Mach Learn (2015) 100:255–283
DOI 10.1007/s10994-015-5484-1

Policy gradient in Lipschitz Markov Decision Processes

Matteo Pirotta · Marcello Restelli · Luca Bascetta

Received: 28 September 2014 / Accepted: 6 February 2015 / Published online: 3 March 2015

Abstract This paper is about the exploitation of Lipschitz continuity properties for Markov Decision Processes to safely speed up policy-gradient algorithms. Starting from assumptions about the Lipschitz continuity of the state-transition model, the reward function, and the policies considered in the learning process, we show that both the expected return of a policy and its gradient are Lipschitz continuous w.r.t. policy parameters. By leveraging such properties, we define policy-parameter updates that guarantee a performance improvement at each iteration. The proposed methods are empirically evaluated and compared to other related approaches using different configurations of three popular control scenarios: the linear quadratic regulator, the mass-spring-damper system and the ship-steering control.

Keywords Reinforcement learning · Markov Decision Process · Lipschitz continuity · Policy gradient algorithm

1 Introduction

In the last years, policy-gradient methods have emerged among the most effective Reinforcement-Learning (RL) techniques for complex real-world control problems with continuous, high-dimensional, and partially-observable properties, such as robotic control systems (Peters and Schaal 2006). Given a parameterized policy space, usually designed to incorporate domain knowledge, policy-gradient algorithms update policy parameters along

Editors: Concha Bielza, João Gama, Alípio M. Jorge, and Indrė Žliobaitė.

M. Pirotta (✉) · M. Restelli · L. Bascetta
Department of Electronics, Information, and Bioengineering,
Politecnico di Milano, Piazza Leonardo Da Vinci, 32, I-20133 Milan, Italy
e-mail: matteo.pirotta@polimi.it

M. Restelli
e-mail: marcello.restelli@polimi.it

L. Bascetta
e-mail: luca.bascetta@polimi.it

an estimated ascent direction of the expected return. Under some mild assumptions on the *step size* used to update the parameters (Moré and Thuente 1994), policy-gradient methods are guaranteed to converge at least to a locally optimal solution.

The research in policy gradient has mainly focused on defining convenient ascent directions and low-variance, model-free estimators of the policy gradient. The oldest policy-gradient approaches are finite-difference methods (Spall 1992), that estimate gradient direction by resolving a regression problem based on the performance evaluation of policies associated to different small perturbations of the current parametrization. Finite-difference methods have some advantages: they are easy to implement, do not need assumptions on the differentiability of the policy w.r.t. the policy parameters, and are efficient in deterministic settings. On the other hand, when used on real systems, the choice of parameter perturbations may be difficult and critical for system safeness. Furthermore, the presence of uncertainties may significantly slow down the convergence rate. Such drawbacks have been overcome by likelihood ratio methods (Williams 1992; Baxter and Bartlett 2001; Sutton et al. 1999), since they do not need to generate policy parameter variations and quickly converge even in highly stochastic systems. Several studies have addressed the problem to find minimum variance estimators by the computation of optimal baselines (Peters and Schaal 2008b). To further improve the efficiency of policy-gradient methods, natural-gradient approaches (where the steepest ascent is computed w.r.t. the Fisher information metric) have been considered (Kakade 2001; Peters and Schaal 2008a). Natural gradients still converge to locally optimal policies, are independent from the policy parametrization, need less data to attain good gradient estimates, and are less affected by plateaus. For recent and comprehensive surveys on policy search and policy gradient methods we refer the reader to Grondman et al. (2012) and Deisenroth et al. (2013).

Unfortunately, a good estimate of the policy gradient is not enough to guarantee effective learning. In fact, even when the exact policy gradient is known, the choice of the step size strongly influences the number of iterations needed to attain a (local) maximum or, even worse, can make convergence unfeasible (Wagner 2011). In general unconstrained programming, the value of the step size is determined through line-search algorithms (Moré and Thuente 1994), that require to evaluate the function to be optimized at points generated along the gradient direction by a sequence of candidate values for the step size. In the policy-gradient framework, being policy evaluations quite expensive, line search is impractical and step-size parameters are usually kept fixed or decreased over time according to some annealing schedule, requiring significant amounts of hand tuning. Convergence issues can be solved by making the step-size parameter decrease according to the Robbins-Monro conditions (Robbins and Monro 1951), but it usually turns out to show very slow convergence.

In spite of the strong impact of the step size over the performance of policy-gradient methods, so far little research has addressed such issue, with a few notable exceptions. Kober and Peters (2008) and Vlassis et al. (2009) studied policy-search methods based on expectation-maximization. Under some assumptions on the reward and policy models, expectation-maximization algorithms have properties similar to the ones of policy gradients, but without the need of specifying any step size. In Pirotta et al. (2013) we have directly addressed the problem of computing a step size that guarantees a policy improvement at any iteration. The idea is to use the data collected using the current policy to lower bound the expected return of any policy. The step size is then chosen to maximize such lower bound along the policy-gradient direction.

The main limitation of previous approaches is the looseness of the lower bounds to the expected return, that usually leads to conservative policy updates. In order to mitigate this drawback, we focus on Lipschitz-continuous MDPs, that represent a relevant subclass of

MDPs. In fact, many real-world problems are characterized by continuous state and action spaces (e.g., robotics, automatic control problems, natural resource management, etc.), where it is reasonable that when similar actions are executed in similar states their effects will be similar. That is what the Lipschitz assumptions want to capture. In this paper, we show that, under Lipschitz continuity assumptions on the Markov Decision Process (MDP) and the policy model (Sect. 2), the expected return of each policy and its policy-gradient components are Lipschitz w.r.t. policy parameters (Sects. 3 and 4). As shown by Armijo (1966), the Lipschitz continuity of the gradient can be used to select the value of the step size so as to guarantee a performance improvement at each iteration. In particular, the smaller are the Lipschitz constants, the larger are the step sizes and the expected improvements. We introduce how to compute the Lipschitz constant related to each component of the gradient and we show how such constants can be used to guarantee a performance improvement either by automatically identifying a proper step size along the gradient direction, or by defining new ascent directions with better guarantees (Sect. 5). Besides the theoretical contributions, we will also provide an empirical analysis to highlight advantages and limitations of the proposed approach (Sect. 6).

2 Preliminaries

In this section, we introduce notation and basic concepts about MDPs, Lipschitz MDPs, and policy gradients.

2.1 Markov Decision Process

A discrete-time continuous MDP is defined as a 6-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu \rangle$, where $\mathcal{S}$ is the continuous state space, $\mathcal{A}$ is the continuous action space, $\mathcal{P}$ is a Markovian transition model where $\mathcal{P}(s'|s,a)$ defines the transition density between state s and s' under action a, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [-R, R]$ is the reward function, such that $\mathcal{R}(s,a)$ is the expected immediate reward for the state-action pair (s,a) and R is the maximum absolute reward value, $\gamma \in [0,1)$ is the discount factor for future rewards, and μ is the initial state distribution. We assume state and action spaces to be complete, separable metric (Polish) spaces $(\mathcal{S}, d_{\mathcal{S}})$ and $(\mathcal{A}, d_{\mathcal{A}})$, equipped with their σ-algebras $\sigma_{\mathcal{S}}$, $\sigma_{\mathcal{A}}$ of Borel sets, respectively. We assume—as done in Hinderer (2005)—that joint state-action space is endowed with the following taxicab norm: $d_{\mathcal{SA}}\left((s,a),(\widehat{s},\widehat{a})\right) = d_{\mathcal{S}}(s,\widehat{s}) + d_{\mathcal{A}}(a,\widehat{a})$. A stationary policy $\pi(\cdot|s)$ specifies for each state s the density function over the Borel action space $(\mathcal{A}, d_{\mathcal{A}}, \sigma_{\mathcal{A}})$.

We consider infinite-horizon problems where the future rewards are exponentially discounted with γ. For each state s, we define the utility of following a stationary policy π as:

$$V^{\pi}(s) = \mathop{\mathbb{E}}_{\substack{a_t \sim \pi \\ s_t \sim \mathcal{P}}} \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) | s_0 = s \right].$$

It is known that, under mild assumptions (Bertsekas and Shreve 1978), V^{π} solves the following recursive (Bellman) equation:

$$V^{\pi}(s) = \int_{\mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \int_{\mathcal{S}} V^{\pi}(s') P(\mathrm{d}s'|s,a) \right) \pi(\mathrm{d}a|s).$$

For model-free control purposes, the value function V is usually replaced by the action-value function Q, where action value $Q^{\pi}(s,a)$ is the expected return of taking action a in state s and following a policy π thereafter:

$$Q^\pi(s,a) = \mathcal{R}(s,a) + \gamma \iint_{\mathcal{SA}} Q^\pi(s',a')\pi(\mathrm{d}a'|s')\mathcal{P}(\mathrm{d}s'|s,a).$$

Policies can be ranked by their expected discounted reward starting from the state distribution μ:

$$J_\mu^\pi = \int_{\mathcal{S}} V^\pi(s)\mu(\mathrm{d}s) = \frac{1}{1-\gamma}\iint_{\mathcal{SA}} \mathcal{R}(s,a)\pi(\mathrm{d}a|s)\delta_\mu^\pi(\mathrm{d}s)$$
$$= \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\delta_\mu^\pi(\cdot)}\,\mathbb{E}_{a\sim\pi(\cdot|s)}\left[\mathcal{R}(s,a)\right],$$

where $\delta_\mu^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr(s_t = s|\pi,\mu)$ is the γ-discounted future state distribution for a starting state distribution μ (Sutton et al. 1999). It is possible to rewrite previous equation in terms of the joint distribution $\zeta(\delta_\mu^\pi,\pi)$ between the future state distribution δ_μ^π and the stationary policy π, that can be written as a function of only π. Let $\zeta(\delta_\mu^\pi,\pi) = \zeta_\mu^\pi$ be a probability distribution over $\mathcal{S}\times\mathcal{A}$, such that:

$$J_\mu^\pi = \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim\zeta_\mu^\pi}\left[\mathcal{R}(s,a)\right].$$

Solving an MDP means finding a policy π^* that maximizes the expected long-term reward: $\pi^* \in arg\max_{\pi\in\Pi} J_\mu^\pi$. For any MDP there exists at least one deterministic stationary optimal policy that simultaneously maximizes $V^\pi(s)$, $\forall s \in \mathcal{S}$ (Puterman 1994).

2.2 Lipschitz MDP

In this section, we introduce the basic concepts of Lipschitz continuity. Given two metric sets (X, d_X) and (Y, d_Y), where d_X and d_Y denote the corresponding metric functions, a function $f: X \to Y$ is called L_f-Lipschitz continuous (L_f-LC) if

$$\forall(x_1,x_2)\in X^2,\quad d_Y\left(f(x_1),f(x_2)\right) \le L_f d_X\left(x_1,x_2\right). \tag{1}$$

The smallest constant L_f for which (1) holds is called the Lipschitz constant of f. Define $\|f\|_L = \sup_{x_1\neq x_2}\left\{\frac{d_Y(f(x_1),f(x_2))}{d_X(x_1,x_2)} : x_1,x_2\in X\right\}$ to be the Lipschitz semi-norm over the function space $\mathcal{F}(X,Y)$. Furthermore, we call f pointwise Lipschitz continuous [1] (PLC) in state x if there exists a constant $L_f(x)$ such that:

$$\forall x'\in X, d_Y\left(f(x),f(x')\right) \le L_f(x) d_X\left(x,x'\right) \text{ where } \forall x\in X, L_f(x)\le L_f.$$

For real-valued functions (e.g., the reward function), we will use the Euclidean distance as metric for the codomain. On the other hand, for the state-transition model and the policies we need to introduce a distance between probability distributions. Following Hinderer (2005) and Rachelson and Lagoudakis (2010), we will consider the Kantorovich or L^1-Wasserstein metric on probability measures p and q:

$$\mathcal{K}(p,q) = \sup_f\left\{\left|\int_X f\,\mathrm{d}(p-q)\right| : \|f\|_L \le 1\right\}. \tag{2}$$

We decided to use this metric, instead of other more common and easier metrics, like the Total Variation (TV) one, because it is "less demanding", that is, MDPs that are Lipschitz according

[1] Notice that our definition of pointwise Lipschitz function differs from the traditional one.

to TV are always Lipschitz also w.r.t. the L^1-Wasserstein metric, while the vice versa is not true. For instance, MDPs with deterministic transitions are never Lipschitz according to TV, while they can be Lipschitz using L^1-Wasserstein metric. Finally, the choice of the L^1-Wasserstein metric rather than other sophisticated distribution distances is motivated by the fact that it has been frequently used for MDPs (Rachelson and Lagoudakis 2010; Hinderer 2005; Ferns et al. 2005).

The analysis proposed in this paper is based on the assumption that the MDP is Lipschitz continuous. A Lipschitz MDP is a standard MDP enhanced by the information that the transition model is $L_{\mathcal{P}}$-LC, and the reward model is $L_{\mathcal{R}}$-LC.

Assumption 1 (Lipschitz MDP) A Lipschitz MDP must satisfy the following conditions:

$$\forall\,(s,\widehat{s},a,\widehat{a}) \in \mathcal{S}^2 \times \mathcal{A}^2, \qquad \mathcal{K}\left(\mathcal{P}\left(\cdot|s,a\right),\mathcal{P}\left(\cdot|\widehat{s},\widehat{a}\right)\right) \leq L_{\mathcal{P}}\, d_{\mathcal{SA}}\left((s,a),(\widehat{s},\widehat{a})\right),$$
$$\forall\,(s,\widehat{s},a,\widehat{a}) \in \mathcal{S}^2 \times \mathcal{A}^2, \qquad \left|\mathcal{R}(s,a) - \mathcal{R}(\widehat{s},\widehat{a})\right| \leq L_{\mathcal{R}}\, d_{\mathcal{SA}}\left((s,a),(\widehat{s},\widehat{a})\right).$$

If π is an L_π-LC policy—$\forall(s,\widehat{s}) \in \mathcal{S}^2, \mathcal{K}\left(\pi(\cdot|s),\pi(\cdot|\widehat{s})\right) \leq L_\pi\, d_{\mathcal{S}}\left(s,\widehat{s}\right)$—, under Assumption 1, it is possible to prove the LC of the corresponding value functions.

Lemma 1 *(Rachelson and Lagoudakis 2010, Lemma 1, Theorem 1) Given an $(L_{\mathcal{P}}, L_{\mathcal{R}})$-LC MDP and a L_π-LC stationary policy π, if $\gamma L_{\mathcal{P}}\,(1+L_\pi) < 1$, then the Q-function Q^π is L_{Q^π}-LC and the V-function is L_{V^π}-LC w.r.t. the joint state-action space:*

$$L_{Q^\pi} = \frac{L_{\mathcal{R}}}{1-\gamma L_{\mathcal{P}}\,(1+L_\pi)}; \qquad L_{V^\pi} = L_{Q^\pi}\,(1+L_\pi).$$

All these conditions are related to state and action variables[2]. In particular, the Lipschitz continuity of the V- and Q-functions means that: $\forall(s,\widehat{s},a,\widehat{a}) \in \mathcal{S}^2 \times \mathcal{A}^2$

$$\left|Q^\pi(s,a) - Q^\pi(\widehat{s},\widehat{a})\right| \leq L_{Q^\pi}\, d_{\mathcal{SA}}\left((s,a),(\widehat{s},\widehat{a})\right);$$
$$\left|V^\pi(s) - V^\pi(\widehat{s})\right| \leq L_{V^\pi}\, d_{\mathcal{S}}\left(s,\widehat{s}\right).$$

In the following, we will consider the Lipschitz continuity related to the policy parametrization.

