Tight Performance Guarantees of Imitator Policies with Continuous Actions

Davide Maran, Alberto Maria Metelli, Marcello Restelli

Politecnico di Milano
Piazza Leonardo da Vinci, 32 20133, Milan, Italy
{davide.maran, albertomaria.metelli, macrello.restelli}@polimi.it

Abstract

Behavioral Cloning (BC) aims at learning a policy that mimics the behavior demonstrated by an expert. The current theoretical understanding of BC is limited to the case of finite actions. In this paper, we study BC with the goal of providing theoretical guarantees on the performance of the imitator policy in the case of continuous actions. We start by deriving a novel bound on the performance gap based on Wasserstein distance, applicable for continuous-action experts, holding under the assumption that the value function is Lipschitz continuous. Since this latter condition is hardy fulfilled in practice, even for Lipschitz Markov Decision Processes and policies, we propose a relaxed setting, proving that value function is always Hölder continuous. This result is of independent interest and allows obtaining in BC a general bound for the performance of the imitator policy. Finally, we analyze noise injection, a common practice in which the expert's action is executed in the environment after the application of a noise kernel. We show that this practice allows deriving stronger performance guarantees, at the price of a bias due to the noise addition.

1 Introduction

The degree of interaction of the human in the ecosystem of artificial intelligence is progressively becoming more and more prominent (Zanzotto 2019). In this setting, the human plays the role of an expert that, with different tools, interacts with the artificial agents and allows the agent to leverage their knowledge to improve, quicken, and make the learning process more effective (Jeon, Milli, and Dragan 2020).

Imitation Learning (IL, Osa et al. 2018) can be considered one of the simplest forms of interaction between a human and an artificial agent. This kind of interaction is unidirectional since the human expert provides the agent with a set of demonstrations of behavior that is optimal w.r.t. an unknown objective. The agent, on its part, aims to learn a behavior as close as possible to the demonstrated one. Classically, we distinguish between two realizations of IL: Behavioral Cloning (BC, Bain and Sammut 1995) and Inverse Reinforcement Learning (IRL, Arora and Doshi 2021). BC aims at mimicking the *behavior* of the agent by recovering a policy that matches as much as possible the expert's demonstrated behavior. Instead, IRL has the more ambitious goal of reconstructing a *reward* function that justifies the expert's *intent* rather than their behavior. In this sense, IRL is more challenging than BC, as its output, the reward function, is a more powerful tool that succeeds in being deployed even in the presence of a modification of the environment.

Although IL techniques have been successfully applied to a large variety of real-world applications (e.g., Asfour et al. 2008; Geng, Lee, and Hülse 2011; Rozo, Jiménez, and Torras 2013; Likmeta et al. 2021), their theoretical understanding in terms of performance of the imitation policy is currently limited. Recently, in (Xu, Li, and Yu 2020), a first analysis of the error bounds has been provided for BC and Generative Adversarial Imitation Learning (Ho and Ermon 2016). However, these results involve the presence of an f-divergence (Rényi et al. 1961), usually total variation (TV) or KL-divergence, between the expert's policy and the imitator one. Consequently, they are significant only when the action space is finite, while becoming vacuous for experts with continuous actions. To further argue on the limitation of this analysis, consider the case in which BC is reduced to minimize the *mean squared error* (MSE) between the expert's action and the imitator one. Even in this simple scenario, as we shall see, the current analysis based on TV cannot relate MSE with the performance of the imitator policy. This represents a relevant limitation since many of the applications of IL are naturally defined with continuous-actions context.

Original Contributions In this paper, we aim to take a step forward to a more comprehensive theoretical understanding of BC. Specifically, we devise error bounds that relate the performance difference $J^{\pi_E} - J^{\pi_I}$ between the expert's policy π_E and the imitator one π_I to their divergence. Our bounds are based on the Wasserstein distance (Villani 2009) and, for this reason, are meaningful even in the presence of continuous-action spaces (Section 3). Our work contains the following contributions:

1. We will prove a performance bound for standard BC in case of Lipschitz reward-transition for the MDP (see (Rachelson and Lagoudakis 2010)) and Lipschitz continuity of the value function.
2. Since the latter assumption is often violated in practice,[1],

[1] It is well-known that the value function is Lipschitz continuous

we extend the result by only requiring Lipschitzness of the MDP, even if it requires a weaker performance bound. We also show that, the less regularity on the value function, the slower the convergence of BC (Section 4).

