneqq \mathrm{diag}_j\left\{\delta^{\kappa}(j, s_k^z) I_{n_\theta}\right\}, \tag{3.30a}$$

$$V_k \coloneqq \mathrm{diag}_j\left\{\delta^{\kappa}(j, s_k^y) I_{n_\omega}\right\}. \tag{3.30b}$$

The next lemma establishes the global exponential stability of the origin for (3.29).

Lemma 5. *There exists a continuous function $\mathcal{V}^Z : \mathbb{R}^{m(n_\theta + n_\omega)} \to \mathbb{R}$ that satisfies* (2.24) *along the trajectories of* (3.29)*.*

Proof. Under assumption 12, it holds $\tilde{z}_k = 0$ and $\tilde{y}_k = 0$ for all $k \geq T$. In fact $\tilde{z}_j$ is set to zero the first time $\mathcal{I}_{\mathcal{U}}(\sigma_k^z) = j$, remaining unchanged henceforth, and the same happens to $\tilde{y}_j$ when $\mathcal{I}_{\mathcal{U}}(\sigma_k^y) = j$ for the first time. Hence, for all $\beta \in (0, 1)$ and regardless of the initial conditions, it holds

$$\|\tilde{z}_k\| \leq \|\tilde{z}_0\| \beta^{k-T},$$
$$\|\tilde{y}_k\| \leq \|\tilde{y}_0\| \beta^{k-T}.$$

The statement follows from the converse Lyapunov theorem (see, e.g., [47, Theorem 3.8]). □

Analysis of the Mid Subsystem

We now study the behavior of the mid subsystem (3.27b) by considering $\theta_k = \theta$ and $y_{j,k} = c\left(u \middle| \omega_k, \pi_\theta\right)$ for all $u \in \mathcal{U}$ and $k \geq 0$. Namely, in view of (3.12), we analyze the auxiliary dynamics

$$\omega_{k+1} = \big(I_{n_\omega} - \alpha_2 A(\pi_\theta)\big)\omega_k + \alpha_2 b, \tag{3.31}$$

By introducing the error coordinates

$$\tilde{\omega}_k \coloneqq \omega_k - \bar{\omega}\left(\pi_\theta\right) \tag{3.32}$$

with respect of the equilibrium $\bar{\omega}\left(\pi_\theta\right)$, we rewrite system (3.31) as

$$\tilde{\omega}_{k+1} = \left(I_{n_\omega} - \alpha_2 A\left(\pi_\theta\right)\right)\tilde{\omega}_k. \tag{3.33}$$

Lemma 6. *There exist a continuous function $\mathcal{V}^\Omega : \mathbb{R}^{n_\omega} \to \mathbb{R}$ and constant $\bar{\alpha}_2 > 0$ such that, for all $\alpha \in (0, \bar{\alpha}_2)$, W_ζ satisfies* (2.21) *along the trajectories of system* (3.33)*.*

Proof. Consider the error coordinates (3.32) and let $\tilde{\mathcal{V}}^\Omega(\tilde{\omega}) \coloneqq \frac{1}{2}\|\tilde{\omega}\|^2$. Besides (2.21a), which is trivially satisfied, $\tilde{\mathcal{V}}^\Omega$ satisfies (2.21c) by the triangle inequality:

$$\begin{aligned}
|\tilde{\mathcal{V}}^\Omega(\tilde{\omega}_1) - \tilde{\mathcal{V}}^\Omega(\tilde{\omega}_2)| &= \left|\|\tilde{\omega}_1\|^2 - \|\tilde{\omega}_2\|^2\right| \\
&= \left|\|\tilde{\omega}_1\| - \|\tilde{\omega}_2\|\right| \cdot \left(\|\tilde{\omega}_1\| + \|\tilde{\omega}_2\|\right) \\
&\leq \|\tilde{\omega}_1 - \tilde{\omega}_2\| \cdot \left(\|\tilde{\omega}_1\| + \|\tilde{\omega}_2\|\right).
\end{aligned}$$

Furthermore, evaluating its increment along the trajectories of (3.33), yields

$$\begin{aligned}
\tilde{\mathcal{V}}^\Omega(\tilde{\omega}_{k+1}) - \tilde{\mathcal{V}}^\Omega(\tilde{\omega}_k) &= -\frac{\alpha_2}{2}\left\langle \tilde{\omega}_k,\ \left(A\left(\pi\right) + A\left(\pi\right)^\top - \alpha_2 A\left(\pi\right)^\top A\left(\pi\right)\right)\tilde{\omega}_k \right\rangle \\
&\overset{(a)}{\leq} -\frac{\alpha_2}{2}\left(\mu - \alpha_2 L\right)\|\tilde{\omega}_k\|^2,
\end{aligned} \tag{3.34}$$

where (a) uses Assumption 11. Now, if we arbitrarily choose $\tilde{\mu} \in (0, \mu)$, and we set $\bar{\alpha}_2 \coloneqq \frac{(\mu - \tilde{\mu})}{L}$, we obtain

$$\tilde{\mathcal{V}}^\Omega(\tilde{\omega}_{k+1}) - \tilde{\mathcal{V}}^\Omega(\tilde{\omega}_k) \leq -\alpha_2 \tilde{\mu}\|\tilde{\omega}_k\|^2, \tag{3.35}$$

for all $\alpha_2 \in (0, \bar{\alpha}_2)$, which shows that also (2.21b) is satisfied. □

The "Pushforward" of the Slow Dynamics

Before analyzing the slow system individually, we leverage the fact that the parametrization map $\wp$ is an embedding to continuously transfer the dynamics from the parameter space $\mathbb{R}^{n_\theta}$ to the policy space $\Gamma_s(\mathcal{X}, \mathcal{U})$. Here the policy space is treated as a compact subspace of $\mathbb{R}^{n \times m}$ and $\wp$ acts as a compactification of $\mathbb{R}^{n_\theta}$. This reformulation is also motivated by the fact that the mid and the fast dynamics depend directly on π_θ, and only indirectly on the parameter vector θ. Hence, by defining $\pi_k \coloneqq \wp(\theta_k)$, we can rewrite the slow dynamics as:

$$\pi_{k+1} = \hat{\varphi}^{\alpha_1}\left(\pi_k, z\right), \tag{3.36}$$

where the map $\hat{\varphi}^{\alpha_1} : \wp(\mathbb{R}^{n_\theta}) \times \mathbb{R}^{m n_\theta} \to \wp(\mathbb{R}^{n_\theta})$ is defined by

$$\hat{\varphi}^{\alpha_1}\big(\pi, z\big) \coloneqq \wp\left(\wp^{-1}(\pi) + \alpha_1 \sum_{j=1}^{m} z_j\right).$$

We also make the following technical assumption on the induced policy dynamics.

Assumption 13. *The map $\hat{\varphi}^{\alpha_1} : \wp(\mathbb{R}^{n_\theta}) \times \mathbb{R}^{mn_\theta} \to \wp(\mathbb{R}^{n_\theta})$ is uniformly Lipschitz continuous in both arguments.*

This assumption assures that $\hat{\varphi}^{\alpha_1}$ is uniformly continuous, and, as a consequence, we obtain the following result.

Lemma 7. *Under Assumption 13, the dynamics $\hat{\varphi}^{\alpha_1}$ allows a unique uniformly continuous extension φ^{α_1} to the closure of $\wp(\mathbb{R}^{n_\theta}) \times \mathbb{R}^{mn_\theta}$.*

Proof. The result follows directly from [29, p.195, Th.26], which guarantees the existence and uniqueness of a uniformly continuous extension of a map defined on a subset of a uniform space to the closure of its domain, provided that the codomain is a complete Hausdorff uniform space. In our setting, the domain is a subspace of a metric space, and thus a uniform space, whereas the target space $\Gamma_s(\mathcal{X},\mathcal{U})$ is a compact metric space, which is uniform, complete and Hausdorff. □

We also make two conjectures on the regularity of the extended dynamics.

Conjecture 1. *The extended dynamics φ^{α_1} is uniformly Lipschitz continuous in each argument.*

The existence of Lipschitz continuous extension should be inferred from the Kirszbraun theorem (see, e.g., [48]), which assures the existence of a Lipschitz continuous extension of a Lipschitz continuous map defined on a subset of a Hilbert space to the entire space, with the same Lipschitz constant. In the present setting, one may proceed by embedding $\wp(\mathbb{R}^{n_\theta}) \subset \Gamma_s(\mathcal{X},\mathcal{U})$ into $\mathbb{R}^{n\times m}$ apply the theorem, and then restrict the extension to the closure of $\wp(\mathbb{R}^{n_\theta})$. However, we leave a rigorous verification of this construction as a subject for future work.

Remark 2. *Note that by* (3.8)*, the gradient $\nabla_\theta J^w_\wp$ can be expressed as:*

$$\nabla_\theta J^w_\wp\big|_{\bar{\theta}} = \left[\bar{G} \circ \wp\right](\theta), \quad \forall \theta \in \mathbb{R}^{n_\theta}, \tag{3.37}$$

for some $\bar{G} : \Gamma_s(\mathcal{X},\mathcal{U}) \to \mathbb{R}^{n_\theta}$. Hence, as a function of π, it is already defined on the entire closure $\mathbb{P}$. Moreover, $\bar{G}$ is smooth, since both the summation in $p_{t,\pi}$ and the Q-function are smooth in π:

$$\sum_{t=0}^{\infty} \gamma^t p_{t,\pi}(x) = W^\top \sum_{t=0}^{\infty} \gamma^t (\Pi F)^t \hat{e}_i = W^\top (I_n - \gamma \Pi F)^{-1} \hat{e}_i$$
$$Q_{\pi,r} = (I_m - \gamma F \Pi)^{-1} R,$$

where $i = \mathcal{I}_\mathcal{X}(x)$. As a consequence it is also Lipschitz continuous on $\Gamma_s(\mathcal{X},\mathcal{U})$, since the latter is a compact subset of $\mathbb{R}^{n\times m}$. We denote the 2-norm of this gradient extension scaled by the Lipschitz constant $L_\wp$ of the policy parametrization as D. In particular, for any $\pi \in \mathbb{P}$:

$$D(\pi) = L_\wp \left\|\bar{G}(\pi)\right\|_2 = L_\wp \lim_{\tau\to\infty} \left\|\nabla_\theta J^w_\wp\big|_{\theta_\tau}\right\|_2,$$

where θ_τ is any sequence such that $\wp(\theta_\tau) \to \pi$ as $\tau \to \infty$.

Conjecture 2. *The set $\mathbb{P} \setminus \wp(\mathbb{R}^{n_\theta})$ is invariant for the extended dynamics.*

Analysis of the Slow Subsystem

We now examine the behavior of the slow subsystem (3.36), by setting $\omega_k = \bar{\omega}(\pi_k)$ and $z_{j,k} = a\left(u \middle| \pi_k, \bar{\omega}(\pi_k)\right)$ for all $u \in \mathcal{U}$ and $k \geq 0$. In particular, the considered dynamics reduces to:

$$\pi_{k+1} = \bar{\varphi}^{\alpha_1}(\pi_k), \tag{3.38}$$

where:

$$\bar{\varphi}^{\alpha_1}(\pi_k) \coloneqq \varphi^{\alpha_1}(\pi_k, z)\big|_{z=a(\pi_k, \bar{\omega}(\pi_k))} \tag{3.39}$$

with z being the stack of z_j. Assuming Conjecture 1 holds, we finally obtain the following result.

Lemma 8. *The dynamics* (3.36) *and its reduced system* (3.38) *satisfy Assumption 3 on* $\mathbb{P} \coloneqq \kappa(\wp(\mathbb{R}^{n_\theta}))$.

Proof. Let us define D as in Remark 2 and $\mathcal{V}^{\mathbb{P}} : \mathbb{P} \to \mathbb{R}$ as

$$\mathcal{V}^{\mathbb{P}} \coloneqq -J^w|_{\mathbb{P}}.$$

If $\pi = \wp(\theta)$ for some $\theta \in \mathbb{R}^{n_\theta}$, then for every $\alpha_1 \in (0, 1/L_V)$ it holds:

$$\begin{aligned}
\mathcal{V}^{\mathbb{P}}(\bar{\varphi}^{\alpha_1}(\pi)) - \mathcal{V}^{\mathbb{P}}(\pi) &= -J^w_\wp\left(\theta - \alpha_1 \nabla_\theta(-J^w_\wp)\big|_\theta\right) - (-J^w_\wp(\theta)) \\
&\overset{(a)}{\leq} -\alpha_1 \left\|\nabla_\theta J^w_\wp\big|_\theta\right\|_2^2 + \frac{L_V}{2}\alpha_1^2 \left\|\nabla_\theta J^w_\wp\big|_\theta\right\|_2^2 \\
&= -\alpha_1\left(1 - \alpha_1\frac{L_V}{2}\right)\left\|\nabla_\theta J^w_\wp\big|_\theta\right\|_2^2 \\
&\leq -\frac{\alpha_1}{2}\left\|\nabla_\theta J^w_\wp\big|_\theta\right\|_2^2
\end{aligned}$$

where in (a) we leveraged Assumption 10 to apply the descent lemma (see, e.g. [49, Proposition A.24]). Furthermore, for each accumulation point $\hat{\pi} \in \mathbb{P} \setminus \wp(\mathbb{R}^{n_\theta})$, there exists a sequence of policies $\{\pi_\tau = \wp(\theta_\tau)\}_{\tau \in \mathbb{N}}$, such that $\pi_\tau \to \hat{\pi}$ as $\tau \to \infty$. Now, since for every τ the following inequality holds

$$\mathcal{V}^{\mathbb{P}}(\bar{\varphi}^{\alpha_1}(\pi_\tau)) - \mathcal{V}^{\mathbb{P}}(\pi_\tau) \leq -\frac{\alpha_1}{2}\left\|\nabla_\theta J^w_\wp\big|_{\theta_\tau}\right\|_2^2,$$

then, it still holds in the limit:

$$\begin{aligned}
\mathcal{V}^{\mathbb{P}}(\bar{\varphi}^{\alpha_1}(\hat{\pi})) - \mathcal{V}^{\mathbb{P}}(\hat{\pi}) &\leq -\frac{\alpha_1}{2}\lim_{\tau\to\infty}\left\|\nabla_\theta J^w_\wp\big|_{\theta_\tau}\right\|_2^2 \\
&= -\frac{\alpha_1}{2}\lim_{\tau\to\infty}\left\|\bar{G}(\wp(\theta_\tau))\right\|_2^2 \\
&\overset{(a)}{=} -\frac{\alpha_1}{2}\left\|\bar{G}(\hat{\pi})\right\|_2^2,
\end{aligned}$$

where (a) follows from the continuity of $\bar{G}$. Hence, recalling Remark 2, we can write:

$$\mathcal{V}^{\mathbb{P}}(\bar{\varphi}^{\alpha_1}(\pi)) - \mathcal{V}^{\mathbb{P}}(\pi) \leq -\frac{\alpha_1}{2L^2_\wp}D(\pi)^2, \quad \forall \pi \in \mathbb{P},$$

which is precisely (2.17a). Furthermore, observe that, if $\pi_\tau = \wp(\theta_\tau) \to \pi \in \mathbb{P}$ as $\tau \to \infty$, by $L_\wp$-Lipschitz continuity of $\wp$ (Assumption 8):

$$\begin{aligned}\lim_{\tau\to\infty} \|\bar{\varphi}^{\alpha_1}(\pi_\tau) - \pi_\tau\| &= \lim_{\tau\to\infty} \left\|\wp\left(\theta_\tau + \alpha_1 \nabla_\theta J_\wp^w\big|_{\theta_\tau}\right) - \wp(\theta_\tau)\right\| \\ &\le \lim_{\tau\to\infty} \alpha_1 L_\wp \left\|\nabla_\theta J_\wp^w\big|_{\theta_\tau}\right\| \\ &= \alpha_1 D(\pi),\end{aligned}$$

and thus (2.16) is satisfied. Finally, if $v \coloneqq (\omega, z, y)$, we have for all $\theta \in \mathbb{R}^{n_\theta}$:

$$\begin{aligned}\mathcal{V}^{\mathbb{P}}(\varphi^{\alpha_1}(\pi_\theta, z_1)) - \mathcal{V}^{\mathbb{P}}(\varphi^{\alpha_1}(\pi_\theta, z_2)) &\overset{(a)}{\le} d\mathcal{V}^{\mathbb{P}}|_{\pi_\theta}(\varphi^{\alpha_1}(\pi_\theta, z_1) - \varphi^{\alpha_1}(\pi_\theta, z_2)) \\ &\quad + \frac{L_V}{2} \underbrace{\|\varphi^{\alpha_1}(\pi_\theta, z_1) - \pi_\theta\|^2}_{=:\alpha_1^2\Delta^\alpha(\pi_\theta, z_1)^2} + \frac{L_V}{2} \underbrace{\|\varphi^{\alpha_1}(\pi_\theta, z_2) - \pi_\theta\|^2}_{=:\alpha_1^2\Delta^\alpha(\pi_\theta, z_2)^2} \\ &\overset{(b)}{\le} c\alpha_1 \|\nabla_\theta J_\wp^w|_\theta\| \cdot \left\|\sum_{i=1}^m \left(z_1^i - z_2^i\right)\right\| + \alpha_1^2 \frac{L_V}{2}\left(\Delta^\alpha(\pi_\theta, z_1)^2 + \Delta^\alpha(\pi_\theta, z_2)^2\right) \\ &\le mc\alpha_1 \|\nabla_\theta J_\wp^w|_\theta\| \cdot \|z_1 - z_2\| + \alpha_1^2 \frac{L_V}{2}\left(\Delta^\alpha(\pi_\theta, z_1)^2 + \Delta^\alpha(\pi_\theta, z_2)^2\right) \\ &\le mc\alpha_1 \|\nabla_\theta J_\wp^w|_\theta\| \cdot \|v_1 - v_2\| + \alpha_1^2 \frac{L_V}{2}\left(\Delta^\alpha(\pi_\theta, z_1)^2 + \Delta^\alpha(\pi_\theta, z_2)^2\right)\end{aligned}$$

where (a) follows from the L_V-smoothness of $J_\wp^w$ and [49, Proposition A.24], while (b) follows from inequality (3.6) of Assumption 8. As before, since the limit operation preserves inequalities, this condition holds for accumulation points as well. Hence:

$$\mathcal{V}^{\mathbb{P}}(\varphi^{\alpha_1}(\pi, z_1)) - \mathcal{V}^{\mathbb{P}}(\varphi^{\alpha_1}(\pi, z_2)) \le \frac{mc}{L_\wp}\alpha_1 D(\pi)\|v_1 - v_2\| + \alpha_1^2 \frac{L_V}{2}\left(\Delta^\alpha(\pi, z_1)^2 + \Delta^\alpha(\pi, z_2)^2\right)$$

for all $\pi \in \mathbb{P}$. This condition is a special case of (2.17b), where the slow dynamics are affected directly only by the z component of v. □

Convergence of the Interconnected System

We are now ready to characterize the convergence properties of Algorithm 1.