2.3 Policy space

We consider the problem of finding a policy that maximizes the expected discounted reward over a class of parameterized policies $\Pi_\Theta = \left\{\pi^{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\right\}$, where $\pi^{\boldsymbol{\theta}}$ is a compact representation of $\pi^{\boldsymbol{\theta}}(a|s)$. Moreover, we assume that (Θ, d_Θ) is a metric space. For ease of notation, in the following we will use $\boldsymbol{\theta}$ to denote the dependence on $\pi^{\boldsymbol{\theta}}$ where possible.

The exact gradient of the expected discounted reward $J_\mu^{\boldsymbol{\theta}}$ w.r.t. the policy parameters is (Sutton et al. 1999):

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}} &= \tfrac{1}{1-\gamma} \iint_{\mathcal{SA}} \nabla_{\boldsymbol{\theta}} \log \pi^{\boldsymbol{\theta}}(a|s)\, Q^{\boldsymbol{\theta}}(s,a)\, \pi^{\boldsymbol{\theta}}(\mathrm{d}a|s)\, d_\mu^{\boldsymbol{\theta}}(\mathrm{d}s) \\ &= \tfrac{1}{1-\gamma}\, \mathbb{E}_{(s,a)\sim\zeta_\mu^{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}} \log \pi^{\boldsymbol{\theta}}(a|s)\, Q^{\boldsymbol{\theta}}(s,a)\right]. \end{aligned} \tag{3}$$

[2] It can be noticed that the results in Lemma 1 are quite similar to the ones obtained by Hinderer (2005, Theorem 4.1a). The main difference is due to the fact that Hinderer focuses on the Lipschitz continuity of the optimal value function under state-dependent action spaces. In this case, the role of the Lipschitz constant L_π is taken by the Lipschitz constant due to state-dependent action spaces. Since we consider policy-based value functions, the eventual differences between action spaces is implicitly coded in the policy.

Several studies have focused on computing the value of this gradient from sample trajectories, trying to produce estimators with low variance (Peters and Schaal 2008b). The policy parameters can be updated by following the direction of the gradient of the expected discounted reward: $\boldsymbol{\theta}' = \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}}$, where α is a parameter used to control the step size.

For proving the results in the next section, we need to introduce the following assumptions on the parameterized policy model.

Assumption 2 (Lipschitz policies) The policy model must satisfy the following conditions:

1) state-action LC: $\forall (s, \widehat{s}) \in \mathcal{S}^2, \mathcal{K}\left(\pi^{\boldsymbol{\theta}}(\cdot|s), \pi^{\boldsymbol{\theta}}(\cdot|\widehat{s})\right) \leq L_{\pi^{\boldsymbol{\theta}}}\, d_{\mathcal{S}}(s, \widehat{s})$
2) parametric PLC: $\forall s \in \mathcal{S},\ \forall (\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \in \Theta^2, \mathcal{K}\left(\pi^{\boldsymbol{\theta}}(\cdot|s), \pi^{\widehat{\boldsymbol{\theta}}}(\cdot|s)\right) \leq L_{\pi}(\boldsymbol{\theta})\, d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right)$

Assumption 3 (Lipschitz gradient of policy logarithm) The gradient of the policy logarithm must satisfy the following conditions of:

1) uniformly bounded gradient: $\forall (s, a) \in \mathcal{S} \times \mathcal{A},\ \forall \boldsymbol{\theta} \in \Theta,\ \forall i = 1, \ldots, d$

$$\left|\nabla_{\boldsymbol{\theta}_i} \log \pi^{\boldsymbol{\theta}}(a|s)\right| \leq M_{\boldsymbol{\theta}}^{i}$$

2) state-action LC: $\forall (s, \widehat{s}, a, \widehat{a}) \in \mathcal{S}^2 \times \mathcal{A}^2,\ \forall \boldsymbol{\theta} \in \Theta,\ \forall i = 1, \ldots, d$

$$\left|\nabla_{\boldsymbol{\theta}_i} \log \pi^{\boldsymbol{\theta}}(a|s) - \nabla_{\boldsymbol{\theta}_i} \log \pi^{\boldsymbol{\theta}}(\widehat{a}|\widehat{s})\right| \leq L^{i}_{\nabla \log \pi^{\boldsymbol{\theta}}}\, d_{\mathcal{SA}}((s, a), (\widehat{s}, \widehat{a}))$$

3) parametric PLC: $\forall (\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \in \Theta^2,\ \forall (s, a) \in \mathcal{S} \times \mathcal{A},\ \forall i = 1, \ldots, d$

$$\left|\nabla_{\boldsymbol{\theta}_i} \log \pi^{\boldsymbol{\theta}}(a|s) - \nabla_{\boldsymbol{\theta}_i} \log \pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right| \leq L^{i}_{\nabla \log \pi}(\boldsymbol{\theta})\, d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right)$$

Notice that some previously defined Lipschitz constants become $\boldsymbol{\theta}$-dependent when a parametric policy model is used, i.e., $L_{\pi^{\boldsymbol{\theta}}}$, $L_{Q^{\boldsymbol{\theta}}}$ and $L_{V^{\boldsymbol{\theta}}}$, whereas $L_{\mathcal{P}}$ and $L_{\mathcal{R}}$ are $\boldsymbol{\theta}$-independent since they are not affected by the policy.

3 Lipschitz continuity of the expected return

In this section, we will show that under Assumptions 1, 2, and 3 the expected return $J_{\mu}^{\boldsymbol{\theta}}$ is a Lipschitz function w.r.t. policy parameters $\boldsymbol{\theta}$. Besides being an interesting objective by itself, it allows us to introduce some preliminary results that will be reused in the next section to make the proof of the Lipschitz property of the policy gradient easier.

The performance distance between two policies corresponding to parameters $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$ is measured by the absolute difference of their expected returns:

$$\left|J_{\mu}^{\boldsymbol{\theta}} - J_{\mu}^{\widehat{\boldsymbol{\theta}}}\right| = \frac{1}{1-\gamma}\left|\mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\boldsymbol{\theta}}}[\mathcal{R}(s, a)] - \mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}}[\mathcal{R}(s, a)]\right|.$$

If the Lipschitz constant of the reward function $\mathcal{R}$ were less or equal to 1 (i.e., $\|\mathcal{R}\|_L \leq 1$), it follows from (2) that the performance distance between policies $\pi^{\boldsymbol{\theta}}$ and $\pi^{\widehat{\boldsymbol{\theta}}}$ would be upper bounded by the Kantorovich distance between the distributions $\zeta_{\mu}^{\boldsymbol{\theta}}$ and $\zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}$. On the other hand, it can be easily shown that if $\mathcal{R}$ is $L_{\mathcal{R}}$-LC then $\left\|\frac{\mathcal{R}}{L_{\mathcal{R}}}\right\|_L \leq 1$.

The following proposition gives an upper bound to absolute difference in performance between policies. Proof can be founded in "Proof of Proposition 1" section of Appendix.

Proposition 1 *Given an $L_{\mathcal{R}}$-LC MDP, for any pair of stationary policies corresponding to parameters $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, the absolute difference between the performance of policy $\pi^{\boldsymbol{\theta}}$ and policy $\pi^{\widehat{\boldsymbol{\theta}}}$ can be bounded as follows:*

$$\left|J_{\mu}^{\boldsymbol{\theta}} - J_{\mu}^{\widehat{\boldsymbol{\theta}}}\right| \leq \frac{L_{\mathcal{R}}}{1-\gamma}\, \mathcal{K}\left(\zeta_{\mu}^{\boldsymbol{\theta}}, \zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right).$$

As a consequence, to prove the Lipschitz continuity of the expected return w.r.t. $\boldsymbol{\theta}$ it suffices to show the Lipschitz continuity of the distribution $\zeta_{\mu}^{\boldsymbol{\theta}}$ w.r.t. $\boldsymbol{\theta}$. It is worth recalling that the distribution $\zeta_{\mu}^{\boldsymbol{\theta}}$ is defined over the joint state-action space and the probability of drawing a state-action pair (s, a) is $\delta_{\mu}^{\boldsymbol{\theta}}(s) \cdot \pi^{\boldsymbol{\theta}}(a|s)$. As it can be noticed, the probability distribution over actions and the one over states are not independent. This means that the Kantorovich distance of the joint distribution cannot be simply upper bounded by the sum of the Kantorovich distances of the γ-discounted future state distribution δ_{μ} and the policy π. The following Lemma gives an upper bound to $\mathcal{K}\left(\zeta_{\mu}^{\boldsymbol{\theta}}, \zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right)$.

Lemma 2 *Given an $L_{\pi^{\boldsymbol{\theta}}}$-LC and $L_{\pi}(\boldsymbol{\theta})$-PLC stationary policy $\pi^{\boldsymbol{\theta}}$, the Kantorovich distance between a pair of joint distributions $\zeta_{\mu}^{\boldsymbol{\theta}}$ and $\zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}$ is bounded by:*

$$\mathcal{K}\left(\zeta_{\mu}^{\boldsymbol{\theta}}, \zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right) \leq L_{\pi}(\boldsymbol{\theta})\; d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) + \left(1 + L_{\pi^{\boldsymbol{\theta}}}\right)\; \mathcal{K}\left(\delta_{\mu}^{\boldsymbol{\theta}}, \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right).$$

Proof The proof is divided into two parts. The first part is devoted to the analysis of the Lipschitz continuity of a term involved in the definition of the L^1-Wasserstein metric between the joint distributions. The second part exploits this result to prove the lemma.

Define $b_f^{\boldsymbol{\theta}}(s) = \mathbb{E}_{a\sim\pi^{\boldsymbol{\theta}}} f(s, a)$, where the function f is 1-LC w.r.t. the joint state-action space. Given an $L_{\pi^{\boldsymbol{\theta}}}$-LC policy model, $b(s)$ is Lipschitz continuous:

$$\begin{aligned}
\left|b_f^{\boldsymbol{\theta}}(s) - b_f^{\boldsymbol{\theta}}(\widehat{s})\right| &= \left|\int_{\mathcal{A}} \pi^{\boldsymbol{\theta}}(a|s) f(s,a) - \pi^{\boldsymbol{\theta}}(a|\widehat{s}) f(\widehat{s},a)\mathrm{d}a\right| \\
&= \left|\int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\boldsymbol{\theta}}(a|\widehat{s})\right) f(s,a) + \left(f(s,a) - f(\widehat{s},a)\right)\pi(a|\widehat{s})\mathrm{d}a\right| \\
&\leq \left|\int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\boldsymbol{\theta}}(a|\widehat{s})\right) f(s,a)\mathrm{d}a\right| + \left|\int_{\mathcal{A}} \left(1 \cdot d_{\mathcal{S}}(s,\widehat{s})\right)\pi(a|\widehat{s})\mathrm{d}a\right| \\
&\leq \mathcal{K}\left(\pi^{\boldsymbol{\theta}}(\cdot|s), \pi^{\boldsymbol{\theta}}(\cdot|\widehat{s})\right) + d_{\mathcal{S}}(s,\widehat{s}) \\
&\leq \left(L_{\pi^{\boldsymbol{\theta}}} + 1\right) d_{\mathcal{S}}(s,\widehat{s}). \qquad (4)
\end{aligned}$$

Recall that the function f in the definition of the L^1-Wasserstein metric for the joint distributions is 1-LC w.r.t. every pair (s, a), but as a consequence it is, at most, 1-LC for the single variables s and a. The proof follows from the previous result and some algebraic manipulations:

$$\begin{aligned}
\mathcal{K}\left(\zeta_{\mu}^{\boldsymbol{\theta}}, \zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right) &= \sup_f \left\{\left|\int_{\mathcal{S}} \delta_{\mu}^{\boldsymbol{\theta}}(s) \int_{\mathcal{A}} \pi^{\boldsymbol{\theta}}(a|s) f(s,a)\mathrm{d}a\mathrm{d}s\right.\right. \\
&\quad \left.\left. - \int_{\mathcal{S}} \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \pi^{\widehat{\boldsymbol{\theta}}}(a|s) f(s,a)\mathrm{d}a\mathrm{d}s\right| : \|f\|_L \leq 1\right\} \qquad (5) \\
&= \sup_f \left\{\left|\int_{\mathcal{S}} \left(\delta_{\mu}^{\boldsymbol{\theta}}(s) - \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s)\right) \int_{\mathcal{A}} \pi^{\boldsymbol{\theta}}(a|s) f(s,a)\mathrm{d}a\mathrm{d}s\right.\right.
\end{aligned}$$

$$+\int_{\mathcal{S}} \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right) f(s,a)\mathrm{d}a\mathrm{d}s \Bigg| : \|f\|_L \leq 1 \Bigg\} \tag{6}$$

$$\leq \sup_f \left\{ \left| \int_{\mathcal{S}} \left(\delta_{\mu}^{\boldsymbol{\theta}}(s) - \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s)\right) b_f^{\boldsymbol{\theta}}(s)\mathrm{d}s \right| : \|f\|_L \leq 1 \right\}$$

$$+\sup_f \left\{ \left| \int_{\mathcal{S}} \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right) f(s,a)\mathrm{d}a\mathrm{d}s \right| : \|f\|_L \leq 1 \right\} \tag{7}$$

$$= \left(L_{\pi^{\boldsymbol{\theta}}} + 1\right) \sup_f \left\{ \left| \int_{\mathcal{S}} \left(\delta_{\mu}^{\boldsymbol{\theta}}(s) - \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s)\right) \frac{b_f^{\boldsymbol{\theta}}(s)}{\left(L_{\pi^{\boldsymbol{\theta}}} + 1\right)}\mathrm{d}s \right| : \|f\|_L \leq 1 \right\}$$

$$+\sup_f \left\{ \left| \int_{\mathcal{S}} \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right) f(s,a)\mathrm{d}a\mathrm{d}s \right| : \|f\|_L \leq 1 \right\} \tag{8}$$

$$\leq \left(L_{\pi^{\boldsymbol{\theta}}} + 1\right) \sup_g \left\{ \left| \int_{\mathcal{S}} \left(\delta_{\mu}^{\boldsymbol{\theta}}(s) - \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s)\right) g(s)\mathrm{d}s \right| : \|g\|_L \leq 1 \right\}$$

$$+\int_{\mathcal{S}} \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s) \sup_f \left\{ \left| \int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right) f(s,a)\mathrm{d}a \right| : \|f\|_L \leq 1 \right\} \mathrm{d}s \tag{9}$$

$$\leq \left(L_{\pi^{\boldsymbol{\theta}}} + 1\right) \mathcal{K}\left(\delta_{\mu}^{\boldsymbol{\theta}}, \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right) + \sup_s \mathcal{K}\left(\pi^{\boldsymbol{\theta}}(\cdot|s), \pi^{\widehat{\boldsymbol{\theta}}}(\cdot|s)\right)$$

$$\leq \left(L_{\pi^{\boldsymbol{\theta}}} + 1\right) \mathcal{K}\left(\delta_{\mu}^{\boldsymbol{\theta}}, \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right) + L_{\pi^{\boldsymbol{\theta}}}(\boldsymbol{\theta})\, d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right). \tag{10}$$

Equality (6) is obtained by manipulation of (5) after insertion of the quantity $\pm \int_{\mathcal{S}} \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s) \int_{\mathcal{A}} \pi^{\boldsymbol{\theta}}(a|s) f(s,a)\mathrm{d}a\mathrm{d}s$. Eq. (8) is obtained by exploiting definition of $b_f^{\boldsymbol{\theta}}(s)$ and adding the (identity) factor $\frac{L_{\pi^{\boldsymbol{\theta}}}+1}{L_{\pi^{\boldsymbol{\theta}}}+1}$. In (9) we rename $\frac{b_f^{\boldsymbol{\theta}}(s)}{L_{\pi^{\boldsymbol{\theta}}}+1}$ to $g(s)$ and we note that, according to Eq. (4), $\|g(s)\|_L = \left\| \frac{b_f^{\boldsymbol{\theta}}(s)}{L_{\pi^{\boldsymbol{\theta}}}+1} \right\|_L \leq 1$. By noting that $\delta_{\mu}^{\widehat{\boldsymbol{\theta}}}$ is always positive, a valid upper bound to the second term in (8) is obtained by pushing the supreme over function space into the state integral. Finally, definition of the Kantorovich distance is used to obtain inequality (10) together with a maximization over the state space. The proof follows from Assumption 2. □

The first term of the upper bound derives from the bound on the Kantorovich distance between policies w.r.t. parameters $\boldsymbol{\theta}$. The second term involves the Kantorovich distance between γ-discounted future state distributions w.r.t. parameters $\boldsymbol{\theta}$ and the factor $(1 + L_{\pi^{\boldsymbol{\theta}}})$ accounts for the dependence between the distribution over the actions and the one over the states: the larger is the $L_{\pi^{\boldsymbol{\theta}}}$ constant the stronger is the dependence between $\pi^{\boldsymbol{\theta}}$ and $\delta_{\mu}^{\boldsymbol{\theta}}$. As expected, when the policy does not depend on the state (i.e., $L_{\pi^{\boldsymbol{\theta}}} = 0$), the bound reduces to the sum of the two Kantorovich distances. The following lemma shows that under Assumptions 1 and 2 also $\mathcal{K}\left(\delta_{\mu}^{\boldsymbol{\theta}}, \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right)$ is Lipschitz w.r.t. $\boldsymbol{\theta}$.