3. Finally, we focus on a popular practice employed in imitation learning, i.e., *noise injection* (Laskey et al. 2017a). In this setting, the expert's action, before being executed in the environment, is corrupted with noise to make the imitation process more robust. We show that noise injection allows achieving stronger theoretical guarantees at the price of competing against a noisy expert, which could have a lower performance (Section 5).

In particular, in the second point, we show that the value function of a Lipschitz MDP is always Hölder continuous, with a suitable choice of the exponent depending on the properties of the MDP and policy. This represents a result of independent interest that overcomes a well-known limitation of the Lipschitz continuity of the value function (Rachelson and Lagoudakis 2010; Pirotta, Restelli, and Bascetta 2015), with possible applications outside BC.

2 Preliminaries

In this section, we provide the background (Section 2.1) and the foundations of Markov Decision Processes (Section 2.2).

2.1 Mathematical Background

Notation Let $\mathcal{X}$ be a set and $\mathfrak{F}$ be a σ-algebra over $\mathcal{X}$, we denote with $\mathcal{P}(\mathcal{X})$ the set of probability measures over the measurable space $(\mathcal{X}, \mathfrak{F})$. Let $x \in \mathcal{X}$, we denote the Dirac delta measure centered in x as δ_x. Let $f : \mathcal{X} \to \mathbb{R}$ be a function, we denote the L_∞-norm as $\|f\|_\infty = \sup_{x \in \mathcal{X}} f(x)$ and with $\|f\|_i$ the L_i-norm for $i \in \{1, 2\}$.

Lipschitz Continuity Let $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ be two metric spaces and $L > 0$. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be *L-Lipschitz continuous* (L-LC) if:

$$d_\mathcal{Y}(f(x), f(x')) \leqslant L d_\mathcal{X}(x, x'), \quad \forall x, x' \in \mathcal{X}.$$

We denote the Lipschitz semi-norm of function f as $\|f\|_L = \sup_{x,x' \in \mathcal{X}, x \neq x'} d_\mathcal{Y}(f(x), f(x'))/d_\mathcal{X}(x, x')$. In the real space ($X \subseteq \mathbb{R}^n$), we use the Euclidean distance, i.e., $d_\mathcal{X}(x, x') = \|x - x'\|_2$. For probability measures ($\mathcal{X} = \mathcal{P}(\Omega)$), the most intuitive distance is the *total variation* (TV), defined as:

$$\mathrm{TV}(\mu, \nu) = \sup_{\|f\|_\infty \leqslant 1} \left| \int_\Omega f(\omega)\,(\mu - \nu)\,(\mathrm{d}\omega) \right| \; \forall \mu, \nu \in \mathcal{P}(\Omega)$$

However, with continuous deterministic distributions, the TV takes its maximum value 1 (Figure 1). Thus, we introduce L_1-*Wasserstein* distance (Villani 2009), defined as:

$$\mathcal{W}(\mu, \nu) = \sup_{\|f\|_L \leqslant 1} \left| \int_\Omega f(\omega)(\mu - \nu)(\mathrm{d}\omega) \right| \quad \forall \mu, \nu \in \mathcal{P}(\Omega)$$

It is worth noting that, for deterministic distributions, we have $\mathcal{W}(\delta_x, \delta_{x'}) = d_\mathcal{X}(x, x')$.

Hölder Continuity The notion of Lipschitz continuity is generalized by Hölder continuity. Let $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ be two metric spaces and $L, \alpha > 0$. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be (α, L)-*Hölder continuous* ((α, L)-HC) if:

$$d_\mathcal{Y}(f(x), f(x')) \leqslant L d_\mathcal{X}(x, x')^\alpha, \quad \forall x, x' \in \mathcal{X}.$$

It is worth noting that: (i) Lipschitz continuity is obtained by Hölder continuity for $\alpha = 1$; (ii) only constant functions are $(\alpha, L)-$HC for $\alpha > 1$; (iii) in bounded domains, the higher the value of α the more restrictive the condition.

Convolution Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be two functions, their *convolution* is defined for all $x \in \mathbb{R}^n$ as:

$$(f * g)(x) := \int_{\mathbb{R}^n} f(x - y) g(y) \mathrm{d}y = \int_{\mathbb{R}^n} f(y) g(x - y) \mathrm{d}y.$$

We introduce the following regularity assumption regarding the probability measures.