Theorem 2. *Consider Algorithm 1 and let Assumptions 8, 9, 10, 11, 12 and 13 hold. Then, there exists $\bar{\alpha}_1, \bar{\alpha}_2 > 0$ such that, for all $\alpha_1 \in (0, \bar{\alpha}_1)$, $\alpha_2 \in (0, \bar{\alpha}_2)$, $\theta_0 \in \mathbb{R}^{n_\theta}$, $\omega_0 \in \mathbb{R}^{n_\omega}$, $z_0^u \in \mathbb{R}^{n_\theta}$, and $y_0^u \in \mathbb{R}^{n_\omega}$ for all $u \in \mathcal{U}$, it holds*

$$\lim_{k\to\infty} \min_{\pi\in\mathcal{S}} \|\wp(\theta_k) - \pi\| = 0,$$

where

$$\mathcal{S} \coloneqq \{\pi \in \mathbb{P} : D(\pi) = 0\} \subseteq \{\pi \in \mathbb{P} : \bar{\varphi}^{\alpha_1}(\pi) = \pi\} =: \Theta^{\alpha_1}.$$

Proof. Lemmas 5-6-8, together with the Lipschitz continuity of the dynamics, allows us to directly apply Theorem 1. □

Stochastic Sampled-Average Actor-Critic

In this section we address the scenario in which the measurement unit does not have access to the MDP model. Specifically, following the standard practice in the actor-critic framework, we exploit the structure of the ideal actor and critic updates to reformulate them as expectations of two random variables defined over the action space. Although these random variables are somewhat artificial, their samples can be obtained by running experiments, since their distributions strictly depend on the current policy and on the transition probabilities of the MDP.

The Ideal Actor Update as Expectation

With a few algebraic manipulations, the i-th partial derivative of the objective function can be rewritten as an expectation. Specifically, recalling (3.8), we obtain:

$$\begin{aligned}\left.\frac{\partial J_\pi^w}{\partial \theta^i}\right|_\theta &= \sum_{u\in\mathcal{U}} \left(\sum_{t=0}^{\infty} \gamma^t \big[p_{t,\pi_\theta} \circ s\big](u)\right) q_{\pi_\theta,r}(u)\big[\psi_i \circ \pi_\theta\big](u) \\ &= \sum_{u\in\mathcal{U}} \left(\frac{1}{m_{s(u)}} \sum_{t=0}^{\infty} \frac{\gamma^t}{1-\gamma}\big[p_{t,\pi_\theta} \circ s\big](u)\right) (1-\gamma) m_{s(u)} q_{\pi_\theta,r}(u)\big[\psi_i \circ \pi_\theta\big](u) \\ &= \mathbb{E}\left[(1-\gamma) m_x\, q_{\pi_\theta}(u)\,\big[\psi_i \circ \pi_\theta\big](u) \,\middle|\, \begin{array}{l} t \sim g_\gamma \\ x \sim p_{t,\pi_\theta} \\ u \sim \nu_x \end{array}\right], \end{aligned} \tag{3.40}$$

for all $i \in \{1, \dots, n_\theta\}$, where $g_\gamma : t \mapsto \gamma^t/(1-\gamma)$ is a geometric probability mass function on $\mathbb{N}$, whereas $\nu_x : u \mapsto 1/m_x$ is the uniform distribution on $\mathcal{U}_x$.

Remark 3. *In case $\pi_\theta(\cdot, x) > 0$ for all $\theta \in \mathbb{R}^{n_\theta}$ and $x \in \mathcal{X}$, we can also use the identity*

$$\left.\frac{\partial}{\partial \theta}\right|_{\bar\theta} [\wp(\theta)](u,x) = \pi_{\bar\theta}(u,x) \left.\frac{\partial}{\partial \theta}\right|_{\bar\theta} \log([\wp(\theta)](u,x))$$

to make the policy appear as a weighting probability over actions (see, e.g., [50]). In this case ν_x should be replaced by $\pi_\theta(\cdot, x)$. Both reformulations are valid for the analysis presented in the following subsections.

Leveraging Assumption 9, we can rewrite (3.40) in the more compact form:

$$\left.\frac{\partial J_\pi^w}{\partial \theta^i}\right|_\theta = \mathbb{E}\left[\hat a^i(u|\pi_\theta, \bar\omega(\pi_\theta)) \,\middle|\, u \sim p_{\hat a}(\cdot|\pi_\theta)\right], \tag{3.41}$$

where $p_{\hat a}(\cdot|\pi_\theta)$ is the probability distribution on $\mathcal{U}$ resulting from the three subsequent samplings in (3.40) and $\hat a^i(\cdot|\pi_\theta, \bar\omega(\pi_\theta)) : \mathcal{U} \to \mathbb{R}$ is defined as

$$\hat a^i(u|\pi_\theta, \bar\omega(\pi_\theta)) \coloneqq (1-\gamma) m_{s(u)}\, \hat q_{\bar\omega(\pi_\theta)}(u)\,\big[\psi^i \circ \pi_\theta\big](u).$$

Hence, provided that the action is sampled from the distribution $p_{\hat a}(\cdot|\pi_\theta)$ at each iteration $k \in \mathbb{N}$, we can use these realizations to produce measurements and update a current estimate z_k of the expectation (3.41) with a convex combination step

$$z_{k+1}^i = (1-\alpha_3) z_k^i + \alpha_3 \hat a^i(u_k|\pi_{\theta_k}, \omega_k), \quad i \in \{1, \cdots, n_\theta\},$$

where we also replaced the unavailable $\omega_{\pi_{\theta_k}}$ with an estimate ω_k. Hence, we can use z_k in place of the exact gradient to approximate a gradient method and improve the policy parameter θ_k:

$$\theta_{k+1} = \theta_k + \alpha_1 \underbrace{z_k}_{\approx \nabla_\theta J_{\wp}^{w}|_{\theta_k}} .$$

The Ideal Critic Update as Expectation

With similar algebraic manipulations, we can do the same with the ideal critic update. Specifically, recalling (3.10), we have:

$$\begin{aligned}
e(u|\omega,\pi) &= r(u) + \gamma \sum_{z\in\mathcal{X}} f(z,u) \sum_{v\in\mathcal{U}_z} \pi(v,z)\hat{q}_\omega(v) - \hat{q}_\omega(u) \\
&= \underbrace{\sum_{z\in\mathcal{X}} f(z,u) \sum_{v\in\mathcal{U}_z} \pi(v,z)}_{=1}\big(r(u) - \hat{q}_\omega(u)\big) + \gamma \sum_{z\in\mathcal{X}} f(z,u) \sum_{v\in\mathcal{U}_z} \pi(v,z)\hat{q}_\omega(v) \\
&= \sum_{z\in\mathcal{X}} f(z,u) \sum_{v\in\mathcal{U}_z} \pi(v,z)\big(r(u) + \gamma\hat{q}_\omega(v) - \hat{q}_\omega(u)\big) \\
&= \mathbb{E}\left[r(u) + \gamma\hat{q}_\omega(v) - \hat{q}_\omega(u) \,\middle|\, \begin{array}{l} z \sim f(\cdot,u) \\ v \sim \pi(\cdot,z) \end{array} \right].
\end{aligned}$$

Therefore, we obtain:

$$\begin{aligned}
\sum_{u\in\mathcal{U}} e(u|\omega,\pi)\phi^i(u) &= \sum_{u\in\mathcal{U}} \frac{1}{m}\big(m\, e(u|\omega,\pi)\phi^i(u)\big) \\
&= \mathbb{E}\left[m\, e(u|\omega,\pi)\phi^i(u) \middle| u \sim \nu \right] \\
&= \mathbb{E}\left[m\big[r(u) + \gamma\hat{q}_\omega(v) - \hat{q}_\omega(u)\big]\phi^i(u) \,\middle|\, \begin{array}{l} u \sim \nu \\ z \sim f(\cdot,u) \\ v \sim \pi(\cdot,z) \end{array} \right] \\
&= \mathbb{E}\left[\hat{c}^i(\sigma^y|\omega) \middle|\ \sigma^y \sim p_c(\cdot|\omega,\pi) \right], \qquad (3.42)
\end{aligned}$$

where ν is the uniform distribution on $\mathcal{U}$, $\sigma^y = (u,v)$, and

$$\hat{c}^i\big((s_1,s_2)|\omega\big) = m\big[r(s_1) + \gamma\hat{q}_\omega(s_2) - \hat{q}_\omega(s_1)\big]\phi^i(s_1).$$

Hence, as done in the previous subsection, at each iteration k we can sample a pair of actions σ_k^y from the distribution $p_c(\cdot|\omega_k,\pi_{\theta_k})$ and update the estimate of the expectation (3.42) through

$$y_{k+1}^i = (1-\alpha_3)y_k^i + \alpha_3\hat{c}^i(\sigma_k^y|\omega_k), \quad \forall i \in \{1,\cdots,n_\omega\}.$$

Hence, we can replace the ideal critic update (3.11) with

$$\omega_{k+1} = \omega_k + \alpha_2 y_k, \quad \forall i \in \{1,\cdots,n_\omega\},$$

where $y = (y^1,\cdots,y^{n_\omega})$.

The Algorithm

By combining the considerations done in the previous two subsections, we obtain Algorithm 2. Compared with Algorithm 1, this new version is considerably less computationally demanding, as the number of auxiliary variables has been reduced from $2m$ to only 2. However, this reduction in complexity comes at the cost of introducing stochasticity into the updates, making the subsequent analysis more involved. To address this, it is convenient to interpret the resulting interconnected dynamics as a stochastic perturbation of a nominal system, obtained by averaging the dynamics at each iteration. In this way, the system can be viewed as the averaged one subject to the zero-mean stochastic inputs defined as:

$$\mathrm{w}_k^z \coloneqq \hat{a}(\pi_{\theta_k}, \omega_k) - G(\wp(\theta_k), \omega_k), \tag{3.43}$$
$$\mathrm{w}_k^y \coloneqq \hat{c}(\omega_k, \pi_{\theta_k}) - B(\omega_k, \wp(\theta_k)), \tag{3.44}$$

where we introduced the shorthands:

$$G(\pi, \omega) \coloneqq \mathbb{E}\left[\hat{a}(u|\pi, \omega)|\ u \sim p_{\hat{a}}(\cdot|\pi)\right]$$
$$B(\omega, \pi) \coloneqq \mathbb{E}\left[\hat{c}(\sigma^y|\omega)|\ \sigma^y \sim p_c(\cdot|\omega, \pi)\right].$$

To formalize this idea, we rewrite Algorithm 2 in system-theoretic form:

$$\theta_{k+1} = \theta_k + \alpha_1 z_k \tag{3.45a}$$
$$\omega_{k+1} = \omega_k + \alpha_2 y_k \tag{3.45b}$$
$$z_{k+1} = (1-\alpha_3) z_k + \alpha_3 G(\wp(\theta_k), \omega_k) + \alpha_3 \mathrm{w}_k^z \tag{3.45c}$$
$$y_{k+1} = (1-\alpha_3) y_k + \alpha_3 B(\omega_k, \wp(\theta_k)) + \alpha_3 \mathrm{w}_k^y, \tag{3.45d}$$

from which the corresponding averaged system is immediately seen to be

$$\theta_{k+1}^{\mathrm{avg}} = \theta_k^{\mathrm{avg}} + \alpha_1 z_k^{\mathrm{avg}} \tag{3.46a}$$
$$\omega_{k+1}^{\mathrm{avg}} = \omega_k^{\mathrm{avg}} + \alpha_2 y_k^{\mathrm{avg}} \tag{3.46b}$$
$$z_{k+1}^{\mathrm{avg}} = (1-\alpha_3) z_k^{\mathrm{avg}} + \alpha_3 G(\wp(\theta_k^{\mathrm{avg}}), \omega_k^{\mathrm{avg}}) \tag{3.46c}$$
$$y_{k+1}^{\mathrm{avg}} = (1-\alpha_3) y_k^{\mathrm{avg}} + \alpha_3 B(\omega_k^{\mathrm{avg}}, \wp(\theta_k^{\mathrm{avg}})), \tag{3.46d}$$

which will be analyzed in the following subsection.

Analysis of the Averaged System

First, observe that system (3.46) is fully deterministic and naturally extends to the closure of the parametrization image. Moreover, due to the presence of α_1 and α_2, it can be analyzed using timescale separation arguments, as done for (3.27). Furthermore, the auxiliary systems associated with the slow dynamics (3.46a) and of the mid-level dynamics (3.46b) correspond to (3.38) and (3.33), respectively. Therefore, Lemmas 6-8 also hold in this framework. As a result, the only difference with respect to the system analyzed in the previous section lies in the measurement update, whose auxiliary system for arbitrary π and ω reads as

$$z_{k+1}^{\mathrm{avg}} = (1-\alpha_3) z_k^{\mathrm{avg}} + \alpha_3 G(\pi, \omega) \tag{3.48a}$$
$$y_{k+1}^{\mathrm{avg}} = (1-\alpha_3) y_k^{\mathrm{avg}} + \alpha_3 B(\omega, \pi), \tag{3.48b}$$

for which we state the following lemma.

Algorithm 2 Sampled-Average Actor-Critic

for $k = 0, 1, 2 \dots$ **do**

Policy Improvement Step

$$\theta_{k+1} = \theta_k + \alpha_1 z_k \tag{3.47a}$$

Policy Evaluation Step

$$\omega_{k+1} = \omega_k + \alpha_2 y_k \tag{3.47b}$$

Auxiliary Variables Update

1) Sample

$$u_k \sim p_a(\cdot|\pi_{\theta_k}) \tag{3.47c}$$
$$\sigma_k^y \sim p_c(\cdot|\omega_k, \pi_{\theta_k}) \tag{3.47d}$$

2) Update

$$z_{k+1} = (1-\alpha_3) z_k + \alpha_3 \hat{a}\,(u_k|\pi_{\theta_k}, \omega_k) \tag{3.47e}$$
$$y_{k+1} = (1-\alpha_3) y_k + \alpha_3 \hat{c}\,(\sigma_k^y|\omega_k) \tag{3.47f}$$

Lemma 9. *The boundary-layer system* (3.48) *satisfies Assumption 7 for* $\alpha_3 \in (0, 2)$.

Proof. Consider the error coordinates

$$\tilde{z}^{\text{avg}} \coloneqq z^{\text{avg}} - G(\pi, \omega) \tag{3.49}$$
$$\tilde{y}^{\text{avg}} \coloneqq y^{\text{avg}} - B(\omega, \pi). \tag{3.50}$$

In these chart, system (3.48a) reads as

$$\tilde{z}_{k+1}^{\text{avg}} = (1-\alpha_3)\tilde{z}_k^{\text{avg}}$$
$$\tilde{y}_{k+1}^{\text{avg}} = (1-\alpha_3)\tilde{y}_k^{\text{avg}},$$

which is an autonomous linear time-invariant system with state matrix $(1-\alpha_3) I_{n_\zeta}$, $n_\zeta = n_\theta + n_\omega$. Now, for $\alpha_3 \in (0, 2)$, let $\tilde{\zeta}^{\text{avg}} \coloneqq (\tilde{z}^{\text{avg}}, \tilde{y}^{\text{avg}})$ and define $\tilde{\mathcal{V}}^Z : \mathbb{R}^{n_\zeta} \to \mathbb{R}$ as

$$\tilde{\mathcal{V}}^Z\left(\tilde{\zeta}^{\text{avg}}\right) \coloneqq \frac{1}{\alpha_3(2-\alpha_3)} \left\|\tilde{\zeta}^{\text{avg}}\right\|^2. \tag{3.51}$$

This function trivially satisfies (2.24a) and (2.24c). Furthermore, observe that:

$$\bar{P} = \frac{1}{\alpha_3(2-\alpha_3)} I_{n_\zeta},$$

is the unique solution to the Lyapunov equation:

$$A^\top P A - P = -I_{n_\zeta}$$

with $A = (1-\alpha_3) I_{n_\zeta}$. It follows that, along the trajectories of (3.48), it holds:

$$\tilde{\mathcal{V}}^Z\left(\tilde{\zeta}_{k+1}^{\text{avg}}\right) - \tilde{\mathcal{V}}^Z\left(\tilde{\zeta}_k^{\text{avg}}\right) = \left\langle \tilde{\zeta}_k^{\text{avg}}, \left[(2-\alpha_3)^2 \bar{P} - \bar{P}\right] \tilde{\zeta}_k^{\text{avg}} \right\rangle \leq -\left\|\tilde{\zeta}_k^{\text{avg}}\right\|_2^2.$$

Hence, also (2.24b) is satisfied. □

Finally, Lemmas 6-8-9 enable us to give the following convergence result on the avreaged system (3.46).