Lemma 3 *Given an $L_{\mathcal{P}}$-LC MDP and an $(L_{\pi^{\boldsymbol{\theta}}}, L_{\pi}(\boldsymbol{\theta}))$-LC stationary policy model, if $\gamma L_{\mathcal{P}}\left(1 + L_{\pi^{\boldsymbol{\theta}}}\right) < 1$, then the Kantorovich distance between a pair of γ-discounted future-state distributions is PLC w.r.t. paramters $\boldsymbol{\theta}$: $\forall(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \in \Theta^2$,*

$$\mathcal{K}\left(\delta_{\mu}^{\boldsymbol{\theta}}, \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right) \leq L_{\delta}(\boldsymbol{\theta})\, d_{\Theta}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}), \quad \textit{where} \quad L_{\delta}(\boldsymbol{\theta}) = \frac{\gamma L_{\mathcal{P}} L_{\pi}(\boldsymbol{\theta})}{1 - \gamma L_{\mathcal{P}}\left(1 + L_{\pi^{\boldsymbol{\theta}}}\right)}.$$

Proof We start the proof with some preliminary results that will be used in the rest of the proof. Let the function $g_f(s,a) = \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} f(s')$ where $\|f\|_L \leq 1$. Then, $g_f(s,a)$ is LC

w.r.t. the action variable:

$$\left|g_f(s,a) - g_f(s,\widehat{a})\right| = \left|\int_{\mathcal{S}} \left(\mathcal{P}(s'|s,a) - \mathcal{P}(s'|s,\widehat{a})\right) f(s')\mathrm{d}s'\right| \leq \mathcal{K}\left(\mathcal{P}\left(\cdot|s,a\right), \mathcal{P}\left(\cdot|s,\widehat{a}\right)\right) \leq L_{\mathcal{P}}\, d_{\mathcal{A}}\left(a,\widehat{a}\right). \tag{11}$$

The second result involves the expectation of $g_f(s,a)$ w.r.t. policy $\pi^{\boldsymbol{\theta}}$. Let $h_f^{\boldsymbol{\theta}}(s) = \mathbb{E}_{a\sim\pi^{\boldsymbol{\theta}}(\cdot|s)}\, g_f(s,a)$, we can prove that it is Lipschitz continuous:

$$\begin{aligned}
\left|h_f^{\boldsymbol{\theta}}(s) - h_f^{\boldsymbol{\theta}}(\widehat{s})\right| &= \left|\iint_{\mathcal{AS}} f(s')\left(\pi^{\boldsymbol{\theta}}(a|s)\mathcal{P}\left(s'|s,a\right) - \pi^{\boldsymbol{\theta}}(a|\widehat{s})\mathcal{P}\left(s'|\widehat{s},a\right)\right)\mathrm{d}s'\mathrm{d}a\right| \\
&= \left|L_{\mathcal{P}}\int_{\mathcal{A}}\left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\boldsymbol{\theta}}(a|\widehat{s})\right)\int_{\mathcal{S}}\frac{\mathcal{P}\left(s'|s,a\right)}{L_{\mathcal{P}}} f(s')\mathrm{d}s'\mathrm{d}a\right. && (12)\\
&\quad \left. + \int_{\mathcal{A}}\pi^{\boldsymbol{\theta}}(a|\widehat{s})\int_{\mathcal{S}}\mathcal{P}\left(s'|s,a\right) f(s') - \mathcal{P}\left(s'|\widehat{s},a\right) f(s')\mathrm{d}s'\mathrm{d}a\right| && (13)\\
&\leq L_{\mathcal{P}}\,\mathcal{K}(\pi^{\boldsymbol{\theta}}(\cdot|s), \pi^{\boldsymbol{\theta}}(\cdot|\widehat{s})) + \sup_a \mathcal{K}(\mathcal{P}(\cdot|s,a), \mathcal{P}(\cdot|\widehat{s},a)) && (14)\\
&\leq L_{\mathcal{P}}\left(L_{\pi^{\boldsymbol{\theta}}} + 1\right)\, d_{\mathcal{S}}\left(s,\widehat{s}\right), && (15)
\end{aligned}$$

where (13) is obtained by adding and subtracting the term $\iint_{\mathcal{AS}} \pi^{\boldsymbol{\theta}}(a|\widehat{s})\mathcal{P}\left(s'|s,a\right)\mathrm{d}s'\mathrm{d}a$. Inequality (14) follows from Kantorovich distance and bound (11), that is $\left\|\frac{\mathbb{E}_{s'\in\mathcal{P}}\, f(s')}{L_{\mathcal{P}}}\right\|_L \leq 1$, given that f is 1-LC. Then

$$\begin{aligned}
\mathcal{K}\left(\delta_\mu^{\boldsymbol{\theta}}, \delta_\mu^{\widehat{\boldsymbol{\theta}}}\right) &= \sup_f\left\{\left|\int_{\mathcal{S}}\left(\delta_\mu^{\boldsymbol{\theta}}(s) - \delta_\mu^{\widehat{\boldsymbol{\theta}}}(s)\right) f(s)\mathrm{d}s\right| : \|f\|_L \leq 1\right\} \\
&= \sup_f\left\{\left|\int_{\mathcal{S}}\left(\mu(s) + \gamma\int_{\mathcal{S}}\int_{\mathcal{A}}\pi^{\boldsymbol{\theta}}(a|s')\mathcal{P}(s|s',a)\delta_\mu^{\boldsymbol{\theta}}(s')\mathrm{d}a\mathrm{d}s'\right) f(s)\right. \\
&\quad \left. - \left(\mu(s) + \gamma\int_{\mathcal{S}}\int_{\mathcal{A}}\pi^{\widehat{\boldsymbol{\theta}}}(a|s')\mathcal{P}(s|s',a)\delta_\mu^{\widehat{\boldsymbol{\theta}}}(s')\mathrm{d}a\mathrm{d}s'\right) f(s)\mathrm{d}s\right| : \|f\|_L \leq 1\right\} && (16)\\
&= \gamma\sup_f\left\{\left|\int_{\mathcal{S}} f(s)\int_{\mathcal{S}}\int_{\mathcal{A}}\mathcal{P}(s|s',a)\left(\pi^{\boldsymbol{\theta}}(a|s')\delta_\mu^{\boldsymbol{\theta}}(s') - \pi^{\widehat{\boldsymbol{\theta}}}(a|s')\delta_\mu^{\widehat{\boldsymbol{\theta}}}(s')\right)\right. \\
&\quad \left.\times\mathrm{d}a\mathrm{d}s'\mathrm{d}s\right| : \|f\|_L \leq 1\Bigg\} \\
&= \gamma\sup_f\left\{\left|\int_{\mathcal{S}} f(s)\int_{\mathcal{S}}\int_{\mathcal{A}}\mathcal{P}(s|s',a)\left(\left(\pi^{\boldsymbol{\theta}}(a|s') - \pi^{\widehat{\boldsymbol{\theta}}}(a|s')\right)\delta_\mu^{\widehat{\boldsymbol{\theta}}}(s')\right.\right. \\
&\quad \left.\left. + \left(\delta_\mu^{\boldsymbol{\theta}}(s') - \delta_\mu^{\widehat{\boldsymbol{\theta}}}(s')\right)\pi^{\boldsymbol{\theta}}(a|s')\right)\mathrm{d}a\mathrm{d}s'\mathrm{d}s\right| : \|f\|_L \leq 1\Bigg\} \\
&\leq \gamma\sup_f\left\{\left|\int_{\mathcal{S}}\delta_\mu^{\widehat{\boldsymbol{\theta}}}(s')\int_{\mathcal{A}}\left(\pi^{\boldsymbol{\theta}}(a|s') - \pi^{\widehat{\boldsymbol{\theta}}}(a|s')\right)\right. \\
&\quad \left.\times\int_{\mathcal{S}}\mathcal{P}(s|s',a) f(s)\mathrm{d}s\mathrm{d}a\mathrm{d}s'\right| : \|f\|_L \leq 1\Bigg\} && (17)\\
&\quad +\gamma\sup_f\left\{\left|\int_{\mathcal{S}}\left(\delta_\mu^{\boldsymbol{\theta}}(s') - \delta_\mu^{\widehat{\boldsymbol{\theta}}}(s')\right)\int_{\mathcal{A}}\pi^{\boldsymbol{\theta}}(a|s')\right. \\
&\quad \left.\times\int_{\mathcal{S}}\mathcal{P}(s|s',a) f(s)\mathrm{d}s\mathrm{d}a\mathrm{d}s'\right| : \|f\|_L \leq 1\Bigg\} && (18)
\end{aligned}$$

$$= \gamma L_{\mathcal{P}} \sup_{f} \left\{ \left| \int_{\mathcal{S}} \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s') \int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s') - \pi^{\widehat{\boldsymbol{\theta}}}(a|s') \right) \frac{g_f(s',a)}{L_{\mathcal{P}}} \mathrm{d}a\mathrm{d}s' \right| : \|f\|_L \leq 1 \right\}$$

$$+\gamma L_{\mathcal{P}} \left(L_{\pi^{\boldsymbol{\theta}}} + 1\right) \sup_{f} \left\{ \left| \int_{\mathcal{S}} \left(\delta_{\mu}^{\boldsymbol{\theta}}(s') - \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s') \right) \frac{h_f^{\boldsymbol{\theta}}(s')}{L_{\mathcal{P}} \left(L_{\pi^{\boldsymbol{\theta}}} + 1\right)} \mathrm{d}s' \right| : \|f\|_L \leq 1 \right\}$$

$$\leq \gamma L_{\mathcal{P}} \int_{\mathcal{S}} \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s') \sup_{\overline{f}} \left\{ \left| \int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s') - \pi^{\widehat{\boldsymbol{\theta}}}(a|s') \right) \overline{f}(s',a) \mathrm{d}a \right| : \left\| \overline{f} \right\|_{L,a} \leq 1 \right\}$$

$$+\gamma L_{\mathcal{P}} \left(L_{\pi^{\boldsymbol{\theta}}} + 1\right) \sup_{\widetilde{f}} \left\{ \left| \int_{\mathcal{S}} \left(\delta_{\mu}^{\boldsymbol{\theta}}(s') - \delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s') \right) \widetilde{f}(s') \mathrm{d}s' \right| : \left\| \widetilde{f} \right\|_L \leq 1 \right\} \tag{19}$$

$$\leq \gamma L_{\mathcal{P}} \sup_{s'} \mathcal{K} \left(\pi^{\boldsymbol{\theta}}(\cdot|s'), \pi^{\widehat{\boldsymbol{\theta}}}(\cdot|s') \right) + \gamma L_{\mathcal{P}} \left(L_{\pi^{\boldsymbol{\theta}}} + 1\right) \mathcal{K} \left(\delta_{\mu}^{\boldsymbol{\theta}}, \delta_{\mu}^{\widehat{\boldsymbol{\theta}}} \right) \tag{20}$$

$$\leq \gamma L_{\mathcal{P}} \left(L_{\pi}(\boldsymbol{\theta}) + (1 + L_{\pi^{\boldsymbol{\theta}}}) L_{\delta}(\boldsymbol{\theta}) \right) d(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}). \tag{21}$$

By replacing $\delta_{\mu}^{\boldsymbol{\theta}}$ with its definition we get equality (16). By adding and subtracting the term $\iiint_{\mathcal{SSA}} f(s)\mathcal{P}(s|s',a)\pi^{\boldsymbol{\theta}}(a|s')\delta_{\mu}^{\widehat{\boldsymbol{\theta}}}(s')\mathrm{d}a\mathrm{d}s'\mathrm{d}s$ and resorting to the triangle inequality, we derive lines (17) and (18). Such terms can be simplified by noting that they contain definition of $g_f(s,a)$ and $h_f^{\boldsymbol{\theta}}(s)$, respectively. After insertion of invariant scaling factors we rename $\overline{f}(s',a) = \frac{g_f(s',a)}{L_{\mathcal{P}}}$ and $\widetilde{f}(s') = \frac{h_f^{\boldsymbol{\theta}}(s')}{L_{\mathcal{P}}(L_{\pi^{\boldsymbol{\theta}}}+1)}$ in (19). Let $\|z(x,y)\|_{L,y}$ be the Lipschitz semi–norm of function z w.r.t. only variable y (taking the supremum over x). Then, from inequality (11), it is easy to see that $\left\|\overline{f}\right\|_{L,a} \leq 1$. Similarly, from inequality (15), we derive that $\left\|\widetilde{f}\right\|_L \leq 1$. As done in the proof of Lemma 2, we push the supremum into the state integral and we maximize the Kantorovich distance [inequality (20)].

Note that inequality (21) leads to the following fixed point equation:

$$L_{\delta}(\boldsymbol{\theta}) = \gamma L_{\mathcal{P}} \left[L_{\pi}(\boldsymbol{\theta}) + (1 + L_{\pi^{\boldsymbol{\theta}}}) L_{\delta}(\boldsymbol{\theta}) \right]$$

that admits a unique feasible solution only if $\gamma L_{\mathcal{P}} \left(1 + L_{\pi^{\boldsymbol{\theta}}}\right) < 1$. □

As expected, the smoothness of the γ-discounted future state distribution w.r.t. to $\boldsymbol{\theta}$ strongly depends on the smoothness of the state transition model and the policy model. In particular, a relevant role is played by $L_{\mathcal{P}}$ that influences both the numerator and the denominator of L_{δ} (decreasing $L_{\mathcal{P}}$ decreases the value of the numerator and increases the value of the denominator). As in the case of the Lipschitz continuity of the Q- and V-functions (see Lemma 1), the Lipschitz continuity of the γ-discounted future state distribution can be guarantee only when the condition $\gamma L_{\mathcal{P}} \left(1 + L_{\pi^{\boldsymbol{\theta}}}\right) < 1$ holds. This condition emerges from the recursive nature of the considered functions and enforces the discounted Markov kernel underlying policy π to be a contraction w.r.t. the Kantorovich distance.

Finally, combining Proposition 1 with Lemmas 2 and 3, we can derive the Lipschitz continuity of the joint distribution ζ_{μ}.