Definition 1. *A probability measure $\mathcal{L} \in \mathcal{P}(\mathbb{R}^n)$ is L-TV-Lipschitz continuous (L-TV-LC) if:*

$$\mathrm{TV}(\mathcal{L}(\cdot + h), \mathcal{L}(\cdot)) \leqslant L \|h\|_2, \qquad \forall h \in \mathbb{R}^n.$$

Under this assumption, we can prove that the convolution regularizes bounded and possibly irregular functions.

Proposition 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function such that $\|f\|_\infty \leqslant M$, and let $\mathcal{L} \in \mathcal{P}(\mathbb{R}^n)$ be an L-TV-LC probability measure that admits density function $\ell : \mathbb{R}^n \to \mathbb{R}_{\geqslant 0}$. Then, the convolution $f * \ell$ is $2LM$-LC continuous.*

2.2 Markov Decision Processes

A discrete-time discounted Markov Decision Process (MDP, Puterman 2014) is a 6-tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma, \mu)$ where $\mathcal{S}$ and $\mathcal{A}$ are the measurable sets of states and actions, $p : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the transition model that defines the probability measure $p(\cdot|s, a)$ of the next state when playing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function defining the reward $r(s, a)$ upon playing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, $\gamma \in [0, 1)$ is the discount factor, and $\mu \in \mathcal{P}(\mathcal{S})$ is the initial-state distribution. The agent's behavior is modeled by a policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, which assigns a probability measure $\pi(\cdot|s)$ of the action to be taken in state $s \in \mathcal{S}$. When the policy is deterministic, we denote with $\pi(s)$ the action played in state $s \in \mathcal{S}$. A policy determines a γ-*discounted visitation distribution*, defined as: $d^\pi(s) := (1 - \gamma) \sum_{t=0}^{+\infty} \gamma^t \, \mathbb{P}(s_t = s | \pi, \mu)$ for every $s \in \mathcal{S}$.

Value Functions The state-action value function (or *Q-function*) which quantifies the expected discounted sum of the rewards obtained under a policy π, starting from a state $s \in \mathcal{S}$ and fixing the first action $a \in \mathcal{A}$:

$$Q^\pi(s, a) := \mathbb{E}_\pi \left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s, a_0 = a \right], \quad (1)$$

where $\mathbb{E}_\pi$ denotes the expectation w.r.t. to the stochastic process $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim p(\cdot|s_t, a_t)$ for all $t \in \mathbb{N}$. The state value function (or *V-function*) is defined as $V^\pi(s) := \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s, a)]$, for all $s \in \mathcal{S}$. Given an initial state distribution μ, the *expected return* is defined as:

$$J^\pi := \mathop{\mathbb{E}}_{s \sim \mu} [V^\pi(s)] = \frac{1}{1 - \gamma} \mathop{\mathbb{E}}_{s \sim d^\pi, a \sim \pi(\cdot|s)} [r(s, a)].$$

under the demanding assumption that $\gamma L_p(1 + L_\pi) < 1$ (Rachelson and Lagoudakis 2010) requiring the Lipschitz constants of the transition model L_p and of the policy L_π to be very small.

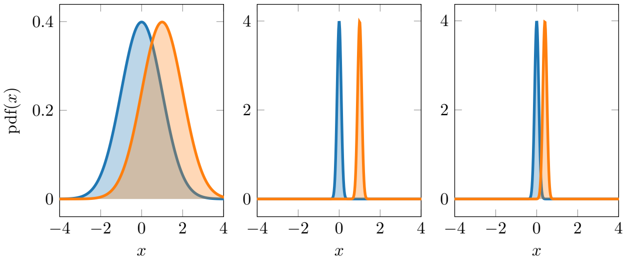

Figure 1: Comparison between TV and Wasserstein distances for two Gaussian distributions μ and ν. Left: $\mathrm{TV}(\mu,\nu) \approx 0.38$, $\mathcal{W}(\mu,\nu) = 1$, Center: $\mathrm{TV}(\mu,\nu) \approx 1$, $\mathcal{W}(\mu,\nu) = 1$, Right: $\mathrm{TV}(\mu,\nu) \approx 1$, $\mathcal{W}(\mu,\nu) = 0.4$

Lipschitz MDPs We now introduce notions that will allow us to characterize the smoothness of an MDP (Rachelson and Lagoudakis 2010). To this end, we assume that the state space $\mathcal{S}$ and action space $\mathcal{A}$ are metric spaces endowed with the corresponding distance functions $d_{\mathcal{S}}$ and $d_{\mathcal{A}}$.