Theorem 3. *Consider system* (3.46) *and let Assumptions 8, 9, 10, 11, 12 and 13 hold. Then, there exists* $\bar{\alpha}_1, \bar{\alpha}_2 > 0$ *such that, for all* $\alpha_1 \in (0, \bar{\alpha}_1)$, $\alpha_2 \in (0, \bar{\alpha}_2)$, $\alpha_3 \in (0, 2)$, $\theta_0^{avg} \in \mathbb{R}^{n_\theta}$, $\omega_0^{avg} \in \mathbb{R}^{n_\omega}$, $z_0^{avg} \in \mathbb{R}^{n_\theta}$, *and* $y_0^{avg} \in \mathbb{R}^{n_\omega}$, *it holds*

$$\lim_{k\to\infty} \min_{\pi\in\mathcal{S}} \|\wp(\theta_k^{avg}) - \pi\| = 0,$$

where

$$\mathcal{S} \coloneqq \{\pi \in \mathbb{P} : D(\pi) = 0\} \subseteq \{\pi \in \mathbb{P} : \bar{\varphi}^{\alpha_1}(\pi) = \pi\} =: \Theta^{\alpha_1}.$$

Theorem 3 follows directly from Theorem 1, which relies on the existence of a Lyapunov-like function, obtained as the sum of the individual Lyapunov-like functions corresponding to the three auxiliary subsystems. In this specific context, the resulting function, expressed in error coordinates, is given by

$$\tilde{\mathcal{V}}\left(\pi, \tilde{\omega}, \tilde{\zeta}\right) \coloneqq -J(\pi) + \frac{1}{2}\|\tilde{\omega}\|^2 + \frac{1}{\alpha_3(2-\alpha_3)}\left\|\tilde{\zeta}\right\|^2. \tag{3.52}$$

and, for α_1 and α_2 sufficiently small, it satisfies

$$\tilde{\mathcal{V}}\left(\pi_{k+1}^{\text{avg}}, \tilde{\omega}_{k+1}^{\text{avg}}, \tilde{\zeta}_{k+1}^{\text{avg}}\right) - \tilde{\mathcal{V}}\left(\pi_k^{\text{avg}}, \tilde{\omega}_k^{\text{avg}}, \tilde{\zeta}_k^{\text{avg}}\right) \leq -q\left(D(\pi_k^{\text{avg}}), \|\tilde{\omega}_k^{\text{avg}}\|, \left\|\tilde{\zeta}_k^{\text{avg}}\right\|\right) \tag{3.53}$$

along the trajectories of (3.46), for some positive definite quadratic function $q : \mathbb{R}^3 \to \mathbb{R}$. This Lyapunov-like function plays a central role in the analysis of the perturbed system (3.45), which is carried out in the following subsection.

Analysis of the Perturbed System

Let us consider system (3.45) in error coordinates and with the actor dynamics transferred in the policy space:

$$\pi_{k+1} = \varphi^{\alpha_1}\left(\pi_k, \tilde{z}_k + G\big(\pi_k, \tilde{\omega}_k + \bar{\omega}(\pi_k)\big)\right) \tag{3.54a}$$

$$\tilde{\omega}_{k+1} = \tilde{\omega}_k + \alpha_2\left(\tilde{y}_k + B\big(\tilde{\omega}_k + \bar{\omega}(\pi_k), \pi_k\big)\right) + \bar{\omega}(\pi_k) - \bar{\omega}(\pi_{k+1}) \tag{3.54b}$$

$$\tilde{z}_{k+1} = (1-\alpha_3)\tilde{z}_k + G(\pi_k, \tilde{\omega}_k + \bar{\omega}(\pi_k)) - G(\pi_{k+1}, \tilde{\omega}_{k+1} + \bar{\omega}(\pi_{k+1})) + \alpha_3 \mathrm{w}_k^z \tag{3.54c}$$

$$\tilde{y}_{k+1} = (1-\alpha_3)\tilde{y}_k + B(\tilde{\omega}_k + \bar{\omega}(\pi_k), \pi_k) - B(\tilde{\omega}_{k+1} + \bar{\omega}(\pi_{k+1}), \pi_{k+1}) + \alpha_3 \mathrm{w}_k^y, \tag{3.54d}$$

where the additional terms account for the drift of the parametrized equilibria. It is evident that the zero-mean disturbances, even in this representation, are additive and affect only the fast system. Moreover, apart from the disturbances, the remaining the dynamics correspond to the averaged dynamics in this representation, as the error coordinates are unaffected by additive disturbances. Therefore, we may compactly rewrite the dynamics representation (3.54) as:

$$\pi_{k+1} = \tilde{\varphi}_{\text{avg}}^{\alpha_1}\big(\pi_k, \big(\tilde{\omega}_k, \tilde{\zeta}_k\big)\big) \tag{3.55a}$$

$$\tilde{\omega}_{k+1} = \tilde{g}_{\text{avg}}^{\alpha_2}\big(\pi_k, \big(\tilde{\omega}_k, \tilde{\zeta}_k\big)\big) \tag{3.55b}$$

$$\tilde{\zeta}_{k+1} = \tilde{h}_{\text{avg}}\big(\pi_k, \big(\tilde{\omega}_k, \tilde{\zeta}_k\big)\big) + \alpha_3 \mathrm{w}_k \tag{3.55c}$$

where w_k denotes the stack of w_k^z and w_k^y. We now characterize the behavior of the stochastic system representation (3.55) with the following proposition.

Proposition 1. *Let $\tilde{\mathcal{V}}$ be the function defined in* (3.52)*, $\alpha_3 \in (0,2)$ and let α_1, α_2 chosen according to Theorem 3, so as to make the averaged system satisfy* (3.53)*. Then, along the trajectories of* (3.55)*, it holds:*

$$\mathbb{E}\left[\tilde{\mathcal{V}}\left(\pi_{k+1}, \tilde{\omega}_{k+1}, \tilde{\zeta}_{k+1}\right)\right] - \tilde{\mathcal{V}}\left(\pi_k, \tilde{\omega}_k, \tilde{\zeta}_k\right) \leq -q\left(D(\pi_k), \|\tilde{\omega}_k\|, \left\|\tilde{\zeta}_k\right\|\right) + \frac{\alpha_3}{2-\alpha_3}\sigma_w^2(\pi_k, \tilde{\omega}_k), \tag{3.56}$$

where $\sigma_w^2(\pi_k, \tilde{\omega}_k)$ is the variance of the zero-mean input w *at iteration k.*

Proof. Let $\left(\pi_{k+1}^{\text{avg}}, \tilde{\omega}_{k+1}^{\text{avg}}, \tilde{\zeta}_{k+1}^{\text{avg}}\right)$ denote the state update that would result from $\left(\pi_k, \tilde{\omega}_k, \tilde{\zeta}_k\right)$ by applying the averaged dynamics obtained from (3.55) by removing the input w. Then, introducing the shorthands

$$\begin{aligned}
\Delta G_k &\coloneqq G(\pi_k, \tilde{\omega}_k + \bar{\omega}(\pi_k)) - G(\pi_{k+1}, \tilde{\omega}_{k+1} + \bar{\omega}(\pi_{k+1})) \\
\Delta B_k &\coloneqq B(\tilde{\omega}_k + \bar{\omega}(\pi_k), \pi_k,) - B(\tilde{\omega}_{k+1} + \bar{\omega}(\pi_{k+1}), \pi_{k+1})
\end{aligned}$$

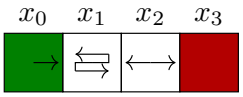
we obtain:

$$\begin{aligned}
\mathbb{E}\left[\tilde{\mathcal{V}}\left(\pi_{k+1}, \tilde{\omega}_{k+1}, \tilde{\zeta}_{k+1}\right)\right] - \tilde{\mathcal{V}}\left(\pi_k, \tilde{\omega}_k, \tilde{\zeta}_k\right) &= \mathbb{E}\left[\tilde{\mathcal{V}}\left(\pi_{k+1}, \tilde{\omega}_{k+1}, \tilde{\zeta}_{k+1}\right)\right] - \tilde{\mathcal{V}}\left(\pi_k, \tilde{\omega}_k, \tilde{\zeta}_k\right) \pm \tilde{\mathcal{V}}\left(\pi_{k+1}^{\text{avg}}, \tilde{\omega}_{k+1}^{\text{avg}}, \tilde{\zeta}_{k+1}^{\text{avg}}\right) \\
&= \mathbb{E}\left[\tilde{\mathcal{V}}\left(\pi_{k+1}, \tilde{\omega}_{k+1}, \tilde{\zeta}_{k+1}\right) - \tilde{\mathcal{V}}\left(\pi_{k+1}^{\text{avg}}, \tilde{\omega}_{k+1}^{\text{avg}}, \tilde{\zeta}_{k+1}^{\text{avg}}\right)\right] \\
&\quad \tilde{\mathcal{V}}\left(\pi_{k+1}^{\text{avg}}, \tilde{\omega}_{k+1}^{\text{avg}}, \tilde{\zeta}_{k+1}^{\text{avg}}\right) - \tilde{\mathcal{V}}\left(\pi_k, \tilde{\omega}_k, \tilde{\zeta}_k\right) \\
&\overset{\text{(a)}}{\leq} \mathbb{E}\left[\tilde{\mathcal{V}}\left(\pi_{k+1}, \tilde{\omega}_{k+1}, \tilde{\zeta}_{k+1}\right) - \tilde{\mathcal{V}}\left(\pi_{k+1}^{\text{avg}}, \tilde{\omega}_{k+1}^{\text{avg}}, \tilde{\zeta}_{k+1}^{\text{avg}}\right)\right] \\
&\quad - q\left(D(\pi_k), \|\tilde{\omega}_k\|, \left\|\tilde{\zeta}_k\right\|\right) \\
&\overset{\text{(b)}}{\leq} \frac{2}{\alpha_3(2-\alpha_3)}\mathbb{E}\left[\left\langle\left(\tilde{z}_{j,k} + \Delta G_k\right), \mathrm{w}_k^z\right\rangle + \left\langle\left(\tilde{y}_{j,k} + \Delta B_k\right), \mathrm{w}_k^y\right\rangle\right] \\
&\quad + \frac{\alpha_3^2}{\alpha_3(2-\alpha_3)}\mathbb{E}\left[\|\mathrm{w}_k\|^2\right] - q\left(D(\pi_k), \|\tilde{\omega}_k\|, \left\|\tilde{\zeta}_k\right\|\right) \\
&\overset{\text{(c)}}{=} \frac{\alpha_3}{(2-\alpha_3)} \underbrace{\mathbb{E}\left[\|\mathrm{w}_k\|^2\right]}_{=:\sigma_w^2(\pi_k, \tilde{\omega}_k)} - q\left(D(\pi_k), \|\tilde{\omega}_k\|, \left\|\tilde{\zeta}_k\right\|\right)
\end{aligned}$$

where (a) follows from (3.53), (b) follows from the decoupled structure of $\tilde{\mathcal{V}}$ as given in (3.52), and from the fact that the perturbed dynamics for the mid and slow systems are identical to those of the averaged system, while (c) holds since both w^z and w^y are zero-mean. □

Proposition 1 shows that, on average, the Lyapunov-like function $\tilde{\mathcal{V}}$ decreases as long as the polynomial q dominates the variance of the disturbance w. However, this bound is not uniform, since the variance depends on both π_k and $\tilde{\omega}_k$. The most we can do is to take its maximum value over the policy space, which is compact, but the dependence on $\tilde{\omega}_k$ cannot be eliminated. Nevertheless, the variance appear in (3.56) multiplied by the step size α_3, and its effect can therefore be mitigated by appropriately tuning this parameter, at the cost of slower convergence. We thus conjecture that, if the system evolves sufficiently slowly, the variance remains bounded and the system exhibits practical stability with respect to the set of equilibria of the averaged system.

Numerical Simulations



Let us take as a benchmark for the simulations the *short corridor with switched actions*, as described in [14, Example 13.1]. This MDP consists of a one-dimensional grid world with 4 states

$$\mathcal{X} = \{x_0, x_1, x_2, x_3\},$$

and an action pool of two available actions

$$\tilde{\mathcal{U}} = \{u_l, u_r\},$$

corresponding to an attempt to move left (u_l) or right (u_r). The resulting action space thus contains 8 actions

$$\mathcal{U} = \{u_l^i, u_r^i : i = 0, \cdots, 3\}.$$

The agent is initialized in state x_0 and aims to reach the absorbing target state x_3 as quickly as possible. Accordingly, the initial distribution is represented by the probability vector $W = (1, 0, 0, 0)$, while a suitable reward function to express this goal is $R = (-1, -1, -1, -1, -1, -1, 0, 0)$. A key feature of this environment is that the dynamics in state x_1 are reversed: if the agent attempts to move left in x_1, it will instead move right, and vice versa. Furthermore the agent cannot fall off the grid, so that if it tries to move left in state x_0, nothing will happen. The overall transition matrix is then the following:

$$F = \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{array}\right). \tag{3.57}$$

To reduce complexity, we parametrize the policy with only two parameters θ^1, θ^2 through a softmax function:

$$\pi_\theta(\cdot, x) = \sigma(\theta) = \frac{1}{e^{\theta^1} + e^{\theta^2}} \left(e^{\theta^1}, e^{\theta^2}\right).$$

In this way the admissible policies are state-independent and no deterministic policy can be optimal. Furthermore, for the discussion previously done on the softmax function, the two parameters are not independent. Specifically, if we remove their average, they are related by $\theta^1 = -p^2$. This defines a line of parameters containing the origin, which is isomorphic to

the quotient $\mathbb{R}^2/\text{span}\{\mathbb{1}_2\}$. Furthermore, we assume a full parametrization for the Q-function, meaning that the span of the basis functions is $\mathbb{R}^8$, to guarantee Assumption 9. We then tested both algorithms on this benchmark to evaluate their convergence behavior for suitably chosen step sizes. Figure 3.2 shows the evolution of the gradient norm and of the policy parameter θ^2 (the other parameter is equal to this with opposite sign) for Algorithm 1, while Figure 3.3 presents the corresponding results for Algorithm 2. In the latter case, observe that, after the transient, the parameter θ^2 remains within a neighborhood of its optimal value. Furthermore, its convergence is considerably slower. This is due to the fact that in order to mitigate the effect of the input variance, the stepsize α_3 had to be reduced, thus slowing down the overall system dynamics.

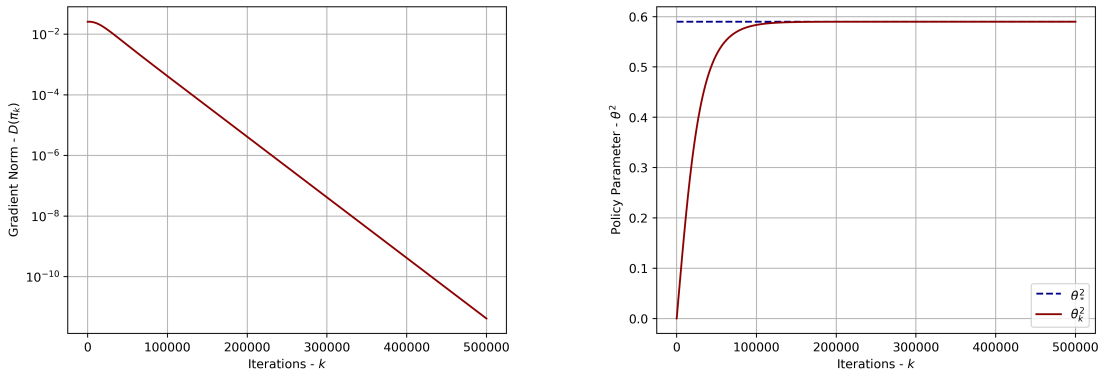


Figure 3.2: Convergence behavior of Algorithm 1 on the presented benchmark. The left plot shows the evolution of the gradient norm throughout the simulation. The right plot shows the evolution of the parameter θ^2 (red line) compared to its optimal value (dashed blue line).

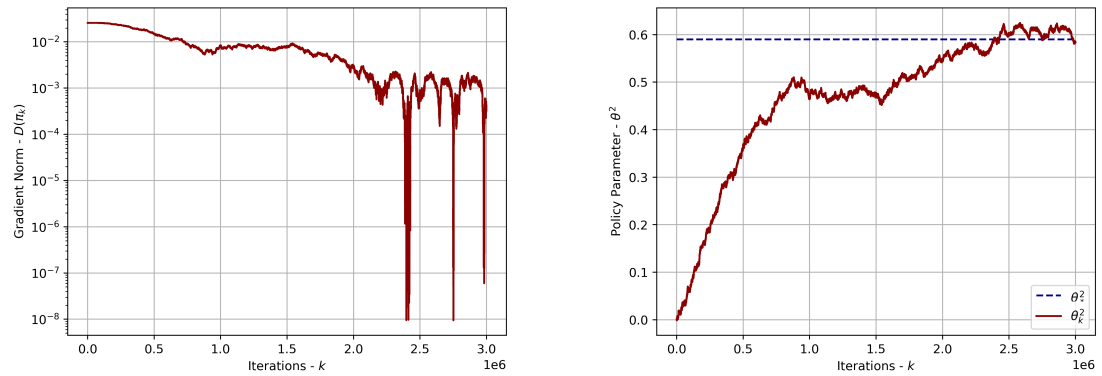


Figure 3.3: Convergence behavior of Algorithm 2 on the presented benchmark. The left plot shows the evolution of the gradient norm throughout the simulation. The right plot shows the evolution of the parameter θ^2 (red line) compared to its optimal value (dashed blue line).

Chapter 4

Reward-Balancing Methods for Reinforcement Learning

In this chapter we thoroughly explore the mathematical framework underlying novel approaches to discounted-return reinforcement learning, referred to as *reward-balancing methods*. We present a detailed analysis of the reward transformations involved, examined through a group-theoretic perspective, and we highlight their key properties both in the exact-model and approximate-model setting. Furthermore, we reinterpret the problem they aim to solve as a control problem, analyzing existing methods under a system-theoretic perspective and introducing an optimal control framework that serves as a principled design tool for the design of reward-balancing methods with additional constraints for, e.g., addressing model uncertainties or a-priori known symmetries.
The present work will appear soon as a preprint on arXiv [51].