Lemma 4 *Under Assumption 1 and 2, if* $\gamma L_{\mathcal{P}} \left(1 + L_{\pi^{\boldsymbol{\theta}}}\right) < 1$, *then the joint distribution* $\zeta_{\mu}^{\boldsymbol{\theta}}$ *is* $L_{\zeta}(\boldsymbol{\theta})$*-PLC w.r.t. the policy parameters* $\boldsymbol{\theta}$, *with:*

$$L_{\zeta}(\boldsymbol{\theta}) = \frac{L_{\pi}(\boldsymbol{\theta})}{1 - \gamma L_{\mathcal{P}} \left(1 + L_{\pi^{\boldsymbol{\theta}}}\right)}.$$

The lemma comes directly from the application of Lemma 3 to Lemma 2. Now we have all the technicalities required to derive the main theorem.

Theorem 1 *Given an* $(L_{\mathcal{P}}, L_{\mathcal{R}})$*-LC MDP and an* $(L_{\pi^{\boldsymbol{\theta}}}, L_{\pi}(\boldsymbol{\theta}))$*-LC stationary policy model, if* $\gamma L_{\mathcal{P}} \left(1 + L_{\pi^{\boldsymbol{\theta}}}\right) < 1$*, then the performance measure* $J_{\mu}^{\boldsymbol{\theta}}$ *is* $L_J(\boldsymbol{\theta})$*-PLC w.r.t. the policies parameters:*

$$\left| J_{\mu}^{\boldsymbol{\theta}} - J_{\mu}^{\widehat{\boldsymbol{\theta}}} \right| \leq L_J(\boldsymbol{\theta})\, d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right),$$

with:

$$L_J(\boldsymbol{\theta}) = \frac{L_{\mathcal{R}}\, L_{\pi}(\boldsymbol{\theta})}{(1-\gamma)\left(1 - \gamma L_{\mathcal{P}}\left(1 + L_{\pi^{\boldsymbol{\theta}}}\right)\right)}.$$

The proof follows from the application of Lemma 4 to Proposition 1. Although L_J resembles the Lipschitz constant of the Q-function (actually $L_J = \frac{L_{\mathcal{R}}}{1-\gamma} L_{Q^{\pi}}$), it is worth underline that they define Lipschitz conditions over different spaces: the expected return is L_J-PLC w.r.t. policy parameters $\boldsymbol{\theta}$, while $L_{Q^{\pi}}$ is a Lipschitz constant w.r.t. the state-action space.

4 Lipschitz continuity of the policy gradient

Leveraging on the results presented in the previous section, here we investigate the Lipschitz continuity of the gradient of the expected discounted reward w.r.t. the policy parameters. Both the expected return of a policy $\pi^{\boldsymbol{\theta}}$ and its gradient can be defined as expected values w.r.t. distribution $\zeta_{\mu}^{\boldsymbol{\theta}}$: in the former case the function to be averaged is the reward function $\mathcal{R}(s,a)$, while in the latter case is $\nabla_{\boldsymbol{\theta}} \log \pi^{\boldsymbol{\theta}}(s,a) Q^{\boldsymbol{\theta}}(s,a)$ (see Eq. 3). For ease of notation, we define the function $\boldsymbol{\eta} : \mathcal{S} \times \mathcal{A} \times \Theta \to \mathbb{R}^d$ as

$$\boldsymbol{\eta}^{\boldsymbol{\theta}}(s,a) = \nabla_{\boldsymbol{\theta}} \log \pi^{\boldsymbol{\theta}}(s,a) Q^{\boldsymbol{\theta}}(s,a).$$

In particular, we consider the component-wise absolute difference between gradients corresponding to different parameterizations:

$$\left| \nabla_{\boldsymbol{\theta}_i} J_{\mu}^{\boldsymbol{\theta}} - \nabla_{\boldsymbol{\theta}_i} J_{\mu}^{\widehat{\boldsymbol{\theta}}} \right| = \frac{1}{1-\gamma} \left| \mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\boldsymbol{\theta}}} \left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) \right] - \mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}} \left[\boldsymbol{\eta}_i^{\widehat{\boldsymbol{\theta}}}(s,a) \right] \right|.$$

It is worth nothing that, differently from the reward function in the expected-return case, functions $\boldsymbol{\eta}^{\boldsymbol{\theta}}$ do depend on the policy parameters $\boldsymbol{\theta}$. This prevents to follow immediately the same steps as done in the previous section and requires to decompose the problem by introducing the following upper bound.

Proposition 2 *For any pair of stationary policies corresponding to parameters* $\boldsymbol{\theta}$ *and* $\widehat{\boldsymbol{\theta}}$*, the component-wise absolute difference between the gradients of the expected return can be upper bounded as follows:*

$$\begin{aligned} \left| \nabla_{\boldsymbol{\theta}_i} J_{\mu}^{\boldsymbol{\theta}} - \nabla_{\boldsymbol{\theta}_i} J_{\mu}^{\widehat{\boldsymbol{\theta}}} \right| \leq\ & \frac{1}{1-\gamma} \left| \mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\boldsymbol{\theta}}} \left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) \right] - \mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}} \left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) \right] \right| \\ & + \frac{1}{1-\gamma} \left| \mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}} \left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) - \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\theta}}}(s,a) \right] \right|. \end{aligned}$$

While the second term requires a further expansion (that will be presented later in the section), the first one can be bounded following a similar argument as the one used for the expected return.

Upper bound to the first term. While the bound of the expected reward (Lemma 1) follows directly from the definition of Lipschitz MDP (Assumption 1), here, we need to prove that the function $\boldsymbol{\eta}_i^{\boldsymbol{\theta}}$ is Lipschitz w.r.t. the joint state-action space. Since the product of two Lipschitz functions is Lipschitz, given Assumption 2 and Lemma 1, we can show that $\boldsymbol{\eta}^{\boldsymbol{\theta}}$ is LC w.r.t. the state-action space (see "Proof of Lemma 5" section of Appendix for the proof).

Lemma 5 *Under Assumptions 1, 2 and 3, the i-th component of $\boldsymbol{\eta}^{\boldsymbol{\theta}}$ is $L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}}$-LC w.r.t. the state-action space, that is: $\forall (s,\widehat{s},a,\widehat{a}) \in \mathcal{S}^2 \times \mathcal{A}^2$,*

$$\left|\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) - \boldsymbol{\eta}_i^{\boldsymbol{\theta}}(\widehat{s},\widehat{a})\right| \leq L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}}\, d_{\mathcal{SA}}\left((s,a),(\widehat{s},\widehat{a})\right),$$

where $L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}} = \frac{R}{1-\gamma} L^i_{\nabla \log \pi^{\boldsymbol{\theta}}} + M^i_{\boldsymbol{\theta}} L_{Q^{\boldsymbol{\theta}}}$.

Since $\boldsymbol{\eta}_i^{\boldsymbol{\theta}}$ is LC, it follows that

$$\left|\mathop{\mathbb{E}}_{(s,a)\sim\zeta_\mu^{\boldsymbol{\theta}}}\left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a)\right] - \mathop{\mathbb{E}}_{(s,a)\sim\zeta_\mu^{\widehat{\boldsymbol{\theta}}}}\left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a)\right]\right| \leq L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}} \mathcal{K}\left(\zeta_\mu^{\boldsymbol{\theta}}, \zeta_\mu^{\widehat{\boldsymbol{\theta}}}\right) \leq L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}} L_\zeta(\boldsymbol{\theta}) d_\Theta\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right).$$

Upper bound to the second term. To prove its Lipschitz continuity w.r.t. policy parameters, we need to introduce an upper bound to $\left|\boldsymbol{\eta}_i^{\boldsymbol{\theta}} - \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\theta}}}\right|$. Refer to "Proof of Lemma 6" section of Appendix for the proof.

Lemma 6 *For any pair of stationary policies corresponding to $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, the absolute difference of the i-th component of functions $\boldsymbol{\eta}^{\boldsymbol{\theta}}$ and $\boldsymbol{\eta}^{\widehat{\boldsymbol{\theta}}}$ is upper bounded by: $\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \forall (\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}) \in \Theta^2$,*

$$\begin{aligned}\left|\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) - \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\theta}}}(s,a)\right| \leq{}& \frac{R}{1-\gamma}\left|\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a|s) - \nabla_{\boldsymbol{\theta}_i}\log\pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right| \\ &+ M^i_{\boldsymbol{\theta}}\left|Q^{\boldsymbol{\theta}}(s,a) - Q^{\widehat{\boldsymbol{\theta}}}(s,a)\right|.\end{aligned}$$

The first term of the bound can be upper bounded in turn by exploiting the LC assumption on $\nabla \log \pi$ w.r.t. policy parameters (see Assumption 3). For what concerns the second term, we need to show further results about the Lipschitz continuity of the Q- and V-functions w.r.t. the policy parameters. Here we extend the standard Lipschitz framework for MDPs to the case in which a parametric policy model is available.

Theorem 2 *Under Assumptions 1 and 2, the V-function and the Q-function are respectively $L_V(\boldsymbol{\theta})$- and $L_Q(\boldsymbol{\theta})$-PLC w.r.t. to the policy parameters, with:*

$$L_V(\boldsymbol{\theta}) = \frac{L_\pi(\boldsymbol{\theta}) L_{\mathcal{R}}}{(1-\gamma)\left(1-\gamma L_{\mathcal{P}}\left(1+L_{\pi^{\boldsymbol{\theta}}}\right)\right)}; \qquad L_Q(\boldsymbol{\theta}) = \gamma L_V(\boldsymbol{\theta}).$$

Proof We need to introduce some preliminary results that will be used to prove the main theorem. First of all, we want to prove that the expected reward under two stationary policies $\pi^{\boldsymbol{\theta}}$ and $\pi^{\widehat{\boldsymbol{\theta}}}$ corresponding to different policy parameterizations is Lipschitz continuous for any state s:

$$\begin{aligned}\left|\mathcal{R}^{\boldsymbol{\theta}}(s) - \mathcal{R}^{\widehat{\boldsymbol{\theta}}}(s)\right| &= \int_{a\in\mathcal{A}}\left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right)\mathcal{R}(s,a)\mathrm{d}a \\ &\leq L_{\mathcal{R}}\, \mathcal{K}\left(\pi^{\boldsymbol{\theta}}(\cdot|s), \pi^{\widehat{\boldsymbol{\theta}}}(\cdot|s)\right) \\ &\leq L_{\mathcal{R}}\, L_\pi(\boldsymbol{\theta})\, d_\Theta\left(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}\right).\end{aligned}$$

The following equation gives an upper bound to the maximum absolute difference of the V-functions associated to two policy parameterizations: $\forall s \in \mathcal{S}$

$$\begin{aligned}
\left|V^{\boldsymbol{\theta}}(s) - V^{\widehat{\boldsymbol{\theta}}}(s)\right| &= \left| \int_{\mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \int_{\mathcal{S}} V^{\boldsymbol{\theta}}(s') \mathcal{P}(\mathrm{d}s'|s,a) \right) \pi^{\boldsymbol{\theta}}(\mathrm{d}a|s) \right. \\
&\quad \left. - \int_{\mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \int_{\mathcal{S}} V^{\widehat{\boldsymbol{\theta}}}(s') \mathcal{P}(\mathrm{d}s'|s,a) \right) \pi^{\widehat{\boldsymbol{\theta}}}(\mathrm{d}a|s) \right| \\
&\leq \left| \int_{\mathcal{A}} \mathcal{R}(s,a) \left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\widehat{\boldsymbol{\theta}}}(a|s) \right) \mathrm{d}a \right| \\
&\quad + \gamma \left| \iint_{\mathcal{S}\mathcal{A}} \mathcal{P}(s'|s,a) \left(\pi^{\boldsymbol{\theta}}(a|s) V^{\boldsymbol{\theta}}(s') - \pi^{\widehat{\boldsymbol{\theta}}}(a|s) V^{\widehat{\boldsymbol{\theta}}}(s') \right) \mathrm{d}a \mathrm{d}s' \right| \\
&\leq \left| \mathcal{R}^{\boldsymbol{\theta}}(s) - \mathcal{R}^{\widehat{\boldsymbol{\theta}}}(s) \right| + \gamma \left| \int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\widehat{\boldsymbol{\theta}}}(a|s) \right) \right. \\
&\quad \left. \times \int_{\mathcal{S}} \mathcal{P}(s'|s,a) V^{\boldsymbol{\theta}}(s') \mathrm{d}s' \mathrm{d}a \right| \\
&\quad + \gamma \left| \int_{\mathcal{A}} \pi^{\widehat{\boldsymbol{\theta}}}(a|s) \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \left(V^{\boldsymbol{\theta}}(s') - V^{\widehat{\boldsymbol{\theta}}}(s') \right) \mathrm{d}s' \mathrm{d}a \right| \\
&= \left| \mathcal{R}^{\boldsymbol{\theta}}(s) - \mathcal{R}^{\widehat{\boldsymbol{\theta}}}(s) \right| + \gamma L_V L_{\mathcal{P}} \left| \int_{\mathcal{A}} \left(\pi^{\boldsymbol{\theta}}(a|s) - \pi^{\widehat{\boldsymbol{\theta}}}(a|s) \right) \right. \\
&\quad \left. \times \int_{\mathcal{S}} \frac{\mathcal{P}(s'|s,a) V^{\boldsymbol{\theta}}(s')}{L_V \; L_{\mathcal{P}}} \mathrm{d}s' \mathrm{d}a \right| \\
&\quad + \gamma \left| \int_{\mathcal{A}} \pi^{\widehat{\boldsymbol{\theta}}}(a|s) \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \left(V^{\boldsymbol{\theta}}(s') - V^{\widehat{\boldsymbol{\theta}}}(s') \right) \mathrm{d}s' \mathrm{d}a \right| \\
&\leq \left| \mathcal{R}^{\boldsymbol{\theta}}(s) - \mathcal{R}^{\widehat{\boldsymbol{\theta}}}(s) \right| + \gamma L_V \; L_{\mathcal{P}} \mathcal{K} \left(\pi^{\boldsymbol{\theta}}(\cdot|s), \pi^{\widehat{\boldsymbol{\theta}}}(\cdot|s) \right) + \gamma L_V(\boldsymbol{\theta}) \; d_{\Theta}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \\
&\leq (L_{\mathcal{R}} + \gamma L_V \; L_{\mathcal{P}}) \, L_{\pi}(\boldsymbol{\theta}) \; d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) + \gamma L_V(\boldsymbol{\theta}) \; d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) \\
&\leq \frac{L_{\pi}(\boldsymbol{\theta})}{1-\gamma} \left(L_{\mathcal{R}} + \gamma L_V \; L_{\mathcal{P}}\right) d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right).
\end{aligned}$$

The proof follows from the substitution of the value L_V with its definition (Rachelson and Lagoudakis 2010). The following equations provides the Lipschitz continuity of the Q-function: $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
\left|Q^{\boldsymbol{\theta}}(s,a) - Q^{\widehat{\boldsymbol{\theta}}}(s,a)\right| &= \gamma \left| \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \left(V^{\boldsymbol{\theta}}(s') - V^{\widehat{\boldsymbol{\theta}}}(s') \right) \mathrm{d}s' \right| \\
&\leq \gamma \int_{\mathcal{S}} \mathcal{P}(s'|s,a) \left| V^{\boldsymbol{\theta}}(s') - V^{\widehat{\boldsymbol{\theta}}}(s') \right| \mathrm{d}s' \\
&\leq \gamma \sup_{s'} \left| V^{\boldsymbol{\theta}}(s') - V^{\widehat{\boldsymbol{\theta}}}(s') \right| \leq \gamma L_V(\boldsymbol{\theta}) \; d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right).
\end{aligned}$$

□

Similarly as done for Lemma 5, we can observe that $\boldsymbol{\eta}^{\boldsymbol{\theta}}$ is the product of two functions that are Lipschitz continuous w.r.t. parameters $\boldsymbol{\theta}$. Proof can be founded in "Proof of Lemma 7" section of Appendix.