Assumption 1 (Lipschitz MDP)**.** *An MDP $\mathcal{M}$ is (L_p, L_r)-LC if, for all $(s,a), (s',a') \in \mathcal{S} \times \mathcal{A}$ it holds that:*

$$\mathcal{W}(p(\cdot|s,a), p(\cdot|s',a')) \leqslant L_p \left(d_{\mathcal{S}}(s,s') + d_{\mathcal{A}}(a,a') \right),$$
$$\left| r(s,a) - r(s',a') \right| \leqslant L_r \left(d_{\mathcal{S}}(s,s') + d_{\mathcal{A}}(a,a') \right).$$

Assumption 2 (Lipschitz Policy)**.** *A (Markovian stationary) policy π is L_π-LC if, for all $s, s' \in \mathcal{S}$ it holds that:*

$$\mathcal{W}(\pi(\cdot|s), \pi(\cdot|s')) \leqslant L_\pi d_{\mathcal{S}}(s,s').$$

Note, if instead of the Wasserstein metric, we had used the TV, these assumptions would be way more restrictive, not holding for deterministic environment/policies with continuous state-action spaces (Munos and Szepesvári 2008). Under Assumptions 1 and 2, provided that $\gamma L_p(1 + L_\pi) < 1$, the Q-function Q^π is L_Q-LC with $L_Q \leqslant \frac{L_r}{1-\gamma L_p(1+L_\pi)}$ (Rachelson and Lagoudakis 2010, Theorem 1).

3 Bound for Imitating Policies Based on Wasserstein Distance

The high-level goal of this work is to find a theoretical guarantee for the imitator policies learned with BC. Specifically, we want to bound the difference in expected return $J^{\pi_E} - J^{\pi_I}$ between the *imitator* policy π_I learned with BC and the *expert* policy π_E in terms of a distributional divergence between the corresponding action distributions.

The best-known results for this kind of analysis, in the case of *discrete* action spaces, are proved in (Xu, Li, and Yu 2020) and we report it below for completeness.[2]

Theorem 2 (Xu, Li, and Yu (2020), Theorem 1)**.** *Let π_E be the expert policy and π_I be the imitator policy. If $|r(s,a)| \leqslant R_{\max}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, it holds that:*

$$J^{\pi_E} - J^{\pi_I} \leqslant \frac{2R_{\max}}{(1-\gamma)^2} \mathop{\mathbb{E}}_{s\sim d^{\pi_E}} [\mathrm{TV}(\pi_E(\cdot|s), \pi_I(\cdot|s))].$$

As anticipated, this result is not suitable for continuous action spaces, since the TV between different policies would take its maximum value 1 whenever one of the two policies is deterministic. The following example clarifies the issue.

Example 1. *Suppose that the action space is a real space $\mathcal{A} \subseteq \mathbb{R}^n$ and that both expert π_E and the imitator π_I policies are deterministic. A common way to perform BC is to minimize the* mean squared error *(MSE) between the expert's action and the imitator one. Suppose we are able to provide the following guarantee on the MSE, for some $\varepsilon > 0$:*

$$\mathop{\mathbb{E}}_{s\sim d^E} \left[\left\| \pi_E(s) - \pi_I(s) \right\|_2^2 \right] \leqslant \varepsilon^2. \tag{2}$$

However, this condition provides no guarantee in TV. Indeed, by taking $\pi_I(s) = \pi_E(s) + \frac{\varepsilon}{\sqrt{n}} \mathbf{1}_n$, being $\mathbf{1}_n$ the vector of all 1s, Equation (2) is fulfilled, but we obtain: $\mathbb{E}_{s\sim d^{\pi_E}}[\mathrm{TV}(\pi_E(\cdot|s), \pi_I(\cdot|s))] = \mathbb{E}_{s\sim d^{\pi_E}}[\mathbb{1}\{\pi_E(s) \neq \pi_I(s)\}] = 1$, where $\mathbb{1}$ is the indicator function.

3.1 A Bound Based on Wasserstein Distance

Even if the existing analysis of Xu, Li, and Yu (2020) cannot be applied in continuous action spaces, as shown in Example 1, it is not hard to leverage the regularity of the MDP to effectively bound the performance difference $J^{\pi_E} - J^{\pi_I}$.