Literature Review

The class of reward-balancing methods treated in this chapter were first introduced in the recent work of Mustafin and collaborators [12]. These methods are particularly interesting because they can be viewed as a sort of dual approach to standard reinforcement learning methods such as actor-critic and Q-learning, while still matching the state-of-the-art sample complexity of Q-learning. Moreover, the version proposed in [12, Section 8] for the unknown-model setting is fully parallelizable, representing a significant improvement over the current state-of-the-art [52].
The idea of modifying rewards without altering the set of optimal policies was first studied by Ng et al. [53], where potential functions were used to define policy-invariant reward transformations. This method, commonly referred to as *reward shaping*, has then become a foundational tool for improving sample efficiency and convergence speed in reinforcement learning. Subsequent works have extended these ideas to more sophisticated shaping functions and learning-based shaping mechanisms, including techniques for automatic reward-balancing [54, 55] and methods that learn shaping potentials [56, 57].
Throughout the analysis, we will make use of fundamental concepts from group theory and optimal control. Readers seeking an introduction on these topics may refer to [58] for group theory and [59] for discrete-time optimal control.

Exact-Model Reward-Balancing Methods

Advantage-Preserving Reward Transformations

In [12] the authors propose a method for modifying the reward function of an MDP in order to transform it into a *normal MDP*, whose definition we recall here for completeness.

Definition 4. *We say that an MDP is in normal form iff its optimal value function is zero.*

Relation (1.9) implies that, in a normal MDP, every state has (at least) one action with zero reward in its fiber, which lies in the image of an optimal policy. Thus, the key advantage of these MDPs is that finding an optimal policy becomes trivial: it is sufficient to select, for each state, an action with zero reward. Now, the challenging aspect is not how to "normalize" a given MDP, but how to do it without altering the problem it is intended to model. In particular, the sought transformation should preserve the relative position of the value functions, which induce a partial order over the policy space. The set of reward transformations satisfying this condition is isomorphic to a subgroup of the affine group of $\mathbb{R}^m$, made by all the transformations of the form

$$\ell_\delta : r \mapsto r - \Gamma(\delta) \tag{4.1}$$

parametrized by the real-valued functions $\delta : \mathcal{X} \to \mathbb{R}$ on the state space through the injective linear operator $\Gamma : \mathrm{Map}(\mathcal{X}, \mathbb{R}) \to \mathrm{Map}(\mathcal{U}, \mathbb{R})$ defined by

$$\Gamma(\delta)(u) = \delta(s(u))\big(1 - \gamma f(s(u), u)\big) - \gamma \sum_{y \neq s(u)} \delta(y) f(y, u),$$

for all $u \in \mathcal{U}$. By using the vector identification introduced in the previous section, we can identify δ with a vector $\Delta \in \mathbb{R}^n$ and write the reward update as

$$\mathcal{L}^\Delta : R \mapsto R + B\Delta, \tag{4.2}$$

where the matrix B is given by

$$B \coloneqq \gamma F - \mathrm{diag}_i\{\mathbb{1}_{m_i}\}, \tag{4.3}$$

in which i ranges from 1 to n. It is interesting to observe that this subgroup of transformations is the result of a faithful group action of the additive group $\mathrm{Map}(\mathcal{X}, \mathbb{R}) \simeq \mathbb{R}^n$ on $\mathrm{Map}(\mathcal{U}, \mathbb{R}) \simeq \mathbb{R}^m$. More precisely, the map $\ell : \mathrm{Map}(\mathcal{X}, \mathbb{R}) \times \mathrm{Map}(\mathcal{U}, \mathbb{R}) \to \mathrm{Map}(\mathcal{U}, \mathbb{R})$, satisfies:

$$\begin{aligned} \ell_0 &= \mathrm{id}_\mathcal{U} \\ \ell_{\delta_1 + \delta_2} &= \ell_{\delta_1} \circ \ell_{\delta_2} \end{aligned}$$

and, since Γ is injective, for every distinct pair $\delta_1 \neq \delta_2$ there correspond distinct transformations $\ell_{\delta_1} \neq \ell_{\delta_2}$. It immediately follows that the space $\mathrm{Map}(\mathcal{U}, \mathbb{R})$ can be partitioned by the orbits of this group action (henceforth referred to as ℓ-orbits), which, since Γ is injective, are affine n-dimensional subspaces. This allows us to align with [12] and define an equivalence relation between model-coupled MDPs.

Definition 5. *We consider the reward functions r_1, r_2 of two model-coupled MDPs to be equivalent, and we write $r_1 \sim_\ell r_2$, iff they belong to the same ℓ-orbit. In symbols*

$$r_1 \sim_\ell r_2 \iff \exists \delta \in Map(\mathcal{X}, \mathbb{R}) : r_2 = \ell_\delta(r_1).$$

Two model-coupled MDPs with equivalent reward functions are said to be affine-equivalent.

Remark 4. *In the setting of model-coupled MDPs, each MDP is uniquely determined by its reward function, allowing us to use the MDP and its reward function interchangeably throughout the discussion.*

As anticipated, this group action is in fact related to the preservation of the relative position of the value functions, which is in turn strictly related to the invariance of the advantage functions. This result was already established in [12] and therefore we recall it here without proof.

Proposition 2 ([12, Theorem 3.5])**.** *If $r \sim_\ell r'$ are two equivalent reward functions, then:*

$$adv_{\pi,r'} = adv_{\pi,r}, \quad \forall \pi \in \Gamma(\mathcal{X}, \mathcal{U}).$$

Moreover, the corresponding value functions are related by:

$$v_{\pi,r'} = v_{\pi,r} - \delta, \quad \forall \pi \in \Gamma(\mathcal{X}, \mathcal{U}).$$

In view of the previous proposition, we will refer to these transformations as *advantage-preserving* throughout the text. Another useful property highlighted in [12] is summarized in the following lemma, of which we omit the proof.

Lemma 10. *For any reward function r and deterministic policy π, it holds*

$$\ell_{v_{\pi,r}}(r) = adv_{\pi,r}.$$

In summary, if we set δ as the value function associated to a given policy and reward function, the corresponding transformation maps the reward function into its advantage function relative to that policy. These facts lead to the following useful result, which follows directly from the properties of the advantage function $\mathrm{adv}_{\pi^*,r}$ relative to an optimal policy π^*.

Corollary 1. *For any reward function r, if $v_{\pi^*,r}$ is the optimal value function, then the transformed reward function $\ell_{v_{\pi^*,r}}(r)$ is nonpositive and the corresponding MDP is in normal form.*

Leveraging these considerations we also provide the following result, which is a consequence of the faithfulness of the action ℓ.

Theorem 4. *Consider a class $\mathfrak{M}(\cdot, \gamma f)$ of model-coupled MDPs. Then:*

(a) In every ℓ-orbit there exists a unique MDP in normal form.

(b) An MDP $\mathfrak{M}(r, \gamma f)$ is in normal form if and only if its reward function satisfies

$$\max_{u \in \mathcal{U}_x} r(u) = 0, \quad \forall x \in \mathcal{X}. \tag{4.4}$$

Proof. (a) The existence follows directly from Corollary 1. To show uniqueness assume π^* is an optimal policy for the MDP and assume the latter to be in normal form. In particular, by (1.9), this implies that $r_{\pi^*} = 0$. Now, consider any normal reward function r' in the same orbit, i.e. such that $r'_{\pi^*} = 0$ and $\ell_\delta(r) = r'$ for some $\delta \in \mathrm{Map}(\mathcal{X}, \mathbb{R})$. From the latter we have, in vector notation:

$$\begin{aligned} R'_{\pi^*} &= \big(\mathcal{L}^\Delta R\big)_{\pi^*} \\ &= \big(R - [\mathrm{diag}_i\{\mathbb{1}_{m_i}\} - \gamma F]\,\Delta\big)_{\pi^*} \\ &= \big(\underbrace{\Pi^* R}_{=0} - \Pi^*[\mathrm{diag}_i\{\mathbb{1}_{m_i}\} - \gamma F]\,\Delta\big) \\ &= \big(I_n - \gamma F_{\pi^*}\big)\Delta \end{aligned}$$

and thus, from the normality of R',

$$0 = R'_{\pi^*} = \big(I_n - \gamma F_{\pi^*}\big)\Delta \iff \Delta = 0$$

since $(I_n - \gamma F_{\pi^*}) \in GL(n, \mathbb{R})$. Hence, $r' = \ell_0(r) = r$, which proves its uniqueness.
(b) By putting together Corollary 1 and Theorem 4-a it follows that the reward function of a normal MDP is nonpositive and maps to zero at least one action in the fiber of each state, implying (4.4). Conversely, if we assume (4.4), it follows that the reward function is nonpositive and that there exists a (greedy) policy $\hat{\pi} \in \Gamma(\mathcal{X}, \mathcal{U})$, satisfying $\hat{\pi}(x) \in \mathrm{argmax}_{u \in \mathcal{U}_x} r(u)$, for all $x \in \mathcal{X}$, such that $v_{\hat{\pi},r} = 0$ (recall (1.9)). Now, by Lemma 10 it follows that

$$\mathrm{adv}_{\hat{\pi},r} = \ell_{v_{\hat{\pi},r}}(r) = \ell_0(r) = r \leq 0$$

meaning that any action differing from the one selected by $\hat{\pi}$ in the same fiber yields a nonpositive advantage. Therefore, $\hat{\pi}$ is optimal and, since its value function is zero, the MDP is in normal form. □

Based on this result, we give the following definition.

Definition 6. *We define the normal set for a class of comparable MDPs as the subset $\mathcal{N} \subset Map(\mathcal{U}, \mathbb{R})$ consisting of all the reward functions satisfying condition* (4.4)*. Explicitly*

$$\mathcal{N} \coloneqq \left\{ r \in Map(\mathcal{U}, \mathbb{R}) : \max_{u \in \mathcal{U}_x} r(u) = 0, \ \forall x \in \mathcal{X} \right\}. \tag{4.5}$$

Theorem 4-b motivates the term "normal" set, as it is exactly the set of all reward functions in normal form. The following theorem characterizes the structure of this set and shows that the reward space $\mathrm{Map}(\mathcal{U}, \mathbb{R})$ can be viewed as a fiber bundle over it.

Theorem 5. *Let $\mathcal{N}$ be the normal set of a class $\mathfrak{M}(\cdot, \gamma f)$ of model-coupled MDPs. Then:*

(a) The normal set $\mathcal{N}$ admits a decomposition as a union

$$\mathcal{N} = \bigcup_{\pi \in \Gamma(\mathcal{X}, \mathcal{U})} \mathcal{N}_\pi, \tag{4.6}$$

where $\mathcal{N}_\pi$, defined as

$$\mathcal{N}_\pi \coloneqq \{ r \in Map(\mathcal{U}, \mathbb{R}) : r \leq 0, r_\pi = 0 \}, \tag{4.7}$$

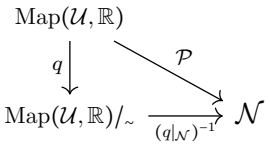
is the set of normal rewards for which the policy π is optimal.

(b) The reward space $Map(\mathcal{U},\mathbb{R})$ is homeomorphic to the product $\mathcal{N}\times Map(\mathcal{X},\mathbb{R})$ and the map

$$\mathcal{P} : Map(\mathcal{U},\mathbb{R}) \to \mathcal{N} \tag{4.8}$$
$$r \mapsto \ell_{v_{\pi^*,r}}(r), \tag{4.9}$$

is a continuous projection which makes it the total space of a principal bundle over $\mathcal{N}$.

Proof. (a) The decomposition naturally follows from the definition of $\mathcal{N}$, while the policy π is optimal for any $r\in\mathcal{N}_\pi$ since it satisfies $r_\pi = 0$ and has zero value function.
(b) Since the action is faithful, each ℓ-orbit is isomorphic to the group $\mathrm{Map}(\mathcal{X},\mathbb{R})$ and also intersects the normal set in just one point by Theorem 4-a. Moreover, by Corollary 1 and Theorem 4-a, the map $\mathcal{P}$ defined above maps each ℓ-orbit in its unique intersection point with $\mathcal{N}$, it is constant in every fiber (since by Theorem 5.a each fiber has a unique set of optimal policies identified by its unique normal representative), it is surjective and idempotent, since for each $r\in\mathcal{N}$ it holds $\mathcal{P}(r) = \ell_{v_{\pi^*,r}}(r) = \ell_0(r) = r$ and thus $\mathcal{P}|_{\mathcal{N}} = \mathrm{id}_{\mathcal{N}}$. Therefore, we can identify $\mathrm{Map}(\mathcal{U},\mathbb{R})$ with the product $\mathcal{N}\times\mathrm{Map}(\mathcal{X},\mathbb{R})$ through the mapping $r\mapsto(\mathcal{P}(r), v_{\pi^*,r})$ whose inverse is continuous and given by $(r_n,\delta)\mapsto r_n+\Gamma(\delta)$, where Γ is the linear transform characterizing the ℓ-action (recall (4.1)). It only remains to prove that $\mathcal{P}$ is continuous. Consider the quotient $\mathrm{Map}(\mathcal{U},\mathbb{R})/_\sim \coloneqq \mathrm{Map}(\mathcal{U},\mathbb{R})/\mathrm{Map}(\mathcal{X},\mathbb{R})$ with quotient map q, whose restriction $q|_{\mathcal{N}}$ is continuous and bijective, since every point of $\mathcal{N}$ lies in one and only one ℓ-orbit. The projection $\mathcal{P}$ can be expressed as $\mathcal{P} = (q|_{\mathcal{N}})^{-1}\circ q$ and it is continuous if (and only if, by the properties of quotient maps) $(q|_{\mathcal{N}})^{-1}$ is continuous (see commutative diagram below).



If $U\subseteq\mathcal{N}$ is an arbitrary $\mathcal{N}$-closed set it is also $\mathrm{Map}(\mathcal{U},\mathbb{R})$-closed, since $\mathcal{N}$ is $\mathrm{Map}(\mathcal{U},\mathbb{R})$-closed. Furthermore, $q^{-1}(q(U)) = \{r : r = r_n+\Gamma(\delta), r_n\in U, \delta\in\mathrm{Map}(\mathcal{X},\mathbb{R})\}$, and it is therefore $\mathrm{Map}(\mathcal{U},\mathbb{R})$-closed, implying that $q(U)$ is $(\mathrm{Map}(\mathcal{U},\mathbb{R})/_\sim)$-closed, being the preimage of a closed set through a quotient map. This further implies that q is also a closed map and its restriction $q|_{\mathcal{N}}$ is, in fact, a homeomorphism. □

The previous theorem has two main consequences. The first result (Theorem 5.a), which holds for all comparable MDPs, implies that a policy π is optimal for a given MDP if and only if the corresponding ℓ-orbit intersects $\mathcal{N}_\pi$. Note that the subsets $\mathcal{N}_\pi$ are not disjoint. Their intersections correspond precisely to normal rewards that admit multiple optimal policies. In particular, if two arbitrary comparable MDPs intersect the normal set in $\mathcal{N}_{\pi^*}\setminus\bigcup_{\pi\neq\pi^*}\mathcal{N}_\pi$, then they share the same unique optimal policy π^*. Hence, they can be regarded as equivalent with respect to the normalization process: replacing one with the other yields, after the normalization process, a policy that is optimal for both. On the other hand, suppose one MDP intersects $\mathcal{N}_{\pi^*}\setminus\bigcup_{\pi\neq\pi^*}\mathcal{N}_\pi$ while the other intersects $\mathcal{N}_{\pi^*}\cap\mathcal{N}_\pi$. Then they cannot be considered equivalent. In fact, normalizing the first still yields the unique optimal policy π^*, which remains optimal for the second (although the other optimal policy π would be ignored). However, normalizing the second would yield two optimal policies, only one of which is optimal for the first. A tie-breaking rule could then select a policy that is suboptimal for the first.

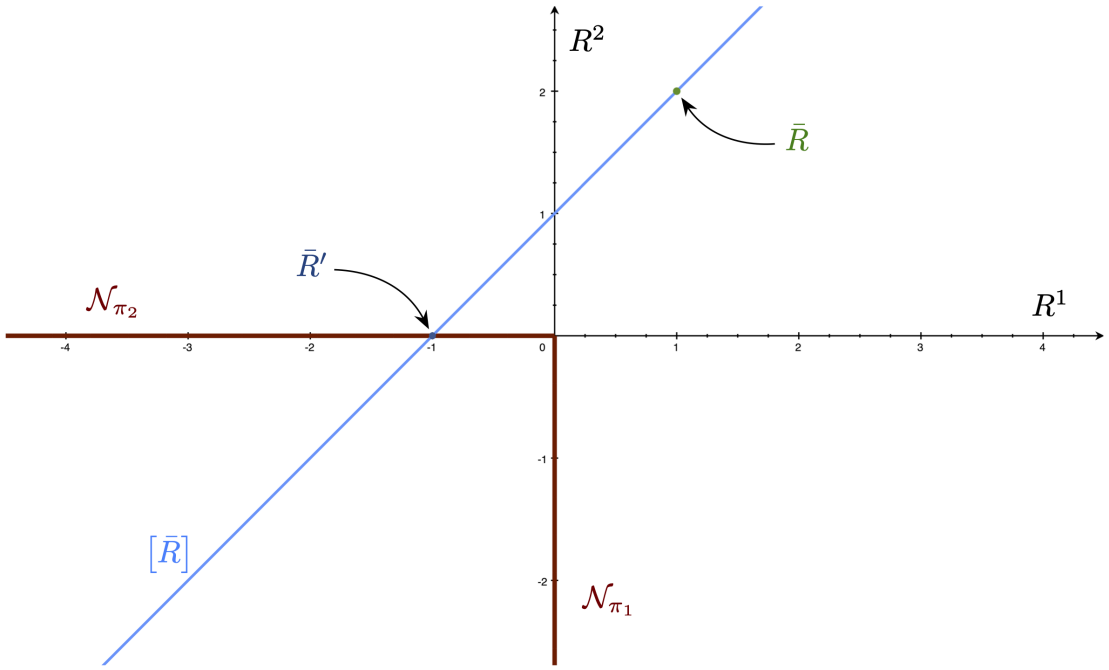


Figure 4.1: Representation of the space $\mathrm{Map}(\mathcal{U}, \mathbb{R})$ of an MDP with just one state and two actions to choose from, for which the optimal policy is greedy regardless of the chosen reward and discount factor, with an optimal value function given by $V_{\pi^*,r}(x) = \max\{R^1, R^2\}/(1-\gamma)$. The dark red set identifies a section of the normal set, in this case defined by $\mathcal{N} = \mathcal{N}_{\pi_1} \cup \mathcal{N}_{\pi_2}$, where $\mathcal{N}_{\pi_i} = \{R \in \mathbb{R}^2 : R^i = 0, R^j \leq 0, j \neq i\}$, while the light blue line is the orbit corresponding to the reward $\bar{R} = (1, 2)$. Notice that the orbit intersects the normal set in just one point, corresponding to its unique normalization $\bar{R}' = (-1, 0)$.