Lemma 7 *Under Assumptions 1, 2 and 3, the i-th component of $\boldsymbol{\eta}$ is $L^i_{\boldsymbol{\eta}}(\boldsymbol{\theta})$-PLC w.r.t. the policy parameters, that is:* $\forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \forall (\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \in \Theta^2$,

$$\left| \boldsymbol{\eta}^{\boldsymbol{\theta}}_i(s,a) - \boldsymbol{\eta}^{\widehat{\boldsymbol{\theta}}}_i(s,a) \right| \leq L^i_{\boldsymbol{\eta}}(\boldsymbol{\theta})\, d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right),$$

where $L^i_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \frac{R}{1-\gamma} L^i_{\nabla \log \pi}(\boldsymbol{\theta}) + M^i_{\boldsymbol{\theta}} L_Q(\boldsymbol{\theta})$.

Finally, combining Lemmas 5 and 7 (see "Proof of Theorem 3" section of Appendix for the proof), we are ready to state our main result about the Lipschitz continuity of the policy gradient w.r.t. policy parameters $\boldsymbol{\theta}$.

Theorem 3 *Under Assumptions 1, 2, and 3, the i-th component of the gradient $\nabla_{\boldsymbol{\theta}} J$ of the expected return is $L^i_{\nabla J}(\boldsymbol{\theta})$-PLC, that is:* $\forall (\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \in \Theta^2$,

$$\left| \nabla_{\boldsymbol{\theta}_i} J^{\boldsymbol{\theta}}_{\mu} - \nabla_{\boldsymbol{\theta}_i} J^{\widehat{\boldsymbol{\theta}}}_{\mu} \right| \leq L^i_{\nabla J}(\boldsymbol{\theta})\, d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right),$$

where $L^i_{\nabla J}(\boldsymbol{\theta}) = \frac{1}{1-\gamma}\left(L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}} L_{\zeta}(\boldsymbol{\theta}) + L^i_{\boldsymbol{\eta}}(\boldsymbol{\theta})\right)$.

Given the vector $\mathbf{L}_{\nabla J}(\boldsymbol{\theta}) = [L^1_{\nabla J}(\boldsymbol{\theta}), \ldots, L^d \nabla J(\boldsymbol{\theta})]$, the policy gradient $\nabla_{\boldsymbol{\theta}} J^{\boldsymbol{\theta}}_{\mu}$ is $L_{\nabla J}(\boldsymbol{\theta})$-LC in Θ, where $L_{\nabla J}(\boldsymbol{\theta}) = \left\| \mathbf{L}_{\nabla J}(\boldsymbol{\theta}) \right\|_2$ when $d_{\Theta}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \left\| \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}} \right\|_2$.

5 Updating policy parameters

In this section, we will show how the Lipschitz continuity of the policy gradient discussed in the previous section can be exploited to update the policy parameters and to guarantee performance improvement. The following lemma exploits the Taylor expansion and the Lipschitz continuity of the policy gradient to derive a lower bound to the policy performance improvement.

Lemma 8 *If the policy gradient is Lipschitz continuous, the policy performance improvement between policy $\widehat{\boldsymbol{\theta}}$ and policy $\boldsymbol{\theta}$ can be lower bounded as follows:*

$$J^{\widehat{\boldsymbol{\theta}}}_{\mu} - J^{\boldsymbol{\theta}}_{\mu} \geq \nabla_{\boldsymbol{\theta}} J^{\boldsymbol{\theta}^T}_{\mu} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) - \frac{1}{2} \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta})^T \cdot \left| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right|,$$

where $|v|$ denotes the component-wise absolute value when v is a vector.

Proof According to the definition of Lipschitz gradient given in Theorem 3, we have that, for some $t \in [0, 1]$:

$$\begin{aligned}
\left| \nabla_{\boldsymbol{\theta}_i} J^{\widehat{\boldsymbol{\theta}}}_{\mu} - \nabla_{\boldsymbol{\theta}_i} J^{\boldsymbol{\theta}}_{\mu} \right| &= \left\| \int_0^1 \nabla \left(\nabla_{\boldsymbol{\theta}_i} J_{\mu}\right) \left(t\, \widehat{\boldsymbol{\theta}} + (1-t)\boldsymbol{\theta}\right)^{\mathrm{T}} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \mathrm{d}t \right\|_p \\
&\leq \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|_p \int_0^1 \left\| \nabla \left(\nabla_{\boldsymbol{\theta}_i} J_{\mu}\right) \left(t\, \widehat{\boldsymbol{\theta}} + (1-t)\boldsymbol{\theta}\right) \right\|_p \mathrm{d}t \\
&\leq L^i_{\nabla J}(\boldsymbol{\theta}) \left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|_p .
\end{aligned}$$

If we only consider the first term of the Taylor's expansion, we can write the following upper bound:

$$
\begin{aligned}
J_\mu^{\widehat{\boldsymbol{\theta}}} &= J_\mu^{\boldsymbol{\theta}} + \int_0^1 \nabla_{\boldsymbol{\theta}} J_\mu^{t\,\widehat{\boldsymbol{\theta}}+(1-t)\boldsymbol{\theta}^{\mathrm{T}}} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \mathrm{d}t \pm \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}^{\mathrm{T}}} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \\
&= J_\mu^{\boldsymbol{\theta}} + \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}^{\mathrm{T}}} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) + \int_0^1 \left(\nabla_{\boldsymbol{\theta}} J_\mu^{t\,\widehat{\boldsymbol{\theta}}+(1-t)\boldsymbol{\theta}} - \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}}\right)^{\mathrm{T}} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \mathrm{d}t \\
&= J_\mu^{\boldsymbol{\theta}} + \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}^{\mathrm{T}}} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) + \sum_i \int_0^1 \left(\nabla_{\boldsymbol{\theta}_i} J_\mu^{t\,\widehat{\boldsymbol{\theta}}+(1-t)\boldsymbol{\theta}} - \nabla_{\boldsymbol{\theta}_i} J_\mu^{\boldsymbol{\theta}}\right) \left(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right) \mathrm{d}t \\
&\geq J_\mu^{\boldsymbol{\theta}} + \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}^{\mathrm{T}}} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) - \sum_i \int_0^1 \left|\left(\nabla_{\boldsymbol{\theta}_i} J_\mu^{t\,\widehat{\boldsymbol{\theta}}+(1-t)\boldsymbol{\theta}} - \nabla_{\boldsymbol{\theta}_i} J_\mu^{\boldsymbol{\theta}}\right)\right| \left|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right| \mathrm{d}t \\
&\geq J_\mu^{\boldsymbol{\theta}} + \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}^{\mathrm{T}}} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) - \int_0^1 \left\| t\,\widehat{\boldsymbol{\theta}} + (1-t)\boldsymbol{\theta} - \boldsymbol{\theta}\right\|_p \mathrm{d}t \sum_i L_{\nabla J}^i \left|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right| \\
&\geq J_\mu^{\boldsymbol{\theta}} + \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}^{\mathrm{T}}} \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) - \frac{1}{2} \left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_p \sum_i L_{\nabla J}^i \left|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right|.
\end{aligned}
$$

□

In Sect. 2.3, we have seen that the policy parameters are updated as follows

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \Delta\boldsymbol{\theta}_t,$$

where, in the steepest ascent approaches, $\Delta\boldsymbol{\theta}_t = \alpha_t \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}$ and α_t is a parameter that determines the step-size length. Recall that the steepest ascent direction of $J_\mu^{\boldsymbol{\theta}_t}$ is defined as the vector $\Delta\boldsymbol{\theta}_t$ that maximizes $J_\mu^{\boldsymbol{\theta}_t+\Delta\boldsymbol{\theta}_t}$ under the constraint that the change in the parameters ($\|\Delta\boldsymbol{\theta}_t\|_p$) is sufficiently small.

In the following we describe three approaches to determine the step size exploiting the Lipschitz continuity of the policy gradient.

5.1 Single step size from single Lipschitz constant (SSS–SLC)

The scenario where a single Lipschitz constant is available for the gradient has been widely studied in the optimization literature. As suggested by Armijo (1966), if policy gradient is $L_{\nabla J}$-LC and the L_2-norm is used as metric, fixing the step size to the reciprocal value of the Lipschitz constant[3]:

$$\alpha = \frac{1}{L_{\nabla J}},$$

guarantees to improve at each iteration. Furthermore, when the function is convex, it allows to converge to an ϵ-optimal solution in $O(\frac{1}{\epsilon})$ iterations. It can be easily shown that similar results hold when the Lipschitz constant is replaced with its pointwise version ($\alpha_t = L_{\nabla J}(\boldsymbol{\theta}_t)^{-1}$), with the advantage of inducing larger step sizes—being $L_{\nabla J} = \sup_{\boldsymbol{\theta}} L_{\nabla J}(\boldsymbol{\theta})$—with better performance improvements. Such result is simply obtained by the maximization w.r.t. the step size of the following quadratic lower bound to the performance improvement derived

[3] See (Armijo 1966) for conditions under which convergence occurs and a proof of convergence.

from Lemma 8 using the L_2-norm as metric:

$$J_\mu^{\boldsymbol{\theta}_{t+1}} - J_\mu^{\boldsymbol{\theta}_t} \geq \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}{}^{\mathrm{T}} \cdot \Delta\boldsymbol{\theta}_t - \frac{L_{\nabla J}(\boldsymbol{\theta}_t)}{2} \|\Delta\boldsymbol{\theta}_t\|_2^2$$
$$= \alpha_t \left\|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right\|_2^2 - \alpha_t^2 \frac{L_{\nabla J}(\boldsymbol{\theta}_t)}{2} \left\|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right\|_2^2.$$

Using the step size α_t that maximizes the above lower bound, we are guaranteed that, at each iteration, the policy improvement is larger than $\frac{\left\|\nabla_{\boldsymbol{\theta}_i} J_\mu^{\boldsymbol{\theta}_t}\right\|_2^2}{2L_{\nabla J}(\boldsymbol{\theta}_t)}$. As expected, the smaller are the Lipschitz constants of the MDP and the policy model, the smaller is the Lipschitz constant of the policy gradient, and the larger are the step size and the guaranteed improvement at each iteration.

5.2 Single step size from multiple Lipschitz constants (SSS–MLC)

When a Lipschitz constant for each gradient component is available, it is convenient to exploit such information. By maximizing the bound in Lemma 8:

$$J_\mu^{\boldsymbol{\theta}_{t+1}} - J_\mu^{\boldsymbol{\theta}_t} \geq \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}{}^{\mathrm{T}} \cdot \Delta\boldsymbol{\theta}_t - \frac{1}{2} \|\Delta\boldsymbol{\theta}_t\|_2 \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \cdot |\Delta\boldsymbol{\theta}_t|$$
$$= \alpha_t \left\|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right\|_2^2 - \frac{1}{2}\alpha_t^2 \left\|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right\|_2 \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \cdot \left|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right|,$$

the following step size is obtained:

$$\alpha_t = \frac{\left\|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right\|_2}{\mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \cdot \left|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right|}.$$

Notice that, differently from the single Lipschitz constant case, the step size directly exploits the information of the gradient at time t. In this case, the policy performance improves at least by $\frac{\left\|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right\|_2^3}{2\mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \cdot \left|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right|}$, that is never worse than the improvement that is guaranteed with the SSS–SLC approach.

5.3 Multiple step sizes from multiple Lipschitz constants (MSS–MLC)

The two previous approaches update the parameters along the steepest ascent direction. However, to mitigate the main drawbacks of the steepest ascent method, many gradient approaches follow different directions (e.g. Amari and Douglas 1998). In particular, when a Lipschitz constant for each gradient component is available, it is interesting to consider a different step size for each parameter: $\Delta\boldsymbol{\theta}_t^i = \alpha_t^i \nabla_{\boldsymbol{\theta}_i} J_\mu^{\boldsymbol{\theta}_t}$. As a result, the new candidate solution $\boldsymbol{\theta}_{t+1}$ may lie outside the policy gradient direction $\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}$. All the step sizes can be obtained by maximizing the following concave[4] function w.r.t. the step sizes $[\alpha_t^1, \ldots, \alpha_t^d]$:

$$J_\mu^{\boldsymbol{\theta}_{t+1}} - J_\mu^{\boldsymbol{\theta}_t} \geq \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}{}^{\mathrm{T}} \cdot \mathbf{A}_t \cdot \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t} - \frac{1}{2} \left\|\mathbf{A}_t \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \cdot \mathbf{A}_t \cdot \left|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right|, \quad (22)$$

where $\mathbf{A}_t = diag(\alpha_t^1, \ldots, \alpha_t^d)$ and $\alpha_t^i \geq 0, \forall i \in \{1, \ldots, d\}$.

[4] The reader may refer to "Proof of Concavity of Function" section of Appendix for the proof.

Algorithm 1 MSS-MLC Algorithm

function MSS- MLC($\nabla_{\boldsymbol{\theta}} J_\mu$, $\mathbf{L}_{\nabla J}$)
  initialize: $V = 0$
  $\phi = \arccos\left(\frac{\mathbf{L}_{\nabla J}{}^{\mathrm{T}} \left|\nabla_{\boldsymbol{\theta}} J_\mu\right|}{\|\mathbf{L}_{\nabla J}\|_2 \|\nabla_{\boldsymbol{\theta}} J_\mu\|_2}\right)$
  if $\phi \leq \arctan(\sqrt{1/8})$ **then**
    $\psi = \arctan\left(\frac{1-\sqrt{1-8\tan(\phi)^2}}{4\tan(\phi)}\right)$
    $V = \frac{\left\|\nabla_{\boldsymbol{\theta}} J_\mu\right\|_2^2 \cos(\psi)^2}{2\|\mathbf{L}_{\nabla J}\|_2 cos(\psi+\phi)}$
    Create vector $\Delta\boldsymbol{\theta} = \frac{\left|\nabla_{\boldsymbol{\theta}} J_\mu\right| \cos(\psi)}{\|\mathbf{L}_{\nabla J}\|_2 cos(\psi+\phi)}$
    Rotate $\Delta\boldsymbol{\theta}$ of angle ϕ around bivector $\mathbf{L}_{\nabla J} \wedge \left|\nabla_{\boldsymbol{\theta}} J_\mu\right|$
  end if
  $i^* = arg\min_i \frac{\left|\nabla_{\boldsymbol{\theta}} J_\mu^i\right|}{\mathbf{L}^i_{\nabla J}}$
  Build $\widetilde{\nabla_{\boldsymbol{\theta}} J_\mu}$ and $\widetilde{\mathbf{L}_{\nabla J}}$ by removing component i^*
  $[\tilde{V}, \tilde{\boldsymbol{\alpha}}]$=MSS- MLC($\widetilde{\nabla_{\boldsymbol{\theta}} J_\mu}$, $\widetilde{\mathbf{L}_{\nabla J}}$)
  if $\tilde{V} > V$ **then**
    return $\tilde{V}$ and $[\tilde{\alpha}^1, \ldots, \tilde{\alpha}^{i^*-1}, 0, \tilde{\alpha}^{i^*}, \ldots, \tilde{\alpha}^d]$
  else
    return V and $\left[\frac{\Delta\boldsymbol{\theta}^1}{\left|\nabla_{\boldsymbol{\theta}} J_\mu^i\right|}{}^1, \frac{\Delta\boldsymbol{\theta}^2}{\left|\nabla_{\boldsymbol{\theta}} J_\mu^i\right|}{}^2, \ldots, \frac{\Delta\boldsymbol{\theta}^d}{\left|\nabla_{\boldsymbol{\theta}} J_\mu^i\right|}{}^d\right]$
  end if
end function

Being the performance-improvement bound a concave function in the step size, we could resort to one of the standard tools in convex optimization to find the optimal learning rates. Nonetheless, we propose a more efficient algorithm (reported in Algorithm 1) that optimizes the performance improvement bound by taking the best among a set of d closed-form candidate solutions (the algorithm is presented as a recursive function for conciseness, but an iterative implementation is straightforward). The computational complexity of MSS-MLC is $\mathcal{O}(d^3)$.