Theorem 3. *Let π_E be the expert policy and π_I be the imitator policy. If that state-action value function Q^{π_I} of the imitator policy π_I is $L_{Q^{\pi_I}}$-LC, then it holds that:*

$$J^{\pi_E} - J^{\pi_I} \leqslant \frac{L_{Q^{\pi_I}}}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi_E}} [\mathcal{W}(\pi_I(\cdot|s), \pi_E(\cdot|s))].$$

Proof. Using the *performance difference lemma* (Kakade and Langford 2002), we have:

$$J^{\pi_E} - J^{\pi_I} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi_E}} \left[\mathop{\mathbb{E}}_{a\sim\pi_E(\cdot|s)} [A^{\pi_I}(s,a)] \right],$$

where $A^{\pi_I}(s,a) = Q^{\pi_I}(s,a) - V^{\pi_I}(s)$ is the advantage function. The inner expectation can be written as:

$$\begin{aligned} &\mathop{\mathbb{E}}_{a\sim\pi_E(\cdot|s)} [A^{\pi_I}(s,a)] \\ &\qquad = \int_{\mathcal{A}} Q^{\pi_I}(s,a)(\pi_E(\mathrm{d}a|s) - \pi_I(\mathrm{d}a|s)) \\ &\qquad \leqslant \sup_{s\in\mathcal{S}} \left\| Q^{\pi_I}(s,\cdot) \right\|_L \mathcal{W}(\pi_E(\cdot|s), \pi_I(\cdot|s)), \end{aligned}$$

where the inequality follows from the definition of Wasserstein metric. The result is obtained by observing that $\sup_{s\in\mathcal{S}} \left\| Q^{\pi_I}(s,\cdot) \right\|_L \leqslant \left\| Q^{\pi_I} \right\|_L \leqslant L_{Q^{\pi_I}}$. □

A similar bound was previously derived by (Pirotta, Restelli, and Bascetta 2015, Theorem 1) and (Asadi, Misra, and Littman 2018, Theorem 2). However, (Pirotta, Restelli, and Bascetta 2015) assume that the policy is LC w.r.t. a policy parametrization. Instead, the result of (Asadi, Misra, and Littman 2018) involves the transition model instead of the policy and requires a bound uniform over $\mathcal{S} \times \mathcal{A}$ on the Wasserstein distance between the true and the estimated models. Let us now revisit Example 1 in light of Theorem 3.

[2] The result reported in (Xu, Li, and Yu 2020) involves the KL-divergence and is obtained, via Pinsker's inequality, from the one we report that is tighter (Appendix A.2 of Xu, Li, and Yu (2020)).

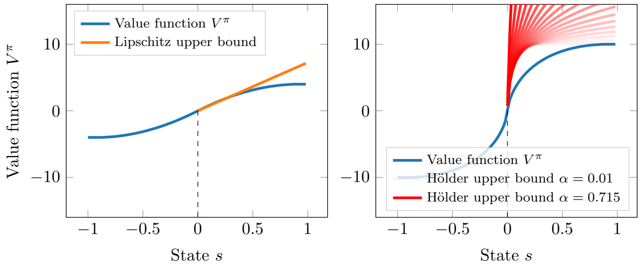

Figure 2: State value functions of Example 2. Left: the bound of (Rachelson and Lagoudakis 2010) hold and it is tight. Right: the bound of (Rachelson and Lagoudakis 2010) does not hold, but our bound based on Hölder continuity holds, for different values of $\alpha \in (0, 1)$.

Example 1 (continued). *Under Equation* (2) *, we can provide an effective guarantee on the Wasserstein distance:*

$$\mathop{\mathbb{E}}_{s\sim d^{\pi_E}}[\mathcal{W}(\pi_E(\cdot|s), \pi_I(\cdot|s))] = \mathop{\mathbb{E}}_{s\sim d^{\pi_E}}[\|\pi_E(s) - \pi_I(s)\|_1]$$

$$\leqslant \mathbb{E}_{s\sim d^{\pi_E}}[\|\pi_E(s) - \pi_I(s)\|_2^2]^{\frac{1}{2}} \leqslant \varepsilon,$$

where in the first inequality, we used Jensen's inequality.