These considerations should be taken into account when dealing with uncertainty in the model used to normalize an unknown or partially known MDP.
The second result (Theorem 5.b) shows that any reward function r is uniquely determined by the pair consisting of the corresponding normal reward function r_n and optimal value function $v_{\pi^*,r}$, which leads to a natural identification:

$$r \leftrightarrow (r_n, v_{\pi^*,r}). \tag{4.10}$$

Figure 4.1 shows the principal bundle geometry of a small dimensional MDP. This represents one of the few instances in which such geometry can be visualized explicitly, as in most other cases the corresponding spaces have dimension greater than three.
We conclude this subsection by briefly characterizing those reward transformations that, when the original MDP is acted upon by a group G, preserve its symmetry. Recalling Definition 3 and (1.11), we obtain the following result.

Lemma 11. *If* $(\mathfrak{M}(r, \gamma f), G)$ *is a* G*-invariant MDP, then:*

$$\left[\ell_\delta(r)\right] \circ \rho_g^{\mathrm{u}} = \ell_{\delta \circ \rho_g^{\mathrm{x}}}(r), \tag{4.11}$$

for all $\delta \in Map(\mathcal{X}, \mathbb{R})$, or, in vector representation:

$$A_g^{\mathrm{u}}\left(\mathcal{L}^{\Delta} R\right) = \mathcal{L}^{A_g^{\mathrm{x}}\Delta} R. \tag{4.12}$$

Proof. Let us apply an arbitrary (left) action A_g^{u} to both sides of the update rule (4.2):

$$\begin{aligned} A_g^{\mathrm{u}}\left(\mathcal{L}^{\Delta} R\right) &= A_g^{\mathrm{u}} R - A_g^{\mathrm{u}}\left[\mathrm{diag}_i\{1_{m_i}\} - \gamma F\right]\Delta \\ &\overset{(a)}{=} R - \left[A_g^{\mathrm{u}}\mathrm{diag}_i\{1_{m_i}\} - \gamma F A_g^{\mathrm{x}}\right]\Delta \\ &= R - \left[A_g^{\mathrm{u}}\mathrm{diag}_i\{1_{m_i}\}\left(A_g^{\mathrm{x}}\right)^{\top} - \gamma F\right] A_g^{\mathrm{x}}\Delta \\ &\overset{(b)}{=} R - \left[\mathrm{diag}_i\left\{1_{m_i}\right\} - \gamma F\right] A_g^{\mathrm{x}}\Delta, \end{aligned}$$

where (a) follows from the invariance properties (1.11) together with the fact that each permutation matrix is an orthogonal matrix, implying $A_g^{\mathrm{u}} F = F A_g^{\mathrm{x}}$. Identity (b) follows from the structure of A_g^{u}. On one hand it permutes the row blocks corresponding to the states. On the other hand it permutes the rows within each block, which correspond to the actions in the corresponding fiber. The latter permutation has no effect on the block-diagonal matrix $\mathrm{diag}_i\{\mathbb{1}_{m_i}\}$, since each block contains identical rows. Hence, the effect of A_g^{u} on this matrix is to permute the state blocks. This can be formalized by thinking of it as a diagonal matrix D, where each row corresponds to a block. In this case:

$$A_g^{\mathrm{x}} D (A_g^{\mathrm{x}})^{\top}$$

remains diagonal, with its diagonal elements permuted according to the action ρ_g^{x}. Finally, the group action ρ_g^{x} permutes states with the same number of actions, and then a diagonal block $\mathbb{1}_{m_i}$ may only be replaced by a block $\mathbb{1}_{m_j}$ with the same size ($m_i = m_j$), ensuring that the block-diagonal structure is preserved under the group action. □

Lemma 11 finally implies the following characterization.

Proposition 3. *Let $(\mathfrak{M}(r, \gamma f), G)$ be a G-invariant MDP. A reward transformation ℓ_δ preserves G-invariance if and only if δ is G-invariant, i.e. $\delta \circ \rho_g^{\mathrm{x}} = \delta$ for all $g \in G$.*

Proof. ($\Rightarrow$) : If $\mathcal{L}^{\Delta} R$ is invariant, $A_g^{\mathrm{u}}(\mathcal{L}^{\Delta} R) = \mathcal{L}^{\Delta} R$ for all $g \in G$. Then, by Lemma 11, we have:

$$\underbrace{\left[-\mathrm{diag}_i\left\{1_{m_i}\right\} + \gamma F\right]}_{=B}\left(A_g^{\mathrm{x}} - I_n\right)\Delta = 0, \quad \forall g \in G,$$

and, since B is injective:

$$(A_g^{\mathrm{x}} - I_n)\Delta = 0, \quad \forall g \in G.$$

Therefore:

$$A_g^{\mathrm{x}}\Delta = \Delta \iff \delta \circ \rho_g^{\mathrm{x}} = \delta.$$

($\Leftarrow$) : Follows directly from Lemma 11. □

The Set of Feasible Transformations

So far we have introduced and analyzed the group of advantage-preserving transformations and their action on $\mathrm{Map}(\mathcal{U}, \mathbb{R})$. However, not every such transformation is useful for the purpose of (iteratively) normalizing the MDP. For instance, if the MDP is already in normal form, the only feasible transformation is the identity, corresponding to $\delta = 0$, by Theorem 4-a. In particular, we are interested in those transformations which, when applied to a given reward function, shift the corresponding optimal value function closer to zero. By Proposition 2 this requirement on ℓ_δ can be expressed by asking for the corresponding (unique) group element δ to belong to the following set:

$$\mathcal{B}_*(r) \coloneqq \{\delta : |v_{\pi^*,r}(x) - \delta(x)| \leq |v_{\pi^*,r}(x)|, \ \forall x \in \mathcal{X}\}. \tag{4.13}$$

However, as we will see in the following subsections, it is convenient to further restrict our attention to nonpositive reward functions. Among other benefits, this restriction allows us to reformulate the normalization of the MDP as an optimal control problem using condition (4.4). Unsurprisingly, it can always be assumed, without loss of generality, that the initial reward function is nonpositive, as justified by the following lemma, which we include for completeness.

Lemma 12. *Given any reward function r, the transformation $r(u) \mapsto r(u) - \max_{a \in \mathcal{U}} r(a) \leq 0$ is advantage-preserving.*

Proof. In vector notation, the transformation corresponds to

$$\Delta = \frac{\max_j R^j}{1-\gamma} \mathbb{1}_n,$$

since

$$\begin{aligned} R' &= R - \frac{\max_j R^j}{1-\gamma} [\mathrm{diag}_i\{\mathbb{1}_{m_i}\} - \gamma F] \mathbb{1}_n \\ &= R - \left(\max_j R^j\right) \mathbb{1}_n \end{aligned}$$

where the last identity follows by the row-stochasticity of the F matrix. □

It is therefore reasonable to require admissible transformations to also preserve nonpositivity, namely to satisfy

$$r \leq 0 \implies \ell_\delta(r) \leq 0, \tag{4.14}$$

which is equivalent to requiring δ to solve the following system of m linear inequalities

$$\Gamma(\delta)(u) \geq r(u), \quad \forall u \in \mathcal{U}. \tag{4.15}$$

We denote by $\mathcal{C}_u(r)$ the closed half-space defined by each of these inequalities and by $\mathcal{C}(r)$ the set

$$\mathcal{C}(r) = \bigcap_{u \in \mathcal{U}} \mathcal{C}_u(r), \tag{4.16}$$

corresponding to the solutions of the system (4.15).

Remark 5. *Each action can be identified with the boundary hyperplane of the corresponding half space $\mathcal{C}_u(r)$ through the map*

$$u \mapsto \partial\mathcal{C}_u(r) = \{\delta : \Gamma(\delta)(u) = r(u)\}.$$

The coefficients of the defining linear-affine function $r(u) - \Gamma(\cdot)(u)$ for $\partial\mathcal{C}_u(r)$ determine the entries of the action vector corresponding to u, as introduced in [12]. Consequently, the actions of the MDP define an arrangement of hyperplanes in the space $Map(\mathcal{X}, \mathbb{R})$. Moreover, for any policy π it holds

$$\{v_{\pi,r}\} = \bigcap_{x\in\mathcal{X}} \partial\mathcal{C}_{\pi(x)}(r). \tag{4.17}$$

The discussion above motivates the following definition.

Definition 7. *For any nonpositive reward function r, we define the admissible set of transformations for the normalization problem as*

$$\mathcal{A}(r) \coloneqq \mathcal{C}(r) \cap \mathcal{B}_*(r). \tag{4.18}$$

We summarize the key properties of the admissible set in the following proposition.

Proposition 4. *For any nonpositive reward function r, the admissible set $\mathcal{A}(r)$ is nonempty, compact and convex.*

Proof. The admissible set is nonempty since both the zero function and, by Corollary 1, the optimal value function belong to this set. The fact that it is compact and convex follows immediately from the fact that $\mathcal{A}(r)$ is the intersection between two closed convex sets, one of which, $\mathcal{B}_*(r)$, is compact. □

Observe that the nonpositivity of the reward function implies the nonpositivity of each value function (recall (1.1)), including the optimal one. Consequently, the set $\mathcal{B}_*(r)$ is a compact subset of the nonpositive orthant. Together with the following proposition, this provides a clear geometric interpretation of each function δ belonging to this set.

Proposition 5. *For any reward function r, every function in $\mathcal{C}(r)$ is an overestimate of the corresponding optimal value function. Formally:*

$$\delta \in \mathcal{C}(r) \implies \delta(x) \geq v_{\pi^*,r}(x), \ \forall x \in \mathcal{X}.$$

Proof. If $\delta \in \mathcal{C}(r)$ then it satisfies

$$r_\pi(x) + (\gamma f_\pi(x,x) - 1)\delta(x) + \sum_{y\neq x} \gamma f_\pi(y,x)\delta(y) \leq 0$$

for every state $x \in \mathcal{X}$ and any fixed policy $\pi \in \Gamma(\mathcal{X},\mathcal{U})$. Since $v_{\pi,r}$ satisfies all of these inequalities with equality (by the Bellman equations (1.2)), they can be rewritten as

$$\sum_{y\in\mathcal{X}} \gamma f_\pi(y,x)\tilde{\delta}_{\pi,r}(y) \leq \tilde{\delta}_{\pi,r}(x) \tag{4.19}$$

where we introduced the shortcut $\tilde{\delta}_{\pi,r} \coloneqq \delta - v_{\pi,r}$. Since the left-hand side of (4.19) is a convex combination scaled by γ, we have

$$\gamma \min_{z \in \mathcal{X}} \tilde{\delta}_{\pi,r}(z) \leq \sum_{y \in \mathcal{X}} \gamma f_\pi(y, x) \tilde{\delta}_{\pi,r}(y) \leq \tilde{\delta}_{\pi,r}(x).$$

Now, since this holds for all $x \in \mathcal{X}$, we can conclude that

$$\gamma \min_{z \in \mathcal{X}} \tilde{\delta}_{\pi,r}(z) \leq \min_{x \in \mathcal{X}} \tilde{\delta}_{\pi,r}(x)$$

or

$$(1 - \gamma) \min_{z \in \mathcal{X}} \tilde{\delta}_{\pi,r}(z) \geq 0,$$

which, since $\gamma < 1$, implies that $\tilde{\delta}_{\pi,r}(x) = \delta(x) - v_{\pi,r}(x) \geq 0, \forall x$. Since the policy π is arbitrary, this holds for the optimal one as well, and this concludes the proof. □

We can therefore interpret the admissible set as the collection of all the nonpositive overestimates of the optimal value function $v_{\pi^*,r}$, whose corresponding transformations preserve the nonpositivity of the given reward function. This finally yields the following characterization.

Proposition 6. *The admissible set of any nonpositive reward function r can be expressed in terms of linear inequalities as follows:*

$$\mathcal{A}(r) = \{\delta \in Map(\mathcal{X}, \mathbb{R}) : \delta \leq 0,\ \ell_\delta(r) \leq 0\}. \tag{4.20}$$

Proof. If we let $\mathcal{C}_*(r)$ denote the affine cone of overestimates of the optimal value function, and let $\mathcal{O}_\leq$ denote the nonpositive orthant of $\mathrm{Map}(\mathcal{X}, \mathbb{R})$, then we have

$$\mathcal{C}_*(r) \cap \mathcal{B}_*(r) = \mathcal{C}_*(r) \cap \mathcal{O}_\leq,$$

since $\mathcal{B}_*(r) \subseteq \mathcal{O}_\leq$ trivially implies $\mathcal{C}_*(r) \cap \mathcal{B}_*(r) \subseteq \mathcal{C}_*(r) \cap \mathcal{O}_\leq$, while:

$$\begin{aligned}
\delta \in \mathcal{C}_*(r) \cap \mathcal{O}_\leq &\iff 0 \geq \delta(x) \geq v_{\pi^*,r}(x),\ \ \forall x \in \mathcal{X} \\
&\iff -v_{\pi^*,r}(x) \geq \delta(x) - v_{\pi^*,r}(x) \geq 0,\ \ \forall x \in \mathcal{X} \\
&\implies \delta \in \mathcal{B}_*(r),
\end{aligned}$$

and thus $\mathcal{C}_*(r) \cap \mathcal{O}_\leq \subseteq \mathcal{C}_*(r) \cap \mathcal{B}_*(r)$, by monotonicity of intersection. This further implies, by Propsosition 5, that

$$\mathcal{C}(r) \cap \mathcal{B}_*(r) = \mathcal{C}(r) \cap \mathcal{O}_\leq.$$

□

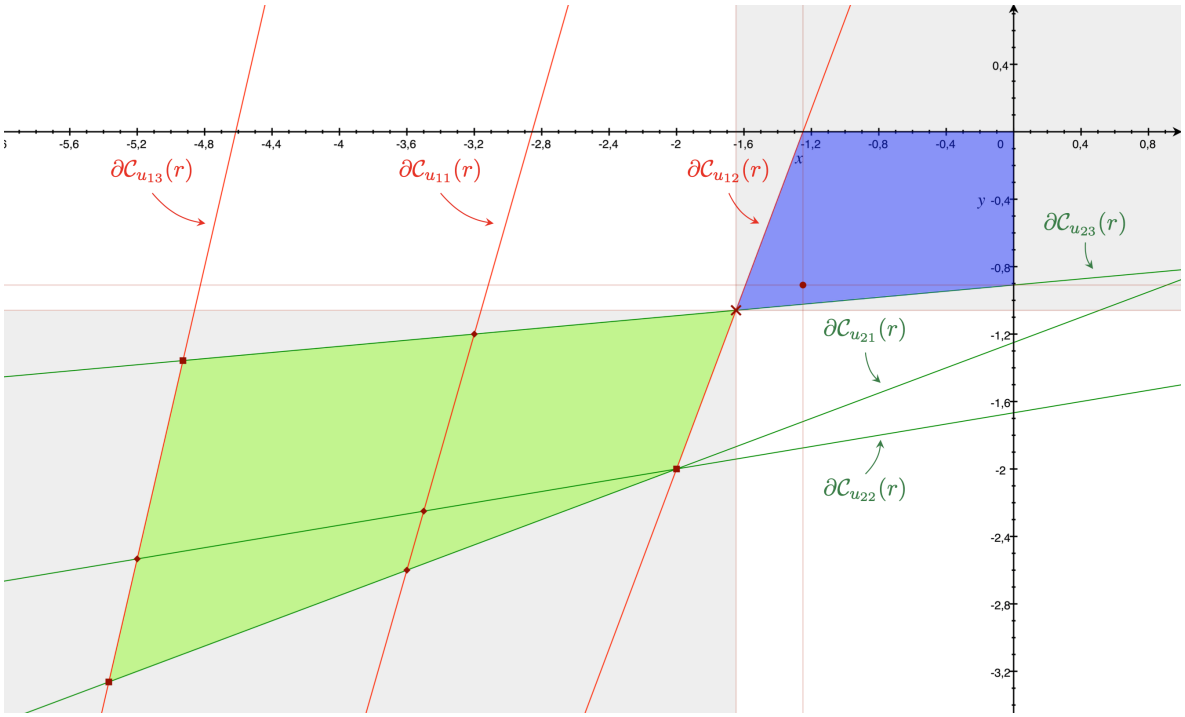


Figure 4.2: Visualization of the admissible set for an MDP with two states x_1, x_2 and an action space $\{u_{ij} : i \in \{1,2\}, j \in \{1,2,3\}\}$, where $u_{ij} \in \mathcal{U}_{x_i}$. Each intersection point $\partial\mathcal{C}_u(r) \cap \partial\mathcal{C}_v(r)$, with $u \in \mathcal{U}_{x_1}$ and $v \in \mathcal{U}_{x_2}$, is the value function corresponding to the policy $\pi : (x_1, x_2) \mapsto (u, v)$. The optimal value function is marked with a red cross, whereas the admissible set is the blueish polygon. The green polygon is the convex hull of the set of deterministic value functions (representing value functions corresponding to stochastic policies as well) and the light-grey cones represent the set of underestimates (bottom-left) and overestimates (top-right) of the optimal value function.