To explain how the algorithm works we need to rewrite the performance improvement bound as follows:

$$\begin{aligned} J_\mu^{\boldsymbol{\theta}_{t+1}} - J_\mu^{\boldsymbol{\theta}_t} &\geq \nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}{}^{\mathrm{T}} \Delta\boldsymbol{\theta}_t - \frac{1}{2} \|\Delta\boldsymbol{\theta}_t\|_2 \ \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} |\Delta\boldsymbol{\theta}_t| \\ &= \left\|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right\|_2 \|\Delta\boldsymbol{\theta}_t\|_2 \cos(\beta_{\nabla J_\mu}) - \frac{1}{2} \|\mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)\|_2 \|\Delta\boldsymbol{\theta}_t\|_2^2 \cos(\beta_L). \end{aligned}$$

To compute the length of the optimal step size $\|\Delta\boldsymbol{\theta}_t\|_2$ w.r.t. to the above bound, we can compute the derivative and put it equal to zero:

$$\|\Delta\boldsymbol{\theta}_t\|_2 = \frac{\left\|\nabla_{\boldsymbol{\theta}} J_\mu^{\boldsymbol{\theta}_t}\right\|_2 \cos(\beta_{\nabla J_\mu})}{\|\mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)\|_2 \cos(\beta_L)},$$

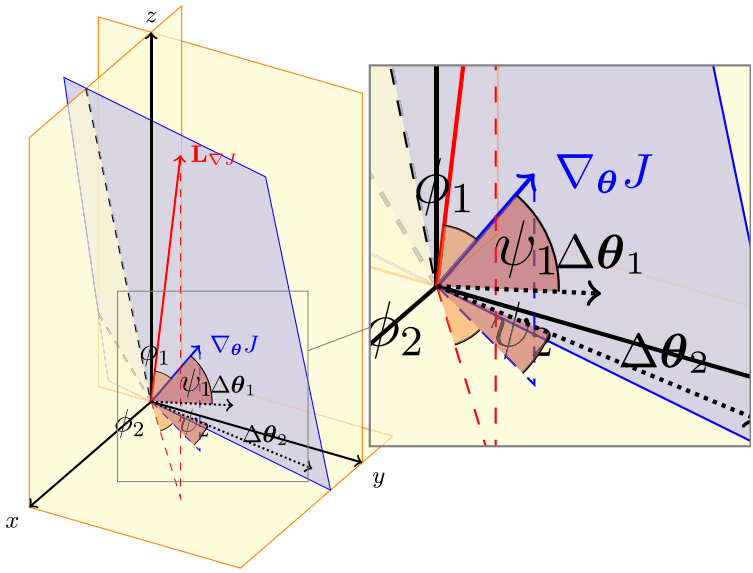

Fig. 1 Geometrical interpretation of the MSS–MLC algorithm

that leads to the following bound:

$$J_{\mu}^{\boldsymbol{\theta}_{t+1}} - J_{\mu}^{\boldsymbol{\theta}_t} \geq \frac{\left\| \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right\|_2^2 \cos(\beta_{\nabla J_{\mu}})^2}{2 \left\| \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t) \right\|_2 \cos(\beta_L)}.$$

To maximize the performance improvement we need to choose the direction of $\Delta\boldsymbol{\theta}_t$ so that it is as close as possible to $\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}$ (to maximize the value of $\cos(\beta_{\nabla J_{\mu}})$) and as far as possible to the vector $\mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)$ (so as to minimize $\cos(\beta_L)$). It can be easily shown that the combination of these two desiderata puts $\Delta\boldsymbol{\theta}_t$ on the plane identified by the vectors $\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}$ and $\mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)$ closer to vector $\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}$ (refer to Fig. 1 for a geometrical interpretation of the algorithm in a three-dimensional problem). It follows that, calling ϕ the angle between vector $\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}$ and $\mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)$, and renaming $\beta_{\nabla J_{\mu}}$ as ψ, we can rewrite the bound on the performance improvement as:

$$J_{\mu}^{\boldsymbol{\theta}_{t+1}} - J_{\mu}^{\boldsymbol{\theta}_t} \geq \frac{\left\| \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right\|_2^2 \cos(\psi)^2}{2 \left\| \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t) \right\|_2 \cos(\psi + \phi)},$$

from which we can compute the optimal value for ψ: if $\phi \leq \arctan(1/8)$ the function presents a local maximum at $\psi = \arctan\left(\frac{1-\sqrt{1-8\tan(\phi)^2}}{4\tan(\phi)}\right)$ associated to a guaranteed performance improvement equal to $\frac{\left\| \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right\|_2^2 \cos(\psi)^2}{2\|\mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)\|_2 cos(\psi+\phi)}$, followed by a local minimum and then increasing with ψ. Otherwise, if $\phi > \arctan(1/8)$, the bound is monotonically increasing with ψ. However, since we have the constraint that the step sizes must be positive, the angle ψ cannot increase arbitrarily, and must stop at the intersection with the planes identified by the Cartesian axes. So we need to explore what happens on the boundaries of the optimization problem (i.e., when one of the step sizes is put to zero). To do this we have to identify which

step size is the first to reach zero as ψ increases, that is the parameter with the smallest $\frac{\left|\nabla_{\boldsymbol{\theta}} J_{\mu}^{i}\right|}{\mathbf{L}_{\nabla J}^{i}}$ ratio. Once identified such parameter, the guaranteed performance improvement is saved, the parameter is removed from the vectors $\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}$ and $\mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)$, and the same procedure is repeated until the vector size is reduced to one. Finally, the optimal solution is the one associated to the largest guaranteed performance improvement among the local maxima found during the execution of the algorithm. Once the optimal $\Delta\boldsymbol{\theta}$ has been identified, the i-th step-size parameter is easily computed through the following ratio: $\frac{\Delta\boldsymbol{\theta}^{i}}{\left|\nabla_{\boldsymbol{\theta}} J_{\mu}^{i}\right|^{i}}$. Notice that, although the resulting update may not follow the steepest ascent direction, since $\alpha_t^i \geq 0$, the algorithm has the same convergence guarantees as the standard policy gradient algorithm.

6 Numerical results

In this section, we empirically evaluate the performance of the three different strategies for choosing the step size introduced in the previous section. The performance is compared against a recently proposed algorithm (Pirotta et al. 2013), that we name Adaptive Gradient Step-Size and will be denoted by $AGSS$. As far as we know, it is the most related algorithm available in RL literature that investigates the choice of the step size. Pirotta et al. (2013) determine the step size by maximizing a lower bound to the expected performance gain. In the case of Gaussian policy model, the bound is a second-order polynomial of the step-size parameter that can be easily maximized in both exact and approximate scenarios. In the exact case, the optimal step size, that is, the value that maximizes a simplified version of the lower bound, is given by:

$$\alpha_{AGSS}^{*} = \frac{(1-\gamma)^3\sqrt{2\pi}\sigma^3 \left\|\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}}\right\|_2^2}{\left(\gamma\sqrt{2\pi}\sigma + 2(1-\gamma)|\mathcal{A}|\right) RM_{\phi}^2 \left\|\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}}\right\|_1^2}, \tag{23}$$

where M_ϕ is the maximum absolute value of the basis functions, i.e., $\forall s \in \mathcal{S}, \forall i |\phi_i(s)| \leq M_\phi$. This step size method ensures a monotonic policy performance improvement. Moreover, Pirotta et al. (2013) have empirically shown that $AGSS$ is more robust than standard step-size selection mechanisms (e.g., constant or time varying) to changes of the MDP parameters. To make the comparison fair, in the following experiments we use a Gaussian policy model as in (Pirotta et al. 2013). This choice is compatible with the Assumptions 2 and 3.

Before showing numerical results about the learning performance of the proposed algorithms, we can state some considerations about computational times. While single step–size algorithms have a very low computational complexity, MSS–MLC has a complexity that is cubic in the number of policy parameters d. The computational gap between single and multi–step algorithms increases as the size of the problem increases. Nonetheless the complexity is dominated by the sampling procedure. In Table 1, we compare the per–iteration times needed to compute the step size using SSS–SLC and MSS–MLC in the Mass–Spring–Damper domain using different parameterizations. Tests are performed on a standard laptop using a C++ implementation. We employed a state–dependent linear policy parametrization in which we have changed the degree of the polynomial. As shown by the experiments the complexity of MSS–MLC is higher than the one of SSS–SLC but it is dominated by the data collection phase, which, in turn, is 4 order of magnitude less than the time that would be required in the real-world scenario. In summary, the time spent for the computation of the step size is negligible w.r.t. the time needed for data collection.

Table 1 Time Complexity. Given the gradient and the Lipschitz constants, the time required by SSS–SLC and MSS–MLC to compute the step size are reported in table. The number of policy parameters have been changed by considering different degrees of the polynomial: 1, 5 and 10, respectively. In addition, the time required to collect the samples is reported both in simulated and real-world scenario

Parameters	SSS–SLC	MSS–MLC	Data collection time	
			Simulated	Real
3	$1.2 \cdot 10^{-6} s$	$9.4 \cdot 10^{-5} s$	$4.4 \cdot 10^{-1} s$	$2.5 \cdot 10^{3} s$
21	$1.6 \cdot 10^{-6} s$	$2.2 \cdot 10^{-3} s$		
66	$1.9 \cdot 10^{-6} s$	$2.0 \cdot 10^{-2} s$		

6.1 Linear-quadratic gaussian regulator

The first scenario is a discrete-time linear-quadratic Gaussian (LQG) regulator as described by Peters and Schaal (2008b). The LQG problem is defined by the following dynamics:

$$s_{t+1} = A(s_t + a_t); \quad a_t \sim \mathcal{N}(\boldsymbol{\theta} \cdot s_t, \sigma^2); \quad r_t = -B(s_t^2 + a_t^2),$$

where s_t and a_t are scalars, and B is 0.5. The range of the state space is bounded to the interval $[-2, 2]$ and the initial state is drawn uniformly at random in the same interval. The scenario is particularly instructive since it allows to easily compute all the Lipschitz constants.

The objective of this test is to show how changes in the model parameters affect the algorithm performance. In particular, we analyze the impact of changes in the transition model (i.e., A), discount factor, and standard deviation of the Gaussian policy. While the values of the discount factor and of the standard deviation of the policy are exploited by all the algorithms, the change in the transition model (that influences the Lipschitz constant $L_{\mathcal{P}}$) is exploited only by Lipschitz approaches. Notice that, since there is a single policy parameter ($d = 1$), the proposed algorithms lead to the same result. For what concerns $AGSS$, in this case its step size is constant over iterations since the $L2$-norm equals the $L1$-norm for scalar values. It is also easy to modify the Lipschitz constant $L_{\mathcal{R}}$ associated to the reward model by changing the value of B. However, any change of such value does not alter the algorithm performance since it is associated to an equal change in the gradient magnitude.

Table 2 shows how the changes in the parameters influence the number of iterations required to learn a near-optimal value of the policy parameter. The variability of the results comes from the estimation of the gradient through GPOMDP algorithm. As $L_{\mathcal{P}}$ decreases the model becomes more and more smooth, so that the proposed approach has a significant advantage w.r.t. $AGSS$. For what concerns the standard deviation of the policy, algorithm performances degrade as the policy gets more deterministic. Note that the step length are inversely related to M_ϕ^i that grows as the policy becomes more deterministic (see Assumption 3). Lipschitz approach is less influenced by changes in the policy than $AGSS$ (it depends quadratically on M_ϕ) as shown by the increased performance gap. This is a positive effect due to the choice of considering the Kantorovich distance instead of simpler metrics between distributions, like the total variation one.

Although the Lipschitz approach outperforms the $AGSS$ algorithm in most of the settings, it cannot be applied to domains with large values of γ and $L_{\mathcal{P}}$ since the MDP needs to be a contraction w.r.t. the Kantorovich distance (i.e., $\gamma L_{\mathcal{P}}(1 + L_{\pi_\theta}) < 1$).

Table 2 LQR domain. Table reports the number of iterations occurred before reaching a 0.01 approximation of the optimal policy parameter. Results are averaged over 10 runs

σ	γ	$L_{\mathcal{P}} = 0.3$		$L_{\mathcal{P}} = 0.5$		$L_{\mathcal{P}} = 0.7$	
		LIP	AGSS	*LIP*	AGSS	*LIP*	AGSS
0.50	0.5	**896** $\pm$ 18	1932 $\pm$ 25	**1430** $\pm$ 15	2637 $\pm$ 12	**1724** $\pm$ 12	2565 $\pm$ 16
0.50	0.7	**3098** $\pm$ 36	7204 $\pm$ 52	**4442** $\pm$ 18	8367 $\pm$ 27	**5809** $\pm$ 19	6988 $\pm$ 22
0.50	0.9	**25676** $\pm$ 2853	114164 $\pm$ 207	**36294** $\pm$ 68	103176 $\pm$ 119	**65605** $\pm$ 86	66381 $\pm$ 50
0.75	0.5	**368** $\pm$ 7	500 $\pm$ 56	**515** $\pm$ 4	686 $\pm$ 7	**636** $\pm$ 5	677 $\pm$ 6
0.75	0.7	**1188** $\pm$ 8	2131 $\pm$ 24	**1435** $\pm$ 7	2068 $\pm$ 8	1770 $\pm$ 5	**1633** $\pm$ 4
0.75	0.9	**9643** $\pm$ 45	33195 $\pm$ 47	**9281** $\pm$ 13	23089 $\pm$ 25	14167 $\pm$ 9	**13086** $\pm$ 13
1.00	0.5	**171** $\pm$ 4	199 $\pm$ 4	**208** $\pm$ 2	219 $\pm$ 2	296 $\pm$ 2	**250** $\pm$ 2
1.00	0.7	**502** $\pm$ 5	761 $\pm$ 6	**529** $\pm$ 2	640 $\pm$ 2	729 $\pm$ 2	**578** $\pm$ 1
1.00	0.9	**3546** $\pm$ 14	11273 $\pm$ 21	**2788** $\pm$ 5	6409 $\pm$ 6	4652 $\pm$ 3	**4216** $\pm$ 3

Bold values are used to denote the best algorithm in each experimental setting

6.2 Mass-spring-damper

The second scenario is a simple mechanical system described by a mass, a linear spring and a damper (Ammar and Taylor 2012). The objective is to reach a desired state s_g by controlling the external force applied to the system. The system dynamic is described by the following equation:

$$m\ddot{x} + c\dot{x} + kx = F,$$

where m is the mass, c is the viscous friction coefficient ($\frac{N \cdot s}{m}$), k is the spring constant ($\frac{N}{m}$) and F is the external force. Let $\omega_0 = \sqrt{\frac{k}{m}}$ and $\zeta = \frac{c}{2\sqrt{km}}$ be the natural frequency and the damping ratio, respectively. Notice that, since in the following $0 < \zeta < 1$, the free system ($F = 0$) is under-damped and will oscillate with a frequency equal to $\omega_d = \omega_0\sqrt{1-\zeta^2}$.

The continuous state space ($\mathcal{S} \in \mathbb{R}^2$) is defined by the position and velocity of the mass ($s = [x, \dot{x}]^{\mathrm{T}}$) and the continuous scalar action a defines the external force applied to the system ($F = b \cdot a$). The control decision is performed every $0.1s$, with the goal to bring and keep the mass to the desired state $s_g = [0.5, 0]^{\mathrm{T}}$. At every time step the agent receives a reward that is proportional to the distance from the goal position x_g and to the control action:

$$r_{t+1} = -w_d \left| x_t - x_g \right| - w_a \left| b \cdot a_t \right|,$$

where, in all the experiments, we set b, w_d, w_a to 10, 1 and 1/20, respectively. Discount factor γ is set to 0.9. Note that the Lipschitz constant associated to the reward function $L_{\mathcal{R}}$ is 1.

Since the problem is continuous we exploited a Gaussian policy model

$$\pi^{\boldsymbol{\theta}}(a|s) = \mathcal{N}\left(\phi(s)^{\mathrm{T}}\boldsymbol{\theta}, \sigma\right), \tag{24}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\phi : \mathcal{S} \to \mathbb{R}^d$ are the basis functions. The standard deviation σ has been fixed to 0.5 in all the experiments. Three polynomial basis functions have been used to approximate the policy mean action: $\phi(s) = [1, x, \dot{x}]^{\mathrm{T}}$. The initial setting is $x_0 = Unif\left(\{-1, 1\}\right)$, $\dot{x}_0 = 0$ and $\boldsymbol{\theta}_0 = \mathbf{0}$. The learning task was performed using GPOMDP with optimal baseline (Peters and Schaal 2008b), exploiting 500 episodes by 50 steps for each gradient estimate.

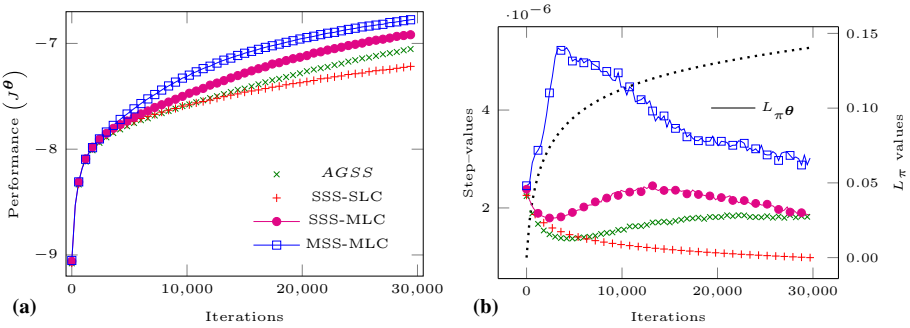

Fig. 2 Mass-Spring-Damper domain with configuration c_1. **a** The score $J_\mu^{\boldsymbol{\theta}}$ as a function of iterations. **b** The step sizes over iterations. For MSS-MLC algorithm, the $L2$-norm of the vector $\boldsymbol{\alpha}$ is drawn. Results are averaged over 100 runs. For sake of readability, the graphs do not display the (very small) confidence intervals

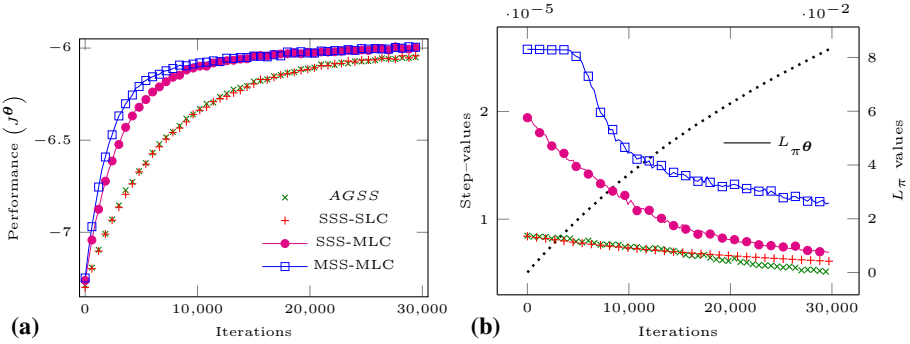

Fig. 3 Algorithm behaviours for the Mass-Spring-Damper domain with configuration c_2. The same structure of Fig. 2 is used here

As done before, the objective is to show how the changes in the parameters affect the algorithm behaviour. In this scenario we concentrate the attention on the Lipschitz continuity of the model by changing the $L_\mathcal{P}$ constant. We defined two configurations: $c_1 = \{m = 0.5, k = 15, c = 0.001\}$ and $c_2 = \{m = 0.1, k = 15, c = 5\}$ with $L_\mathcal{P}$-constants equal to 0.79 and 0.91, respectively. In both the configurations the position, velocity and action were limited in $[-1, 1]$, $[-2, 2]$ and $[-3, 3]$, respectively.

Figures 2a and 3a show that the knowledge of the individual Lipschitz constants can be successfully exploited in order to speed up the learning process. In particular, the algorithms that exploit the individual Lipschitz constants (SSS-MLC and MSS-MLC) outperform both SSS-SLC and $AGSS$ algorithms in all the experiments because they are able to exploit larger learning step for each gradient component. Notice that in the c_1 configuration $AGSS$ learns faster than SSS-SLC, while in the c_2 configuration, characterized by a smaller $L_\mathcal{P}$ value, the performances of the two algorithms get similar. As expected, MSS-MLC outperforms SSS-MLC, thanks to the possibility of updating the policy parameters outside the gradient direction. In particular, the advantage of MSS-MLC on SSS-MLC increases as the angle between the gradient direction and the vector of the Lipschitz constants gets larger. When such angle is zero, the two algorithms lead to the same result.

6.3 Ship steering with water current

In this scenario we adapted a standard control problem where the task is to steer a ship, which is cruising at constant speed, to a goal in minimum time (Rosenstein and Barto 2004). The task is made not trivial by the presence of different water currents around the goal. The continuous state and action spaces are described by the 2-dimensional ship position ($\mathcal{S} \in \mathbb{R}^2$) and the scalar heading ($\mathcal{A} \in \mathbb{R}$), respectively. The following equations describes the continuous motion of the ship:

$$\dot{x} = C\left(\cos\left(\omega + \omega_0\right) - y\right), \quad \dot{y} = C\sin\left(\omega + \omega_0\right)\omega$$

where $s = [x, y]^{\mathrm{T}}$ denotes the state, C is the ship velocity and ω is the ship heading. Notice that the water current influences only the horizontal coordinate x, and its magnitude is proportional to the vertical position y. We changed the problem in order to minimize the travelled distance, by defining the immediate reward as the minimum distance from the goal region:

$$r_{t+1} = \min(g_r - \left\|s_t - s_g\right\|_2, 0),$$

where $s_g = [0, 0]^{\mathrm{T}}$ is the goal position and $g_r = 0.2Km$ is the goal radius. The effects of a real ship inertia and resistance to water are not modelled, i.e., there is no time lag between changes in the desired heading and the actual one. Control decisions were taken every $25s$, with an immediate change of the ship heading. The ship starts at $s_0 = [2.5, -1]^{\mathrm{T}}$.

The policy model is a standard Gaussian model described in Eq. (24) with fixed standard deviation $\sigma = 0.6$. Since the optimal policies for the objectives are not linear in the state variable, a Gaussian radial basis approximation was used:

$$\phi(s) = \left\{\mathcal{N}\left(s; c_i, \Sigma_i\right)\right\}_{i=1}^{3}.$$

The shape and position of the Gaussian basis are defined by the following parameters:

$$\begin{aligned} c_1 &= [2.5, 0]^{\mathrm{T}}, & \Sigma_1 &= diag\,(10, 5) \\ c_2 &= [1.25, 1]^{\mathrm{T}}, & \Sigma_2 &= diag\,(3, 1) \\ c_3 &= [0, 0]^{\mathrm{T}}, & \Sigma_3 &= diag\,(10, 5) \end{aligned}$$

State and action variables are bounded: $x \in [-2, 6]$, $y \in [-2, 2]$ and $a \in [-\pi, \pi]$. As done in the previous domains, we exploited GPOMDP with baseline with 500 episodes for the gradient estimate. Due to the low discount factor ($\gamma = 0.6$), we performed only 30 steps for each episode.

We developed two different scenarios that differ for the reference angle ω_0: $\pi/2$ and 0, respectively. Since the ship heading points to north when $\omega_0 = \pi/2$ and $\omega = 0$, the learning process is easier in the former configuration then in the latter one. Notice that the initial policy, that is the same in both the configurations, is defined by zero-weights, that is, in the first iteration $\omega \sim \mathcal{N}(0, 0.6)$. In contrast, in the second configuration the agent must learn to steer of 90°, moving the ship heading from east to north, but this may be critical under Lipschitz assumptions. We will deeply discuss the second scenario later in the section.

Concerning the first configuration ($\omega_0 = \pi/2$), we have tested two different ship velocities ($C_1 = 0.01^{Km/s}$ and $C_2 = 0.005^{Km/s}$) that lead to $L_{\mathcal{P}}$ constants of 1.13 and 1.06. The Lipschitz constant $L_{\mathcal{R}}$ is 1. As the reader may notice from the $L_{\mathcal{P}}$ constants, the transition model is no more a contraction as in the previous domains. While the overall performance of Lipschitz algorithms does not seem to be affected by the change in the model transitions,

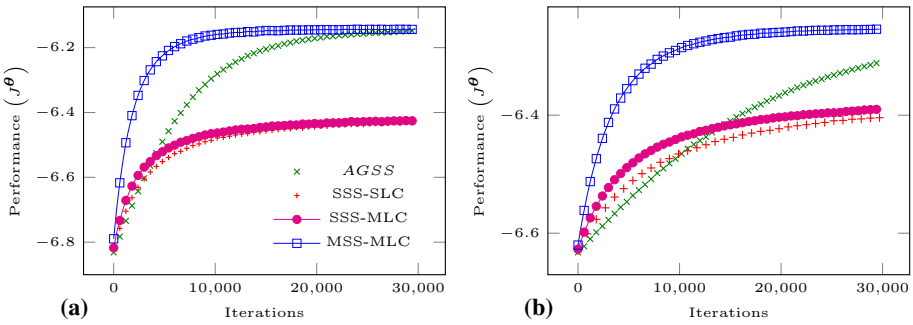

Fig. 4 Score $J_{\mu}^{\boldsymbol{\theta}}$ as a function of iterations. *Error bars* represent the samples standard deviation. The underlying domain consists in the ship steering with reference angle $\omega_0 = \pi/2$ and ship speed of $0.01 Km/s$ (**a**) and $0.005 Km/s$ (**b**), respectively

the $AGSS$ algorithm is no more able to mimic the performance of the best algorithms. This is a direct consequence of the smaller gradient estimates due to slow dynamics. Moreover, as shown in Fig. 4b, the $AGSS$ algorithm is negatively affected by changes in the transition model. In particular, the performance gap between the best algorithms and $AGSS$ increases as the model becomes smoother because the latter approach is not able to exploit such information.

Concerning the second scenario ($\omega_0 = 0$), the $AGSS$ algorithm reaches a better performance than all the Lipschitz algorithms when $C = 0.01^{Km/s}$, see Fig. 5a. As mentioned before, the main issue is the high coefficient required on the first component of the basis functions in order to move the shift heading from east to north in the initial phase of the trial. Although the final policy is quite smooth (the change in the action value is slow and continuous), in the initial phase of the learning process, the first component of the policy is the term that guarantees the highest improvement, thus, it is characterized by large gradient values and parameter steps, that lead to high rate of increase of the Lipschitz constant. As the parameter increases, the policy becomes less Lipschitz (i.e., the L_{π} increases) and the system becomes less contractive, as shown in Fig. 5b. As a consequence, the learning step decreases over iterations at the same rate of increase of the Lipschitz constants.

However, when the ship moves at slow speed, the Lipschitz algorithm are able to exploit the improved smoothness of the system through larger step sizes. On the other hand, the $AGSS$ algorithm exhibits a surprisingly slow learning behavior. The reasons appear clear comparing Figs. 5b, d. The improved Lipschitz continuity of the model directly influences the computation of the step size, leading to larger movements in the initial phase of the learning. At the same time, the rate of increase of the Lipschitz constant of the policy is slower than in the previous setting. Notice that, no changes are visible in the trend of the step sizes shown by the $AGSS$ algorithm that is almost constant over iterations.

7 Conclusions

In this paper we have studied how to automatically set the step-size parameter in policy gradient algorithms under assumptions on the Lipschitz continuity of the MDP and the policy model. We have shown that under such continuity assumptions both the expected return and the policy gradient are Lipschitz in the policy-parameter space and we have derived Lipschitz constants for each component of the gradient. On the basis of such constants, we have

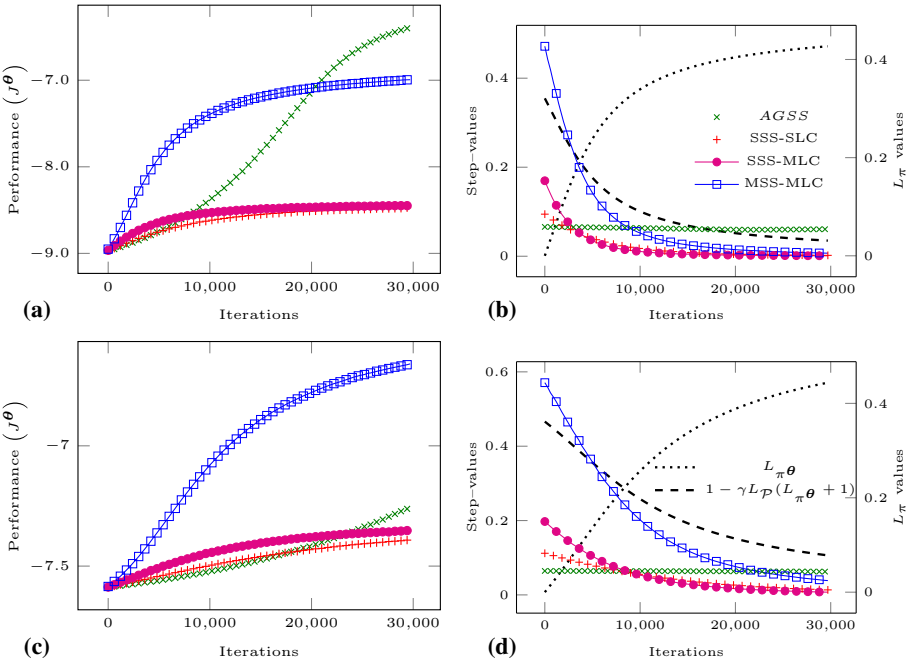

Fig. 5 Algorithm behaviors in the ship domain with $\omega_0 = 0$. While the upper figures refer to the first configuration ($C = 0.01Km/s$), lower figures report the results for $C = 0.005Km/s$. **a**, **c** The performance $J_{\mu}^{\boldsymbol{\theta}}$ as function of the iterations. **b**, **d** The step sizes ($L2$-norm of the vector for MSS-MLC algorithm) and changes in the Lipschitz continuity properties over iterations. In particular, the L_π constant and the contractive property of the system ($1 - \gamma L_{\mathcal{P}}(1 + L_{\pi^{\boldsymbol{\theta}}})$) are shown

proposed to update the policy parameters according to three different learning strategies that guarantee a performance improvement at each iteration. As shown by empirical evaluations, when the MDP has strong continuity properties (i.e., small Lipschitz constants), the proposed approach can take advantage of this by making larger step sizes than the ones made by related approaches that do not exploit such information.

The main drawback of the proposed approach is that it can be applied only when the state transition dynamics of the MDP are a contraction (i.e., $\gamma L_{\mathcal{P}}(1+L_{\pi^{\boldsymbol{\theta}}}) < 1$). To overcome such limitation, it would be interesting to study the effect of replacing the Kantorovich distance with other metrics between distributions. Another way to remove the limitation consists in combining the proposed approach with the $AGSS$ (Pirotta et al. 2013) (e.g., by simply taking at each iteration the step size that guarantees the largest improvement). Finally, future research could address how to use the proposed approach in problems where the Lipschitz constants are unknown and need to be estimated from data.