Control-Theoretic Reformulation

In the previous subsection, we introduced the reward-dependent set of transformations to be used in the design of a normalizing algorithm for a given MDP. Now, we now reformulate this task within a system-theoretic framework, with the idea of restating it as a control problem. To this end, consider the vector representation (4.2) of a general advantage-preserving reward transformation and let us rewrite it as an iterative law

$$R_{t+1} = R_t + B\Delta_t, \tag{4.21}$$

where B is defined as in (4.3), and $t \in \mathbb{N}$. In doing so, we are considering the reward vector R as the state of a linear time-invariant system and Δ as a control input. Hence, our problem boils down to searching for a feedback control law $R \mapsto \Delta_{\text{FB}}(R)$ able to steer the reward R, initialized in the nonpositive orthant, towards the normal set. Hence, the normalization problem can be interpreted as an output stabilization problem, with output map given by

$$R \mapsto \underbrace{(I_n - \gamma F_{\pi^*})^{-1}\Pi^* R}_{=V_{\pi^*,r}} \tag{4.22}$$

where π^* is any optimal policy for the MDP. However, this output map is impractical, since computing it explicitly is essentially equivalent to solving the very problem the normalization

process is meant to address. Therefore, we leverage Theorem 4-b and replace (4.22) with a measurable nonlinear output map $h : \mathbb{R}^m \to \mathbb{R}^n$ with components defined by

$$h^i(R) = \max_{j \in \mathcal{U}_i} R^j, \quad i \in \{1, \cdots, n\}. \tag{4.23}$$

Consequently, we consider the dynamical system

$$R_{t+1} = R_t + B\Delta_t, \tag{4.24a}$$
$$y_t = h(R_t), \tag{4.24b}$$

and seek an admissible control input $\Delta_t \in \mathcal{A}(R_t)$ that steers y_t to zero. Any such control input, which we will refer to as a *normalizing control law*, is characterized by a property that follows from Theorem 5.b and from the faithfulness of the action ℓ.

Proposition 7. *A control law $t \mapsto \Delta_t$ for the dynamics (4.21) is normalizing if and only if it satisfies*

$$\sum_{t=0}^{\infty} \Delta_t = V_{\pi^*, r_0}.$$

Proof. By Theorem 5.b, every reward vector R can be uniquely identified with a tuple $(\bar{R}, V_{\pi^*,r})$, where $\bar{R}$ is its normal representative and $V_{\pi^*,r}$ is its optimal value function. In particular, it can be expressed as

$$R = \bar{R} - BV_{\pi^*,r}.$$

We can exploit this fact to easily put the dynamics in error coordinates with respect to the unique normal reward. In particular, we have:

$$\begin{aligned} R_{t+1} &= R_t + B\Delta_t \\ &= R_0 + B\sum_{\tau=0}^{t} \Delta_\tau \\ &= \bar{R} - BV_{\pi^*,r_0} + B\sum_{\tau=0}^{t} \Delta_\tau, \end{aligned}$$

and thus, by introducing $\tilde{R} \coloneqq R - \bar{R}$, we finally get:

$$\tilde{R}_{t+1} = -B\left(V_{\pi^*,r_0} - \sum_{\tau=0}^{t} \Delta_\tau\right). \tag{4.25}$$

The statement then follows from the injectivity of B. □

In other words, in order to normalize an MDP with reward function r, a control law $\{\Delta_t\}_{t\in\mathbb{N}}$ must constitute a partition of the corresponding optimal value function $v_{\pi^*/!,r}$. We now present three examples of normalizing control laws. Two of them were originally introduced in [12] as normalization algorithms: here, we recast them as control laws within our system-theoretic framework and establish additional structural properties.

I) Ideal Normalizing Control.

The first law we present is impractical, yet it serves as the prototype for all normalizing control laws.

Definition 8. *We define the ideal output-feedback control law by setting*

$$\Delta_t = V_{\pi^*, r_t}. \tag{4.26}$$

This control law is characterized by the following result, which is an immediate consequence of Corollary 1 and Proposition 7.

Proposition 8. *The ideal output-feedback control law is admissible and normalizing. In particular, it solves the problem in one step.*

It follows that the corresponding reward sequence is zero for all $t > 0$, resulting in an input sequence

$$\Delta_t = \begin{cases} (I_n - \gamma F_{\pi^*})^{-1} \Pi^* R_0, & t = 0 \\ 0, & t > 0 \end{cases} \tag{4.27}$$

II) Full-Output Feedback.
The second law we consider is the one used in [12, Algorithm 2], where it was originally applied to an approximate-model setting.

Definition 9. *The full-output feedback reward balancing method is the control law obtained by setting*

$$\Delta_t = y_t. \tag{4.28}$$

We then provide the following result, which not only establishes its normalizing properties but also offers a clear interpretation linking this method to standard dynamic programming.

Proposition 9. *The full-output feedback reward balancing control law is admissible, normalizing and equivalent to value iteration.*

Proof. The control law is feasible by Proposition 6, since if $R_t \leq 0$ then clearly, for every $i \in \{1, \cdots, n\}$, we have $\Delta_t^i = \max_{l \in \mathcal{U}_i} R_t^l \leq 0$, and, for every $j \in \mathcal{U}_i$

$$\begin{aligned} R_{t+1}^j &= R_t^j - \Delta_t^i + \gamma \sum_{k=1}^n F_k^j \Delta_t^k && (4.29) \\ &= R_t^j - \max_{l \in \mathcal{U}_i} R_t^l + \gamma \sum_{k=1}^n F_k^j \max_{l \in \mathcal{U}_k} R_t^l && (4.30) \\ &\leq 0. && (4.31) \end{aligned}$$

Furthermore:

$$\begin{aligned} \underbrace{\max_{j \in \mathcal{U}_i} R_{t+1}^j}_{=\Delta_{t+1}^i} &= \max_{j \in \mathcal{U}_i} \left\{ R_t^j - \Delta_t^i + \gamma \sum_{k=1}^n F_k^j \Delta_t^k \right\} \\ &= \max_{j \in \mathcal{U}_i} \left\{ R_0^j - \sum_{\tau=0}^t \Delta_\tau^i + \gamma \sum_{k=1}^n F_k^j \sum_{\tau=0}^t \Delta_\tau^k \right\} \\ &= \max_{j \in \mathcal{U}_i} \left\{ R_0^j + \gamma \sum_{k=1}^n F_k^j \sum_{\tau=0}^t \Delta_\tau^k \right\} - \sum_{\tau=0}^t \Delta_\tau^i, \end{aligned}$$

which yields

$$\sum_{\tau=0}^{t+1} \Delta_\tau^i = \max_{j \in \mathcal{U}_i} \left\{ R_0^j + \gamma \sum_{k=1}^{n} F_k^j \sum_{\tau=0}^{t} \Delta_\tau^k \right\}. \tag{4.32}$$

By setting $\hat{V}_t^i \coloneqq \sum_{\tau=0}^{t} \Delta_\tau^i$, the update (4.32) can be compactly written as

$$\hat{V}_{t+1} = \mathcal{T}\big(\hat{V}_t\big), \tag{4.33}$$

where $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ is the well known optimal Bellman operator. Consequently, (4.32) is precisely the value iteration update on $\hat{V}_t$, which is a fixed-point iteration of the γ-contraction $\mathcal{T}$. It follows that

$$\begin{aligned} \left\| V_{\pi^*,r_0} - \sum_{\tau=0}^{t} \Delta_\tau \right\|_\infty &\leq \gamma^t \| V_{\pi^*,r_0} - \Delta_0 \|_\infty \\ &= \gamma^t \max_i \left| V_{\pi^*,r_0}^i - \max_{j \in \mathcal{U}_i} R_0^j \right|, \end{aligned} \tag{4.34}$$

and thus, by Proposition 7, the control law is normalizing. □

III) Safe Reward Balancing.
The third law we present is the one used in [12, Algorithm 1], there referred to as *safe reward balancing* algorithm, a name we also retain here.

Definition 10. *The safe reward-balancing (RB-S) method is the nonlinear state-feedback control law obtained by setting*

$$\Delta_t^i = \max_{j \in \mathcal{U}_i} \frac{R^j}{1 - \gamma F_i^j}, \quad i \in \{1, \cdots, n\}. \tag{4.35}$$

The properties of this control law were already established in the analysis in the aforementioned paper. We nevertheless provide them here for completeness.

Proposition 10. *The reward-balancing control law is admissible and normalizing.*

Proof. The control law is admissible by Proposition 6 since clearly if $R_t \leq 0$, then $\Delta_t \leq 0$ and, by [12, Proposition C.1], $R_{t+1} \leq 0$. Moreover, as shown in the proof of [12, Theorem 4.4], the RB-S control law satisfies

$$\left\| V_{\pi^*,r_0} - \sum_{\tau=0}^{t-1} \Delta_\tau \right\|_\infty \leq \max_i \left\{ -y_0^i \right\} \frac{\gamma^t}{1-\gamma}, \tag{4.36}$$

which in turn implies by Proposition 7 that it is normalizing. □

Figure 4.2 provides a geometric intuition of this input choice (the red dot within the admissible set) and it illustrates that, for each nonpositive reward, it corresponds to the best overestimate of the optimal value function whose projections onto the elements of the canonical basis of $\mathbb{R}^n$ remain feasible. This geometric intuition is formalized by the following proposition.

Proposition 11. *For every nonpositive R_t, the RB-S control law satisfies the following*

$$\Delta_t^i = \min_{\alpha \in \mathbb{R}} \alpha \tag{4.37a}$$

$$\textit{subj.to: } R_t + \alpha B \hat{e}_i \leq 0 \tag{4.37b}$$

Proof. If we consider the i-th state and the j-th constraint of (4.37b) then:

- if $j \in \mathcal{U}_i$ we have

$$R^j + \alpha(\gamma F_i^j - 1) \leq 0 \implies \alpha \geq \frac{R^j}{1 - \gamma F_i^j}. \tag{4.38}$$

- if $j \notin \mathcal{U}_i$ we have

$$R^j + \alpha\gamma F_i^j \leq 0 \implies \alpha\gamma F_i^j \leq \underbrace{-R^j}_{\geq 0}. \tag{4.39}$$

Now, every lower bound in (4.38) is nonpositive and therefore also satisfies every inequality (4.39). Hence, the minimum feasible value for α is the maximum among these lower bounds, which is exactly the i-th entry of the RB-S control law. □

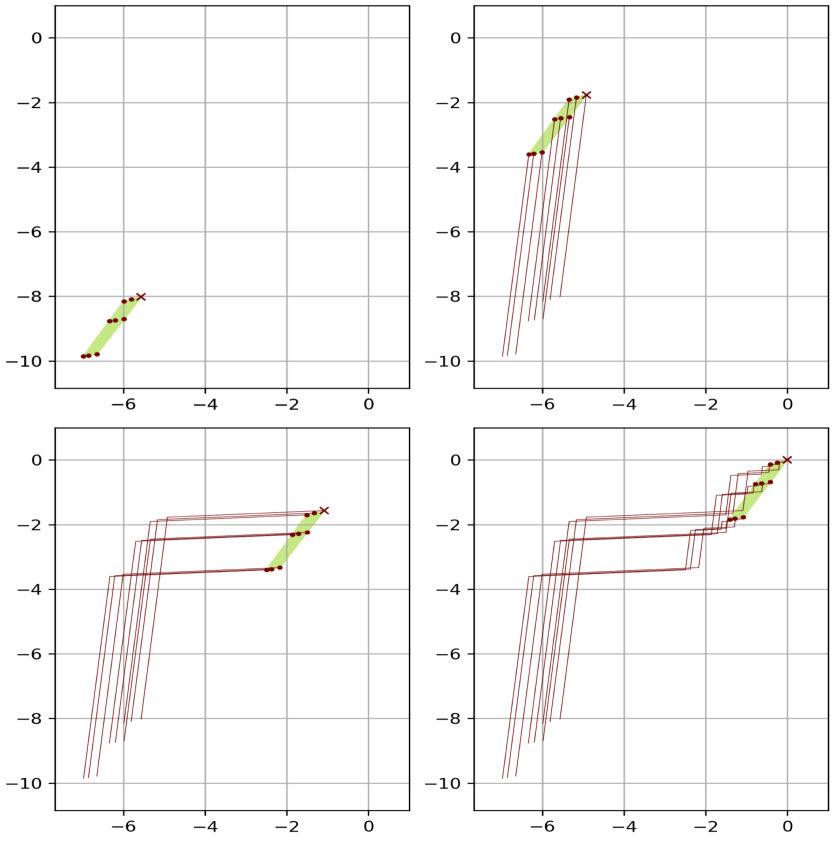


Figure 4.3: Example of the evolution of the set of value functions (green region, including stochastic policies, with the optimal one marked with a cross), for an MDP with two states and six actions in total (three available in each state) when the reward is updated using the RB-S control law computed using the exact model. Each suboptimal value function corresponding to a stationary deterministic policy is marked with a red dot, whereas the optimal one is marked with a cross. Depicted are the original set (top-left) and the set after one (top-right), two (bottom-left) and thirty iterations (bottom-right).

Normalization as an Optimal Control Problem

In this subsection we recast the normalization problem as an optimal control problem, which can be used to design normalizing control laws in either a finite-horizon setting (where normalization can be achieved approximately, or exactly via model predictive control) or in an infinite-horizon setting. This formulation also allows for incorporating a broader variety of constraints and possibly additional penalty terms in the cost. While in the exact-model case one can at most improve the decay rate of y_t, this approach enables the design of more robust control laws in the approximate-model setting. Let us then consider the following optimal control problem

$$\min_{\{(R_t,\Delta_t)\}} -\sum_{t=1}^{T}\sum_{i=1}^{n}\max_{j\in\mathcal{U}_i} R_t^j \tag{4.40a}$$

subj.to:

$$R_0 = \bar{R} \le 0 \tag{4.40b}$$

$$R_{t+1} = R_t + B\Delta_t, \tag{4.40c}$$

$$PR_t + Q\Delta_t \le 0, \tag{4.40d}$$

for all $t \in \{0,\cdots,T-1\}$

where B is defined as in (4.3), $T \in \mathbb{N}_+ \cup \{\infty\}$ and $P \in \mathbb{R}^{d\times m}, Q \in \mathbb{R}^{d\times n}$ satisfy

$$\varnothing \neq \mathcal{F}(R) \coloneqq \{\Delta : PR + Q\Delta \le 0\} \subseteq \mathcal{A}(R), \tag{4.41}$$

for all $R \le 0$. That is, we allow for restrictions to a subset of the admissible set. In particular, if we choose:

$$P = \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad Q = \begin{bmatrix} I \\ B \end{bmatrix}, \tag{4.42}$$

with $d = m + n$, we retrieve the admissible set by Proposition 6. In this case, the optimal solution is trivial as shown by the following result.

Proposition 12. *If P and Q are chosen as in* (4.42)*, then the input sequence* (4.27) *is the unique optimal solution of Problem* (4.40)*, for every $T \ge 1$.*

Proof. The input sequence (4.27) normalizes the MDP in a single step, and the corresponding stage cost is thus zero for all t. Since the reward is constrained to remain nonpositive, the cost is lower-bounded by zero and this input sequence is therefore optimal. Any other optimal solution would also need to normalize the MDP in a single step as well (by Theorem 4-b). However, Proposition 7 shows that there is no Δ_0 other than V_{π^*,r_0} capable of achieving this. □

Although known in closed form, the previous solution is impractical, as already discussed. Moreover, in the presence of model uncertainties, more conservative solutions become particularly valuable. For instance, one could make the solution more robust by restricting the feasible set or by introducing additional penalty terms in the cost. Furthermore, if the MDP exhibits symmetries, it may be desirable to enforce the input to satisfy Proposition 3, which can be expressed as a finite set of constraints of the type $\pm(A_g^{\mathrm{x}} - I_n)\Delta \le 0$, compatible with the structure of (4.40d). For general choices of P and Q, the problem admits no solution in

closed form, and exact normalization of the MDP can be achieved only asymptotically in the infinite-horizon case. In this setting, we provide two sufficient conditions on P and Q that guarantee feasibility of the problem and existence of optimal solutions.

Theorem 6. *Problem* (4.40) *with* $T = \infty$ *is feasible if, for every* $R \leq 0$*, at least one of the following conditions holds:*

(a) There exists $\alpha \in [0, 1)$ *such that:*

$$\min_{\Delta \,:\, PR + Q\Delta \leq 0} \|V_{\pi^*, r} - \Delta\|_\infty \leq \alpha \|V_{\pi^*, r}\|_\infty. \tag{4.43}$$

(b) The output map (4.23) *satisfies*

$$PR + Qh(R) \leq 0. \tag{4.44}$$

Moreover, under either condition, an optimal solution exists and has finite cost.