Appendices

Proof of Proposition 1

Given an $L_{\mathcal{R}}$-LC MDP, for any pair of stationary policies corresponding to parameters $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, the absolute difference between the performance of policy $\pi^{\boldsymbol{\theta}}$ and policy $\pi^{\widehat{\boldsymbol{\theta}}}$ can be bounded as follows:

$$\begin{aligned}\left|J_\mu^{\boldsymbol{\theta}} - J_\mu^{\widehat{\boldsymbol{\theta}}}\right| &= \frac{1}{1-\gamma}\left|\mathop{\mathbb{E}}_{(s,a)\sim\zeta_\mu^{\boldsymbol{\theta}}}[\mathcal{R}(s,a)] - \mathop{\mathbb{E}}_{(s,a)\sim\zeta_\mu^{\widehat{\boldsymbol{\theta}}}}[\mathcal{R}(s,a)]\right| \\ &= \frac{1}{1-\gamma}\left|\iint_{\mathcal{SA}}\left(\zeta_\mu^{\boldsymbol{\theta}}(s,a) - \zeta_\mu^{\widehat{\boldsymbol{\theta}}}(s,a)\right)\mathcal{R}(s,a)\mathrm{d}a\mathrm{d}s\right| \\ &\leq \frac{L_{\mathcal{R}}}{1-\gamma}\sup_f\left\{\left|\iint_{\mathcal{SA}} f\,\mathrm{d}\left(\zeta_\mu^{\boldsymbol{\theta}} - \zeta_\mu^{\widehat{\boldsymbol{\theta}}}\right)\right| : \|f\|_L \leq 1\right\} \\ &= \frac{L_{\mathcal{R}}}{1-\gamma}\mathcal{K}\left(\zeta_\mu^{\boldsymbol{\theta}}, \zeta_\mu^{\widehat{\boldsymbol{\theta}}}\right)\end{aligned}$$

□

Proof of Lemma 5

Under Assumptions 1, 2 and 3, the i-th component of $\boldsymbol{\eta}^{\boldsymbol{\theta}}$ is $L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}}$-LC w.r.t. the state-action space, that is: $\forall(s,\widehat{s},a,\widehat{a}) \in \mathcal{S}^2\times\mathcal{A}^2$,

$$\begin{aligned}\left|\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) - \boldsymbol{\eta}_i^{\boldsymbol{\theta}}(\widehat{s},\widehat{a})\right| &= \left|\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a|s)Q^{\boldsymbol{\theta}}(s,a) - \nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(\widehat{a}|\widehat{s})Q^{\boldsymbol{\theta}}(\widehat{s},\widehat{a})\right| \\ &= \left|\left(\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a|s) - \nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(\widehat{a}|\widehat{s})\right)Q^{\boldsymbol{\theta}}(s,a)\right. \\ &\quad \left. + \left(Q^{\boldsymbol{\theta}}(s,a) - Q^{\boldsymbol{\theta}}(\widehat{s},\widehat{a})\right)\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(\widehat{a}|\widehat{s})\right| \\ &\leq \sup_{s',a'}\left|Q^{\boldsymbol{\theta}}(s',a')\right|\left|\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a|s) - \nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(\widehat{a}|\widehat{s})\right| \\ &\quad + \sup_{s',a'}\left|\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a'|s')\right|\left|Q^{\boldsymbol{\theta}}(s,a) - Q^{\boldsymbol{\theta}}(\widehat{s},\widehat{a})\right| \\ &\leq \left(\frac{R}{1-\gamma}L^i_{\nabla\log\pi^{\boldsymbol{\theta}}} + M^i_{\boldsymbol{\theta}}L_{Q^{\boldsymbol{\theta}}}\right)d_{\mathcal{SA}}\left((s,a),(\widehat{s},\widehat{a})\right).\end{aligned}$$

□

Proof of Lemma 6

For any pair of stationary policies corresponding to $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, the absolute difference of the i-th component of functions $\boldsymbol{\eta}^{\boldsymbol{\theta}}$ and $\boldsymbol{\eta}^{\widehat{\boldsymbol{\theta}}}$ is upper bounded by: $\forall(s,a)\in\mathcal{S}\times\mathcal{A}$, $\forall(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}})\in\Theta^2$,

$$\begin{aligned}\left|\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) - \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\theta}}}(s,a)\right| &= \left|\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a|s)Q^{\boldsymbol{\theta}}(s,a) - \nabla_{\boldsymbol{\theta}_i}\log\pi^{\widehat{\boldsymbol{\theta}}}(a|s)Q^{\widehat{\boldsymbol{\theta}}}(s,a)\right| \\ &= \left|\left(\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a|s) - \nabla_{\boldsymbol{\theta}_i}\log\pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right)Q^{\widehat{\boldsymbol{\theta}}}(s,a)\right. \\ &\quad \left. + \left(Q^{\boldsymbol{\theta}}(s,a) - Q^{\widehat{\boldsymbol{\theta}}}(s,a)\right)\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a|s)\right| \\ &\leq \sup_{s',a'}\left|Q^{\boldsymbol{\theta}}(s',a')\right|\left|\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a|s) - \nabla_{\boldsymbol{\theta}_i}\log\pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right| \\ &\quad + \sup_{s',a'}\left|\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a'|s')\right|\left|Q^{\boldsymbol{\theta}}(s,a) - Q^{\widehat{\boldsymbol{\theta}}}(s,a)\right|\end{aligned}$$

Proof follows from the application of Assumption 3 and from the bound on the supremum norm of the Q-function. □

Proof of Lemma 7

Under Assumptions 1, 2 and 3, the i-th component of $\boldsymbol{\eta}$ is upper–bounded by: $\forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \forall (\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \in \Theta^2$,

$$\begin{aligned}\left|\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) - \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\theta}}}(s,a)\right| &\leq \frac{R}{1-\gamma}\left|\nabla_{\boldsymbol{\theta}_i}\log\pi^{\boldsymbol{\theta}}(a|s) - \nabla_{\boldsymbol{\theta}_i}\log\pi^{\widehat{\boldsymbol{\theta}}}(a|s)\right| + M_{\boldsymbol{\theta}}^i\left|Q^{\boldsymbol{\theta}}(s,a) - Q^{\widehat{\boldsymbol{\theta}}}(s,a)\right| \\ &\leq \left(\frac{R}{1-\gamma}L^i_{\nabla\log\pi}(\boldsymbol{\theta}) + M^i_{\boldsymbol{\theta}}L_Q(\boldsymbol{\theta})\right) d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right),\end{aligned}$$

where inequalities follow from the combination of Lemma 6 with Theorem 2. □

Proof of Theorem 3

Under Assumptions 1, 2, and 3, the i-th component of the gradient $\nabla_{\boldsymbol{\theta}} J$ of the expected return is $L^i_{\nabla J}(\boldsymbol{\theta})$-PLC, that is: $\forall (\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \in \Theta^2$,

$$\begin{aligned}\left|\nabla_{\boldsymbol{\theta}_i} J_{\mu}^{\boldsymbol{\theta}} - \nabla_{\boldsymbol{\theta}_i} J_{\mu}^{\widehat{\boldsymbol{\theta}}}\right| &\leq \frac{1}{1-\gamma}\left|\mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\boldsymbol{\theta}}}\left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a)\right] - \mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}}\left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a)\right]\right| \\ &\quad + \frac{1}{1-\gamma}\left|\mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}}\left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) - \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\theta}}}(s,a)\right]\right| \\ &= \frac{L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}}}{1-\gamma}\left|\iint_{\mathcal{SA}} \frac{\boldsymbol{\eta}_i^{\boldsymbol{\theta}}}{L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}}} \mathrm{d}\left(\zeta_{\mu}^{\boldsymbol{\theta}} - \zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right)\right| + \frac{1}{1-\gamma}\left|\mathop{\mathbb{E}}_{(s,a)\sim\zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}}\left[\boldsymbol{\eta}_i^{\boldsymbol{\theta}}(s,a) - \boldsymbol{\eta}_i^{\widehat{\boldsymbol{\theta}}}(s,a)\right]\right| \\ &\leq \frac{L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}}}{1-\gamma}\mathcal{K}\left(\zeta_{\mu}^{\boldsymbol{\theta}}, \zeta_{\mu}^{\widehat{\boldsymbol{\theta}}}\right) + \frac{L^i_{\boldsymbol{\eta}}(\boldsymbol{\theta})}{1-\gamma} d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right) \\ &\leq \frac{1}{1-\gamma}\left(L^i_{\boldsymbol{\eta}^{\boldsymbol{\theta}}} L_{\zeta}(\boldsymbol{\theta}) + L^i_{\boldsymbol{\eta}}(\boldsymbol{\theta})\right) d_{\Theta}\left(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}\right).\end{aligned}$$

The proof follows by combining Theorem 5 and Theorem 7 with Proposition 2. □

Proof of concavity of function (22)

For sake of readability we report here Eq. (22):

$$\Delta J_{\mu}^{\boldsymbol{\theta}}(\mathbf{A}) = J_{\mu}^{\boldsymbol{\theta}_{t+1}} - J_{\mu}^{\boldsymbol{\theta}_t} \geq \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}{}^{\mathrm{T}} \mathbf{A} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} - \frac{1}{2}\left\|\mathbf{A}\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}\right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \mathbf{A}\left|\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}\right|.$$

We need to prove that $\Delta J_{\mu}^{\boldsymbol{\theta}}(\mathbf{A})$ is a concave function:

$$\begin{aligned}&\Delta J_{\mu}^{\boldsymbol{\theta}}(\lambda\mathbf{A} + (1-\lambda)\mathbf{B}) = \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}{}^{\mathrm{T}} (\lambda\mathbf{A} + (1-\lambda)\mathbf{B}) \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \\ &-\frac{1}{2}\left\|(\lambda\mathbf{A} + (1-\lambda)\mathbf{B}) \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}\right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} (\lambda\mathbf{A} + (1-\lambda)\mathbf{B})\left|\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}\right| \\ &\geq \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}{}^{\mathrm{T}} (\lambda\mathbf{A} + (1-\lambda)\mathbf{B}) \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \\ &-\frac{1}{2}\left(\lambda\left\|\mathbf{A}\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}\right\|_p + (1-\lambda)\left\|\mathbf{B}\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}\right\|_p\right) \\ &\quad \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} (\lambda\mathbf{A} + (1-\lambda)\mathbf{B})\left|\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}\right|\end{aligned}$$

$$
\begin{aligned}
&= \lambda \left({\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}}^{\mathrm{T}} \mathbf{A} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} - \frac{\lambda}{2} \left\| \mathbf{A} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \mathbf{A} \left| \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right| \right) \\
&+ (1-\lambda) \left({\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}}^{\mathrm{T}} \mathbf{B} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} - \frac{1-\lambda}{2} \left\| \mathbf{B} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \mathbf{B} \left| \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right| \right) - \frac{1}{2} \lambda \\
&(1-\lambda) \left(\left\| \mathbf{A} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \mathbf{B} \left| \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right| + \left\| \mathbf{B} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \mathbf{A} \left| \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right| \right) \\
&\geq \lambda \left({\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}}^{\mathrm{T}} \mathbf{A} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} - \frac{\lambda}{2} \left\| \mathbf{A} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \mathbf{A} \left| \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right| \right) \\
&+ (1-\lambda) \left({\nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t}}^{\mathrm{T}} \mathbf{B} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} - \frac{1-\lambda}{2} \left\| \mathbf{B} \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right\|_p \mathbf{L}_{\nabla J}(\boldsymbol{\theta}_t)^{\mathrm{T}} \mathbf{B} \left| \nabla_{\boldsymbol{\theta}} J_{\mu}^{\boldsymbol{\theta}_t} \right| \right) \\
&\geq \lambda \Delta J_{\mu}^{\boldsymbol{\theta}}(\mathbf{A}) + (1-\lambda) \Delta J_{\mu}^{\boldsymbol{\theta}}(\mathbf{B}) \qquad (25)
\end{aligned}
$$

where (25) follows from the Minkowski inequality. □

References

Amari, S., & Douglas, S. (1998). Why natural gradient? In: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE international conference on, vol 2, pp. 1213–1216 vol. 2, doi:10.1109/ICASSP.1998.675489.

Ammar, H., & Taylor, M. (2012). Reinforcement learning transfer via common subspaces. *Adaptive and learning agents, lecture notes in computer science* (Vol. 7113, pp. 21–36). Berlin: Springer.

Armijo, L. (1966). Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, *16*(1), 1–3.

Baxter, J., & Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, *15*, 319–350.

Bertsekas, D. P., & Shreve, S. E. (1978). *Stochastic optimal control: The discrete time case* (Vol. 139). New York: Academic Press.

Deisenroth, M. P., Neumann, G., Peters, J., et al. (2013). A survey on policy search for robotics. *Foundations and Trends in Robotics*, *2*(1–2), 1–142.

Ferns, N., Panangaden, P., & Precup, D. (2005). Metrics for markov decision processes with infinite state spaces. *Proceedings of the twenty-first conference annual conference on uncertainty in artificial intelligence (UAI-05)* (pp. 201–208). Arlington, Virginia: AUAI Press.

Grondman, I., Busoniu, L., Lopes, G. A., & Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *42*(6), 1291–1307.

Hinderer, K. (2005). Lipschitz continuity of value functions in markovian decision processes. *Mathematical Methods of Operations Research*, *62*(1), 3–22.

Kakade, S. (2001). A natural policy gradient. In *Advances in neural information processing systems 14* (Vol. 14, pp. 1531–1538). Vancouver, British Columbia: MIT Press.

Kober, J., & Peters, J. (2008). Policy search for motor primitives in robotics. In *Advances in neural information processing systems 21* (Vol. 21, pp. 849–856). Vancouver, British Columbia: Curran Associates, Inc.

Moré, J. J., & Thuente, D. J. (1994). Line search algorithms with guaranteed sufficient decrease. *ACM Transactions on Mathematical Software*, *20*(3), 286–307.

Peters, J., & Schaal, S. (2006). Policy gradient methods for robotics. In: Intelligent robots and systems, 2006 IEEE/RSJ international conference on, (pp. 2219–2225).

Peters, J., & Schaal, S. (2008a). Natural actor-critic. *Neurocomputing*, *71*(7–9), 1180–1190.

Peters, J., & Schaal, S. (2008b). Reinforcement learning of motor skills with policy gradients. *Neural Networks*, *21*(4), 682–697.

Pirotta, M., Restelli, M., & Bascetta, L. (2013). Adaptive step-size for policy gradient methods. *Advances in Neural Information Processing Systems*, *26*, 1394–1402.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York, NY: Wiley.

Rachelson, E., & Lagoudakis, M.G. (2010). On the locality of action domination in sequential decision making. In: International symposium on artificial intelligence and mathematics.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400–407.

Rosenstein, M. T., & Barto, A. G. (2004). Supervised actor-critic reinforcement learning. In J. Si, A. Barto, W. Powell, & D. Wunsch (Eds.), *Handbook of learning and approximate dynamic programming* (pp. 359–380). John Wiley & Sons, Inc.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *Automatic Control, IEEE Transactions on*, *37*(3), 332–341.

Sutton, R.S., McAllester, D.A., Singh, S.P., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In: *Advances in neural information processing systems 12*, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999], (pp. 1057–1063).

Vlassis, N., Toussaint, M., Kontes, G., & Piperidis, S. (2009). Learning model-free robot control by a monte carlo em algorithm. *Autonomous Robots*, *27*(2), 123–130.

Wagner, P. (2011). A reinterpretation of the policy oscillation phenomenon in approximate policy iteration. *Advances in Neural Information Processing Systems*, *24*, 2573–2581.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*, 229–256.