Proof. (a) If we choose a feasible control law satisfying

$$\Delta_t \in \operatorname*{argmin}_{\Delta \,:\, PR_t + Q\Delta \leq 0} \|V_{\pi^*, r_t} - \Delta\|_\infty,$$

we have

$$\begin{aligned}
\left\| V_{\pi^*, r_0} - \sum_{\tau=0}^{t} \Delta_\tau \right\|_\infty &= \|V_{\pi^*, r_t} - \Delta_t\|_\infty \\
&\leq \alpha \|V_{\pi^*, r_t}\|_\infty \\
&= \alpha \|V_{\pi^*, r_{t-1}} - \Delta_{t-1}\|_\infty \\
&\leq \alpha^{t+1} \|V_{\pi^*, r_0}\|_\infty
\end{aligned}$$

where in we iteratively used the relation $V_{\pi^*, r_t} = V_{\pi^*, r_{t-1}} - \Delta_{t-1}$. It follows that the stage cost is geometrically bounded, since

$$\begin{aligned}
-\max_{j \in \mathcal{U}_i} R^j_{t+1} &= \Delta^i_t - \max_{j \in \mathcal{U}_i} \left\{ R^j_t + \sum_{k=1}^{n} \gamma F^j_k \Delta^k_t \right\} \\
&= \sum_{\tau=0}^{t} \Delta^i_\tau - \max_{j \in \mathcal{U}_i} \left\{ R^j_0 + \sum_{k=1}^{n} \gamma F^j_k \sum_{\tau=0}^{t} \Delta^k_\tau \right\} \\
&= \sum_{\tau=0}^{t} \Delta^i_\tau - \left[\mathcal{T}\left(\sum_{\tau=0}^{t} \Delta_\tau \right) \right]^i - \underbrace{\left[V_{\pi^*, r_0} - \mathcal{T}\left(V_{\pi^*, r_0} \right) \right]^i}_{=0} \\
&\leq \left\| V_{\pi^*, r_0} - \sum_{\tau=0}^{t} \Delta_\tau \right\|_\infty + \left\| \mathcal{T}\left(V_{\pi^*, r_0} \right) - \mathcal{T}\left(\sum_{\tau=0}^{t} \Delta_\tau \right) \right\|_\infty \\
&\leq (1 + \gamma) \left\| V_{\pi^*, r_0} - \sum_{\tau=0}^{t} \Delta_\tau \right\|_\infty \\
&\leq (1 + \gamma) \alpha^{t+1} \|V_{\pi^*, r_0}\|_\infty,
\end{aligned}$$

and thus the summation converges.
(b) If the output map h satisfies (4.44), then the feedback law $\Delta_t = y_t$ is feasible for the

optimal control problem. It follows from (4.34) and an argument analogous to that used in the previous part, that the stage cost is geometrically bounded.
To prove existence of optimal solutions, we regard the cost functional J as an extended real-valued map defined on the set of feasible sequences $\{(R_t, \Delta_t)\}_{t\in\mathbb{N}}$, equipped with the product topology ("extended" refers to the fact that some feasible trajectories have infinite cost). We first show that this space is compact. For each $t \in \mathbb{N}$, the pair (R_t, Δ_t) belongs to the set

$$S_t(\bar{R}) \coloneqq \{(R, \Delta) : R \in \mathfrak{R}_t(\bar{R}), \Delta \in \mathcal{F}(R)\}$$

where $\mathfrak{R}_t(\bar{R})$ denotes the reachable set at time t from the initial reward $\bar{R}$. The set $S_0(\bar{R}) = \{\bar{R}\} \times \mathcal{F}(\bar{R})$ is clearly compact. Assume that $S_t(\bar{R})$ is compact for a generic time $t \in \mathbb{N}$. The reachable set at time $t+1$ is given by

$$\mathfrak{R}_{t+1}(\bar{R}) = G(S_t(\bar{R}))$$

where $G : (R, \Delta) \mapsto R + B\Delta$, and is therefore compact by continuity of G. Consider (without loss of generality) a sequence $(R_k, \Delta_k) \in S_{t+1}(\bar{R})$ such that $R_k \to \hat{R} \in \mathfrak{R}_{t+1}(\bar{R})$ as $k \to \infty$. Since

$$\mathcal{F}(R_k) \subseteq \mathcal{A}(R_k) \subset \mathcal{B}_*(R_k) \subseteq \mathcal{B}_*(\bar{R}),$$

for all $k \in \mathbb{N}$ (recall (4.13)), there exists a subsequence $(R_{k'}, \Delta_{k'})$ converging to some $\left(\hat{R}, \hat{\Delta}\right) \in \mathfrak{R}_{t+1}(\bar{R}) \times \mathcal{B}_*(\bar{R})$. Since $\Delta_{k'} \in \mathcal{F}(R_{k'})$ for every k' we have:

$$\sum_{i=1}^{m} P_i^s R_{k'}^i + \sum_{j=1}^{n} Q_j^s \Delta_{k'}^j \le 0, \quad \forall s \in \{1, \cdots, d\}.$$

Passing to the limit preserves the inequalities, and it follows by continuity that $\hat{\Delta} \in \mathcal{F}(\hat{R})$. Therefore, $S_{t+1}(\bar{R})$ is sequentially compact, and thus compact. By induction $S_t(\bar{R})$ is compact for all $t \in \mathbb{N}$, and thus the space of feasible sequences, given by $\prod_{t=0}^{\infty} S_t(\bar{R})$, is compact in the product topology by Tychonoff theorem. Furthermore, the cost is lower-semicontinuous, since it is the supremum over $\mathbb{N}$ of its partial sums, which are continuous on this space. Now, under either assumption (a) or (b) there exists a feasible solution with finite cost $\bar{J}$. Hence, the sublevel set $J^{-1}((-\infty, \bar{J}]) = J^{-1}([0, \bar{J}])$ is nonempty. By lower semicontinuity this set is closed and therefore compact as a closed subset of a compact topological space. By the extended Weierstrass theorem for lower semicontinuous functions, J attains its minimum on this set, which implies that the minimizer has finite cost. □

Approximate-Model Reward-Balancing Methods

In this section we examine the consequences of performing a reward-balancing transformation when only an approximate model is available. Specifically, we characterize the effect of an advantage-preserving transformation on the entire class of reward-coupled MDPs. Furthermore, we discuss how a sample-based normalization strategy can be used to increase the likelihood of achieving optimality in the absence of an exact model. Finally, we propose a principled framework for the design of robust normalizing laws based on a scenario MPC approach, extending the optimal control formulation introduced in the previous section.

Let $\mathfrak{M}(r,\gamma f)$ be the MDP modelling the problem we aim to solve, and suppose that we have access to an approximation $\hat{f}$ of the transition function, which could be obtained, for example, from experimental data. Consequently, our reference MDP is actually $\mathfrak{M}\big(r,\gamma\hat{f}\big)$ and we can normalize its reward function using one of the methods discussed in the previous section. However, at this point, a question naturally arises: what happens to the original MDP? This question is difficult to answer a priori, as it depends heavily on the specific model $\hat{f}$. Nevertheless, we can reasonably expect that the advantage functions, together with the relative positions of the value functions of $\mathfrak{M}(r,\gamma f)$ will be altered during the normalization procedure. This is because we are shifting the reward function of two reward-coupled MDPs along the orbit of one, thus moving it away from the orbit of the other. Clearly if $\hat{f}$ is a good approximation of f, we might still expect that, in the end, the greedy policy will be optimal for $\mathfrak{M}(r,\gamma f)$. This occurs in the fortunate case in which the orbit of the approximate MDP intersects the same sections of normal set as $\mathfrak{M}(r,\gamma f)$, as already discussed. In any case the normalized reward will inevitably differ from the unique normalization of the original MDP. The next proposition addresses part of the problem, by showing what is the effect of an advantage-preserving reward transformation on any reward-coupled MDP.

Proposition 13. *Let $\mathfrak{M}\big(r,\gamma\hat{f}\big)$ and $\mathfrak{M}(r,\gamma f)$ be two reward-coupled MDPs, and suppose that the reward function is updated with an advantage-preserving transformation of the first MDP:*

$$\hat{R} = R + \big[\gamma\hat{F} - \mathrm{diag}_i\{\mathbb{1}_{m_i}\}\big]\Delta.$$

Then, the value functions of the second MDP change according to the following relations:

$$V_{\pi,\hat{r}} = V_{\pi,r} - \Delta + \gamma\big(I_n - \gamma F_\pi\big)^{-1}\big(\hat{F}_\pi - F_\pi\big)\Delta, \tag{4.45}$$

for all $\pi \in \Gamma(\mathcal{X},\mathcal{U})$.

Proof. Consider an arbitrary policy $\pi \in \Gamma(\mathcal{X},\mathcal{U})$. Both the value functions corresponding to $\hat{r}$ and r, respectively, satisfy the corresponding Bellman equation. In vector notation:

$$0 = R_\pi - (I_n - \gamma F_\pi)V_{\pi,r} \tag{4.46}$$

$$0 = \underbrace{R_\pi - (I_n - \gamma\hat{F}_\pi)\Delta}_{=:\hat{R}_\pi} - (I_n - \gamma F_\pi)V_{\pi,\hat{r}}. \tag{4.47}$$

The comparison between (4.46) and (4.47) leads to

$$V_{\pi,\hat{r}} = V_{\pi,r} + (I_n - \gamma F_\pi)^{-1}\big(\gamma\hat{F}_\pi - I_n\big)\Delta. \tag{4.48}$$

Finally, expanding the second term on the right-hand side we get

$$\begin{aligned}(I_n - \gamma F_\pi)^{-1}\big(\gamma\hat{F}_\pi - I_n\big) &= \left(\sum_{t=0}^{\infty}\gamma^t(F_\pi)^t\right)\big(\gamma\hat{F}_\pi - I_n\big)\\ &= \gamma(I_n - \gamma F_\pi)^{-1}\hat{F}_\pi - I_n - \gamma\left(\sum_{t=0}^{\infty}\gamma^t(F_\pi)^t\right)F_\pi\\ &= -I_n + \gamma(I_n - \gamma F_\pi)^{-1}\big(\hat{F}_\pi - F_\pi\big),\end{aligned}$$

from which the proposition follows. □

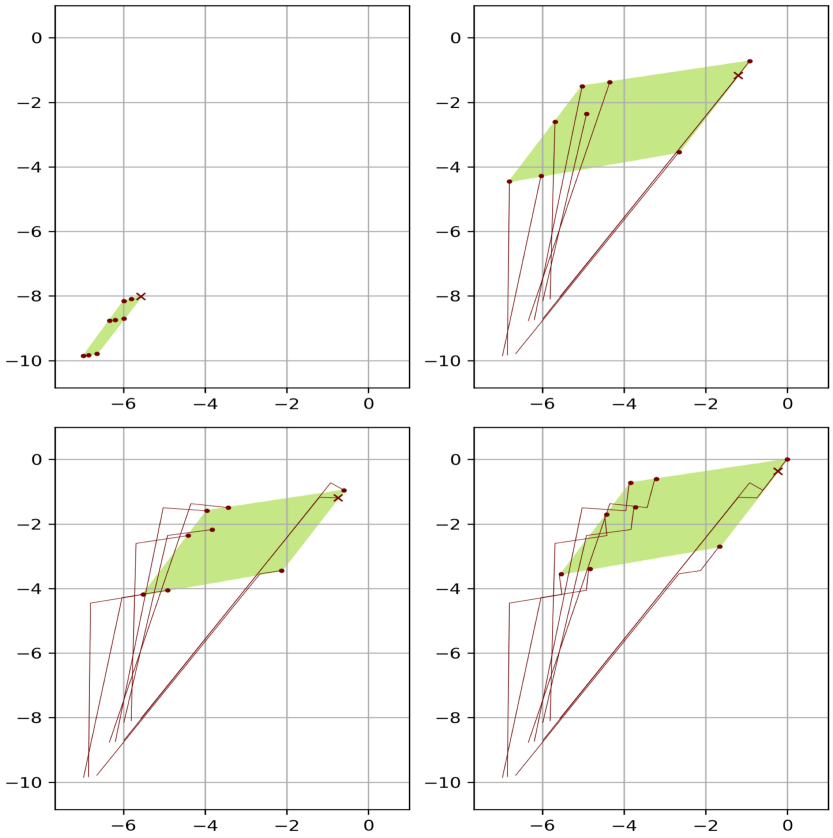


Figure 4.4: Example of the evolution of the value functions set (green region, including stochastic policies) for the same MDP as in Figure 4.3 when the reward is updated using the RB-S control law computed using a wrong model. Each suboptimal value function corresponding to a stationary deterministic policy is marked with a red dot, whereas the optimal one is marked with a cross. Depicted are the original set (top-left) and the set after one (top-right), two (bottom-left) and thirty iterations (bottom-right). In this case a suboptimal value function gets annihilated, precisely the one which is optimal for the model used to perform the reward updates.

The following proposition shows that the additional term in (4.45) can be bounded by a function of the model mismatch and the input amplitude.

Proposition 14. *The deviation of the perturbed update* (4.45) *from the exact-model update admits the following upper bound:*

$$\left\| V_{\pi,\hat{r}} - (V_{\pi,r} - \Delta) \right\|_2 \leq \varepsilon \sqrt{n} \frac{\gamma}{1-\gamma} \|\Delta\|_2 \tag{4.49}$$

where ε *is the operator norm of the model error:*

$$\varepsilon \coloneqq \left\| \hat{F} - F \right\|_2.$$

Proof. Let $\pi \in \Gamma(\mathcal{X},\mathcal{U})$ be any policy. From Proposition 13 it immediately follows that

$$\left\|V_{\pi,\hat{r}} - (V_{\pi,r} - \Delta)\right\|_2 \leq \frac{\gamma}{\sigma^{\min}_{\pi,\gamma}}\left\|\hat{F}_\pi - F_\pi\right\|_2 \|\Delta\|_2 \tag{4.50}$$

$$\leq \frac{\gamma}{\sigma^{\min}_{\pi,\gamma}}\|\Pi\|_2\left\|\hat{F} - F\right\|_2 \|\Delta\|_2 \tag{4.51}$$

$$= \varepsilon\frac{\gamma}{\sigma^{\min}_{\pi,\gamma}}\|\Delta\|_2, \tag{4.52}$$

where $\sigma^{\min}_{\pi,\gamma}$ denotes the minimum singular value of $A \coloneqq (I_n - \gamma F_\pi)$. Now, the matrix A is always diagonally dominant, regardless of the policy and the discount factor, since for all $i \in \{1,\cdots,n\}$ we have

$$\left|A^i_i\right| = 1 - \gamma(F_\pi)^i_i \geq \gamma\left(1 - [F_\pi]^i_i\right) \overset{(a)}{=} \sum_{j\neq i}\gamma[F_\pi]^i_j = \sum_{j\neq i}\left|A^i_j\right|,$$

where (a) follows from the row-stochasticity of F_π. Therefore, by [60, Theorem 1], it satisfies

$$\left\|A^{-1}\right\|_\infty < \frac{1}{\alpha},$$

with

$$\begin{aligned}
\alpha &\coloneqq \min_i\left\{\left|A^i_i\right| - \sum_{j\neq i}\left|A^i_j\right|\right\}\\
&= \min_i\left\{1 - \gamma[F_\pi]^i_i - \sum_{j\neq i}\gamma[F_\pi]^i_j\right\}\\
&= \min_i\left\{1 - \sum_{j=1}^{n}\gamma[F_\pi]^i_j\right\}\\
&= \min_i\{1-\gamma\}\\
&= 1-\gamma.
\end{aligned}$$

Finally, norm equivalence implies

$$\frac{1}{\sigma^{\min}_{\pi,\gamma}} = \left\|A^{-1}\right\|_2 \leq \sqrt{n}\left\|A^{-1}\right\|_\infty < \frac{\sqrt{n}}{1-\gamma},$$

which, used in (4.50), concludes the proof. □

The previous results characterize the structural relationship between a perturbed, advantage-preserving update and the value functions of the true model. We now turn to analyzing the long term behavior of the normalization process under model uncertainties. As already discussed, choosing a single approximate model completely determines the outcome of the normalization process, and no choice of input that can change this outcome. Consequently, in situations where a certain level of model accuracy cannot be guaranteed, a different approach is required. Suppose we have access to random realizations $\hat{B}_t$ of the true model B. Then the reward can be updated as

$$R_{t+1} = R_t + \hat{B}_t\Delta_t = R_t + B\Delta_t + \left(\hat{B}_t - B\right)\Delta_t, \tag{4.53}$$

which falls into the standard framework of stochastic systems with multiplicative noise. In this setting, the outcome of a normalization process is not predetermined by a single fixed model. A first drawback of this approach is that, in general, the admissible set changes at every model realization. It is therefore convenient to identify a model-invariant subset common to *every* admissible set. For example, the set $\{\Delta : h(R) \le \Delta \le 0\}$ is a subset of $\mathcal{A}(R)$ for all f and is therefore independent of the particular model choice. In general, however, this set is not the largest possible subset with this invariance property, as established by the following theorem.

Theorem 7. *For any fixed reward function r and discount factor $\gamma \in [0,1)$ the set:*

$$\mathcal{F}^I(R) \coloneqq \bigcap_{i,k=1}^{n} \mathcal{F}^I_{ik}(R), \tag{4.54}$$

where

$$\mathcal{F}^I_{ik}(R) \coloneqq \left\{\Delta \in \mathbb{R}^n : 0 \ge \Delta^i \ge \gamma\Delta^k + R^j,\ j \in \mathcal{U}_i\right\}, \tag{4.55}$$

satisfies (4.41) *for any model f. In particular, it is the maximal model-independent feasible set in the following sense:*

$$\mathcal{F}^I(R) = \bigcap_{f} \mathcal{A}(R). \tag{4.56}$$

Furthermore, it satisfies condition (b) of Theorem 6, thereby ensuring the existence of an optimal solution with finite cost for the nominal infinite-horizon problem (4.40).

Proof. Consider any $\Delta \in \mathcal{F}^I(R)$ and choose any model F. We need to show that $\Delta \in \mathcal{A}(R)$, i.e., by (4.20), that $\mathcal{L}^\Delta(R) \le 0$ (the other condition, $\Delta \le 0$, is satisfied by definition). Let $j \in \{1,\cdots,m\}$ be any action index and $i = s(j)$ be the corresponding state index. Then, we have:

$$\begin{aligned}
\left[\mathcal{L}^\Delta(R)\right]^j &= R^j + (\gamma F_i^j - 1)\Delta^i + \sum_{k\neq i} \gamma F_k^j \Delta^k \\
&= R^j - \Delta^i + \gamma \sum_{k=1}^{n} F_k^j \Delta^k \\
&\overset{(a)}{=} -\Delta^i + \sum_{k=1}^{n} F_k^j \left(\gamma\Delta^k + R^j\right) \\
&\overset{(b)}{\le} -\Delta^i + \underbrace{\sum_{k=1}^{n} F_k^j}_{=1} \Delta^i \\
&= 0,
\end{aligned}$$

where (a) follows from the row-stochasticity of F, whereas (b) follows from the definition of $\mathcal{F}^I_{ik}(R)$. Since j is arbitrary, this shows that $\mathcal{L}^\Delta(R) \le 0$, thus concluding the first part of the proof. In general, some of the linear inequalities defining the set $\mathcal{F}^I(R)$ could be redundant for its definition. We say that a triple (i,j,k) is active if the corresponding inequality $\Delta^i \ge \gamma\Delta^k + R^j$ is non-redundant (i.e. we cannot remove it without changing the set). Observe

that the inequality corresponding to any active triple (i, j, k) coincides with the inequality $[\mathcal{L}_\Delta(R)]^j \le 0$, for a model F such that $F_k^j = 1$ and $F_h^j = 0$ for all $h \neq k$. Indeed:

$$0 \ge R^j - \Delta^i + \sum_{h=1}^{n} \gamma F_h^j \Delta^h = R^j - \Delta^i + \gamma \Delta^k,$$

which shows that the inequality corresponding to the active triple (i, j, k) is indeed part of the definition of the feasible set for this specific model, and this concludes the proof of the second part. Finally, if $R \le 0$, for all index tuples (i, j, k), with $j \in \mathcal{U}_i$, it holds:

$$0 \ge \max_{s \in \mathcal{U}_i} R^s \ge R^j \ge R^j + \underbrace{\gamma \max_{s \in \mathcal{U}_k} R^s}_{\le 0},$$

and thus $h(R) \in \mathcal{F}^I(R)$, which is precisely condition (b) of Theorem 6. □

If, in addition, we define a model-independent feedback control law that shrinks sufficiently fast, then the stochastic system (4.53) exhibits well-behaved stochastic dynamics. This is formalized in the following result, which characterizes the stochastic process under model-independent input selection and generalizes the stochastic framework of [12].

Proposition 15. *Suppose $\{\hat{F}_t\}_{t \in \mathbb{N}}$ is a sequence of i.i.d. random realizations of F such that $\mathbb{E}\left[\hat{F}_t\right] = F$, and that the reward is updated according to (4.53) with $\hat{B}_t = \gamma \hat{F}_t - \mathrm{diag}_i\{\mathbb{1}_{m_i}\}$. If the input Δ_t is a function of R_t that does not depend on $\hat{F}_t$, then the deviation M_t from the ideal update, defined as*

$$M_0 \coloneqq 0 \tag{4.57}$$

$$M_t \coloneqq R_t - R_0 - \sum_{\tau=0}^{t-1} B \Delta_\tau, \quad t > 0, \tag{4.58}$$

is a martingale. Furthermore, if $\{\Delta_t\}_{t \in \mathbb{N}}$ is square-summable, then M_t is L^2-bounded and converges almost surely and in L^2.

Proof. By combining the definition of M_t with the update (4.53) we obtain

$$M_t = \sum_{\tau=0}^{t-1} \big(\hat{B}_\tau - B\big)\Delta_\tau = \gamma \sum_{\tau=0}^{t-1} \big(\hat{F}_\tau - F\big)\Delta_\tau, \tag{4.59}$$

for all $t > 0$. Let $\mathfrak{F}_t$ be the filtration

$$\mathfrak{F}_t \coloneqq \sigma\big(\Delta_0, \cdots, \Delta_{t-1}, \hat{F}_0, \cdots, \hat{F}_{t-1}\big), \tag{4.60}$$

then:

$$\begin{aligned}
\mathbb{E}\left[M_{t+1} \mid \mathfrak{F}_t\right] &= \mathbb{E}\left[\gamma \sum_{\tau=0}^{t} \big(\hat{F}_\tau - F\big)\Delta_\tau \mid \mathfrak{F}_t\right] \\
&= \mathbb{E}\left[\gamma \sum_{\tau=0}^{t-1} \big(\hat{F}_\tau - F\big)\Delta_\tau + \gamma\big(\hat{F}_t - F\big)\Delta_t \mid \mathfrak{F}_t\right] \\
&= \mathbb{E}\left[M_t + \gamma\big(\hat{F}_t - F\big)\Delta_t \mid \mathfrak{F}_t\right] \\
&\overset{(a)}{=} M_t + \gamma \mathbb{E}\left[\big(\hat{F}_t - F\big)\right]\Delta_t \\
&= M_t,
\end{aligned}$$

where (a) follows from the fact that Δ_t is independent of $\hat{F}_t$ and is $\mathfrak{F}_t$-measurable as a function of R_t, which depends on the past model samples and input sequence. Each component of M_k is then a real-valued martingale. Now, if the input is square-summable we have

$$\begin{aligned}
\sum_{t=0}^{\infty} \mathbb{E}\left[\left(M_t^i - M_{t-1}^i\right)^2\right] &\leq \sum_{t=0}^{\infty} \mathbb{E}\left[\|M_t - M_{t-1}\|_2^2\right] \\
&= \sum_{t=0}^{\infty} \mathbb{E}\left[\gamma^2 \left\|\left(\hat{F}_t - F\right)\Delta_t\right\|_2^2\right] \\
&\overset{(a)}{\leq} \sum_{t=0}^{\infty} \mathbb{E}\left[\gamma^2 \left\|\left(\hat{F}_t - F\right)\Delta_t\right\|_2^2\right] \\
&\leq 4n\gamma^2 \sum_{t=0}^{\infty} \|\Delta_t\|_2^2 \\
&< \infty,
\end{aligned}$$

where in (a) we used norm equivalence for the operator norms $\|A\|_2 \leq \sqrt{n}\|A\|_\infty$, and the fact that, if A and B are row-stochastic, then $\|A - B\|_\infty \leq \|A\|_\infty + \|B\|_\infty \leq 2$. The boundedness and convergence properties of this martingale follow directly from a standard result of the theory of real-valued martingales (see, e.g., [61, Theorem 12.1]) applied to each individual component of M_t. □

The previous proposition highlights the benefit of employing a control law that does not depend on the current model sample and provides a sufficient condition for the almost sure and L^2 convergence of the stochastic deviation from the average, deterministic system. We emphasize, however, that selecting an input in a model-invariant admissible set does not in itself guarantee that the resulting control law is model-independent. This distinction is particularly relevant in the optimal control framework, where the optimal control policy typically depends explicitly on the model. To mitigate this dependence, one may resort to stochastic optimization techniques such as the *scenario approach*, in which a single input trajectory is chosen to minimize the average cost over multiple realizations of the process. An instance of such a problem is the following:

$$\min_{\{(R_t,\Delta_t)\}} \frac{1}{N} \sum_{s=1}^{N} \sum_{t=0}^{T-1} \left(\sum_{i=1}^{n} - \max_{j \in \mathcal{U}_i} R_{s,t+1}^j \right) + \alpha \|\Delta_t\|_2^2 \tag{4.61a}$$

subj.to:

$$R_0 = \bar{R} \leq 0 \tag{4.61b}$$

$$R_{s,t+1} = R_{s,t} + \hat{B}_{s,t}\Delta_t, \tag{4.61c}$$

$$\Delta_t \in \mathcal{F}^I(R_{s,t}) \tag{4.61d}$$

for all $t \in \{0, \cdots, T-1\}, s \in \{1, \cdots, N\}$.

Here, a same input sequence $\{\Delta_t\}_{t=0}^{T-1}$ is applied across N sampled scenarios, and the objective minimizes the corresponding average cost. We also introduced a regularization term $\alpha\|\Delta_t\|_2^2$ to promote square-summability of the input and to reduce the variance of the deviation from the nominal update. This improved robustness, however, comes at the expense of increased computational complexity. In the following subsection we present an example in which implementing an MPC feedback law based on (4.61) outperforms the model-free feedback law of Definition 9.

Numerical Example

Consider the MDP with two states $\{x_1, x_2\}$ with five actions each $\{u_{ij}\}_{j=1}^{5}$, with transition probabilities to the first state

$f(x_1, u_{ij})$	u_{i1}	u_{i2}	u_{i3}	u_{i4}	u_{i5}
$i = 1$	0.35	0.4	0.7	0.3	0.4
$i = 2$	0.75	0.8	0.2	0.55	0.25

and with reward function

$r(u_{ij})$	u_{i1}	u_{i2}	u_{i3}	u_{i4}	u_{i5}
$i = 1$	-0.7	-0.5	-0.9	-0.45	-0.2
$i = 2$	-3.2	-2.75	-2.8	-3	-2.9

We set a scenario MPC based on (4.61) with a prediction horizon $T = 20$ and a regularization parameter $\alpha = 2$. The model samples $\hat{B}_t$ are drawn from a uniform distribution around the true model, with a maximum entry-wise deviation of 0.2. Figure 4.5 compares the percentage of runs that reach the optimal policy under scenario MPC and full-output feedback, demonstrating a significant performance improvement when employing the former, with performance further improving as the number of scenarios increases. We emphasize, however, that this example is not intended to establish the superiority of scenario MPC over the full output feedback. Indeed, the latter retains a clear advantage in terms of computational efficiency and practical deployability. Rather, the purpose is to show that there exist normalization laws capable of achieving improved performance under model uncertainty, although at a higher computational cost. This observation suggests that there remains room for further improvement, and that the proposed framework may serve as a systematic tool for the design and analysis of more effective normalization laws beyond those currently available.

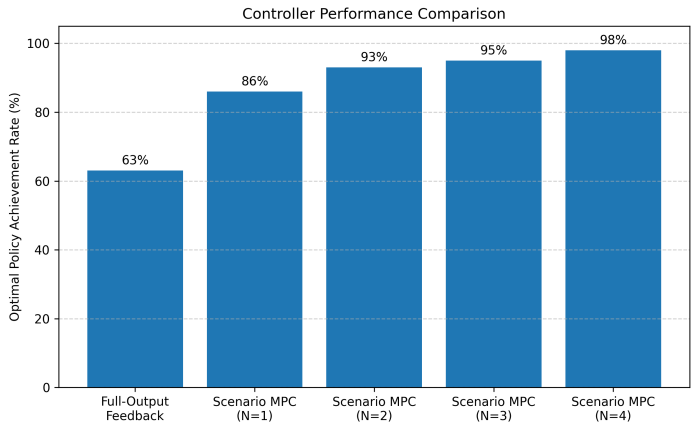


Figure 4.5: Comparison between the full-output feedback control law and scenario MPC.

Conclusions

The central goal of this thesis was to provide a deeper insight into the algebraic and topological structures underlying reinforcement learning methods, and to develop novel RL algorithms grounded in system theoretic principles, leveraging the mathematical structure of the problems they address.

The first major contribution of this work lies in the study of a general class of learning methods that extend the classical two-timescale paradigm to incorporate system measurements. Using tools from singular perturbations theory and LaSalle's invariance principle, a sufficient condition for their convergence was derived. Building on this theoretical framework, two novel actor-critic variants were proposed to accommodate two different types of measurement dynamics. The first involves a measurement unit endowed with knowledge of the system model, and the second is entirely data driven.

The second contribution consists of a thorough analysis of a recently proposed class of reinforcement learning algorithms, called reward-balancing methods. These methods iteratively adjust the reward function to reach a so-called normal form that preserves the same information content of the original reward function, while ensuring that all optimal policies are greedy. A detailed characterization of the transformation group and of the underlying geometry of this process was provided, together with a control-theoretic reformulation of the methodology. Furthermore, an optimal control problem was formulated to provide a principled framework for designing normalizing laws capable of handling constraints and penalty terms, and sufficient conditions guaranteeing its feasibility were established. Finally, the scenario of model uncertainty was addressed, providing an analysis of sample-based methods to handle model mismatches.

Overall, the results of this thesis illustrate that combining mathematical analysis, control theory, and algorithm design provides a powerful framework for developing new reinforcement learning methods. While the work assumes specific structural properties in some cases, it opens several promising directions for future research. Potential extensions could include the application of the singular-perturbation-based results to alternative frameworks, such as average-return or LQR settings. Moreover, the general structural results obtained in Chapter 4 may pave the way for extending reward-balancing methods applied to average-return frameworks, infinite state and action spaces, and nonstationary policies.

In conclusion, this thesis demonstrates that a rigorous, structure-aware approach to reinforcement learning not only deepens the theoretical understanding of existing methods but also enables the principled design of novel algorithms, bridging the gap between theory and practical application.

BIBLIOGRAPHY

[1] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[2] Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Reinforcement learning for humanoid robotics. In *Proceedings of the third IEEE-RAS international conference on humanoid robots*, pages 1–20, 2003.

[3] Yunlong Song, Mats Steinweg, Elia Kaufmann, and Davide Scaramuzza. Autonomous drone racing with deep reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1205–1212. IEEE, 2021.

[4] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

[5] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE transactions on intelligent transportation systems*, 23(6):4909–4926, 2021.

[6] Jianyu Chen, Bodi Yuan, and Masayoshi Tomizuka. Model-free deep reinforcement learning for urban autonomous driving. In *2019 IEEE intelligent transportation systems conference (ITSC)*, pages 2765–2771. IEEE, 2019.

[7] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[8] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[9] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[10] Elise Van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, and Max Welling. Mdp homomorphic networks: Group symmetries in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:4199–4210, 2020.

[11] Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.

[12] Arsenii Mustafin, Aleksei Pakharev, Alex Olshevsky, and Ioannis Ch. Paschalidis. Mdp geometry, normalization and reward balancing solvers, 2025.

[13] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[14] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[15] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1957.

[16] Hassan K Khalil. Nonlinear systems. *Upper Saddle River*, 2002.

[17] Wassim M Haddad and VijaySekhar Chellaboina. *Nonlinear dynamical systems and control: a Lyapunov-based approach*. Princeton university press, 2008.

[18] Petar V Kokotovic, Robert E O'Malley Jr, and Peddapullaiah Sannuti. Singular perturbations and order reduction in control theory—an overview. *Automatica*, 12(2):123–132, 1976.

[19] Petar Kokotović, Hassan K Khalil, and John O'reilly. *Singular perturbation methods in control: analysis and design*. SIAM, 1999.

[20] Eduardus Marie de Jager and JF Furu. *The theory of singular perturbations*, volume 42. Elsevier, 1996.

[21] R Bouyekhf and A El Moudni. On analysis of discrete singularly perturbed non-linear systems: application to the study of stability properties. *Journal of the Franklin Institute*, 334(2):199–212, 1997.

[22] Bakhtiar Litkouhi and Hassan Khalil. Multirate and composite control of two-time-scale discrete-time systems. *IEEE Transactions on Automatic Control*, 30(7):645–651, 1985.

[23] Vivek S Borkar and Vijaymohan R Konda. The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, 22:525–543, 1997.

[24] Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

[25] Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

[26] Harsh Gupta, Rayadurgam Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

[27] Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

[28] Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

[29] John L. Kelley. *General topology*. D. Van Nostrand Company, Inc., Toronto-New York-London, 1955.

[30] Simone Baroncini, Guido Carnevale, Bahman Gharesifard, and Giuseppe Notarstefano. A system-theoretic note on model-based actor-critic. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 1929–1934. IEEE, 2024.

[31] Richard Bellman. Dynamic programming. *science*, 153(3731):34–37, 1966.

[32] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[33] Xin Xu, Lei Zuo, and Zhenhua Huang. Reinforcement learning algorithms with function approximation: Recent advances and applications. *Information sciences*, 261:1–31, 2014.

[34] Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671, 2008.

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

[36] Ivo Grondman, Lucian Busoniu, Gabriel AD Lopes, and Robert Babuska. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, part C (applications and reviews)*, 42(6):1291–1307, 2012.

[37] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[38] John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

[39] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[40] Peter Dayan and CJCH Watkins. Q-learning. *Machine learning*, 8(3):279–292, 1992.

[41] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

[42] Loring W Tu. Manifolds. In *An Introduction to Manifolds*, pages 47–83. Springer, 2011.

[43] John M Lee. Smooth manifolds. In *Introduction to smooth manifolds*, pages 1–29. Springer, 2003.

[44] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[45] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

[46] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

[47] Nicoletta Bof, Ruggero Carli, and Luca Schenato. Lyapunov theory for discrete time systems. *arXiv preprint arXiv:1809.05289*, 2018.

[48] David H Fremlin. Kirszbraun's theorem. *Preprint*, 2011.

[49] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[50] Alex Olshevsky and Bahman Gharesifard. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2):980–1007, 2023.

[51] S Baroncini, B Gharesifard, and G Notarstefano. On reward-balancing methods for reinforcement learning. Preprint, to appear on arXiv, 2026.

[52] Jiin Woo, Gauri Joshi, and Yuejie Chi. The blessing of heterogeneity in federated q-learning: Linear speedup and beyond. *Journal of Machine Learning Research*, 26(26):1–85, 2025.

[53] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer, 1999.

[54] Bhaskara Marthi. Automatic shaping and decomposition of reward functions. In *Proceedings of the 24th International Conference on Machine learning*, pages 601–608, 2007.

[55] Jinsheng Ren, Shangqi Guo, and Feng Chen. Orientation-preserving rewards' balancing in reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6458–6472, 2021.

[56] Haosheng Zou, Tongzheng Ren, Dong Yan, Hang Su, and Jun Zhu. Reward shaping via meta-learning. *arXiv preprint arXiv:1901.09330*, 2019.

[57] Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems*, 33:15931–15941, 2020.

[58] M. Artin. *Algebra*. Pearson Prentice Hall, 2011.

[59] Arthur Earl Bryson. *Applied optimal control: optimization, estimation and control*. Routledge, 2018.

[60] J.M. Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3–5, 1975.

[61] David Williams. *Probability with martingales*. Cambridge university press, 1991.