Comparing Theorem 3 with Theorem 2, we no longer require the uniform bound $R_{\max}$ on the reward function, but we introduce an additional assumption on the regularity of the imitator Q-function Q^{π_I}. Clearly, we should find suitable assumptions under which $L_{Q^{\pi_I}}$ is finite. As we anticipated in Section 2.2, the only known result that provides such an estimate under the assumption of Lipschitz MDP and Lipschitz policy with Wasserstein metric is (Rachelson and Lagoudakis 2010), where the authors proved that, if $\gamma L_p(1 + L_\pi) < 1$ is satisfied, L_{Q^π} can be chosen as:

$$L_{Q^\pi} := \frac{L_r}{1 - \gamma L_p(1 + L_\pi)}. \qquad (3)$$

However, we argue that condition $\gamma L_p(1 + L_\pi) < 1$ is very demanding and often unrealistic. Indeed, to fulfill it we need at least one of these conditions to be satisfied:

(i) $\gamma \ll 1$: in practice, it is almost always false, since the discount factor is often chosen to be close to 1;

(ii) $L_p < 1$: this is a very unrealistic assumption, since it would make all the states shrink exponentially when the same actions are performed;

(iii) $L_\pi \approx 0$: the action depends very little on the state so that there is a very limited possibility of controlling the environment (this condition alone is not even sufficient).

3.2 The Tightness of the Value Function Lipschitz Constant

It is legitimate to question whether the value L_{Q^π} of Equation (3), widely employed in the literature (e.g., Rachelson and Lagoudakis 2010; Pirotta, Restelli, and Bascetta 2015; Asadi, Misra, and Littman 2018), is a tight approximation of the Lipschitz semi-norm $\|Q^\pi\|_L$. In this section, we prove that the result cannot be improved, at least when requiring the Lipschitz continuity of the value function. Example 2 shows that the value function Q^π can be made non-LC even when the MDP and the policy are LC, while Theorem 4 proves that a bound like that of Theorem 3 cannot be obtained for a generic Lipschitz MDPs and policies.

Example 2. *Let $\mathcal{M}$ be an MDP and π be a policy defined as follows, given the constants $L_p, L_r > 0$:*

- *$\mathcal{S} = [-1, 1]$;*
- *$\mathcal{A} = \{0\}$;*
- *The dynamic is deterministic. From every state $s \in \mathcal{S}$, performing action 0, the only possible, the environment moves to the state $s' = \operatorname{clip}(L_p s, -1, 1)$.[3] This means that $p(\mathrm{d}s'|s, a) = \delta_{\operatorname{clip}(L_p s, -1, 1)}(\mathrm{d}s')$;*
- *$r(s, a) = L_r s$;*
- *The initial state distribution is $\mu = \operatorname{Uni}([0, 1])$ (not influential for the derivation that follows).*

This MDP is (L_p, L_r)-LC and the policy has Lipschitz constant equal to $L_\pi = 0$, since there is one action only. Equation (3) *ensures that the state value function V^π (that is equal to the state-action value function Q^π since there is one action only) is LC with constant:*

$$L_{V^\pi} = \frac{L_r}{1 - \gamma L_p}.$$

Since the state space is one dimensional, we can compute the state value function V^π exactly:

$$V^\pi(s) = L_r \sum_{k=0}^{+\infty} \gamma^k \operatorname{clip}\big(L_p^k s, -1, 1\big), \qquad \forall s \in \mathcal{S}.$$

As shown in Figure 2 left, the point of maximal slope $s = 0$. Even if we have employed the specific values $L_p = 1.15$, $L_r = 1$, and $\gamma = 0.75$, it is simple to see that this property is valid in general. Moreover, we have plotted in orange the line which passes through the origin, having slope equal to:

$$L_{V^\pi} = \frac{L_r}{1 - \gamma L_p} = \frac{1}{1 - 0.75 \cdot 1.15} \approx 7.27,$$

which is the tangent line to the state value function in $s = 0$, as it also can be found analytically:

$$\frac{\partial V^\pi}{\partial s}(0) = L_r \sum_{k=0}^{+\infty} \gamma^k L_p^k = \frac{L_r}{1 - \gamma L_p}.$$

This means that, in this case, the choice of the Lipschitz constant provided by the theory (Equation 3) is actually tight.

What happens if we reach the hard edge of $\gamma L_p(1+L_\pi) = \gamma L_p > 1$, where Equation (3) *does not guarantee any property? For instance, by taking $L_p = 1.15$, $L_r = 1$, and $\gamma = 0.9$, we lose any Lipschitz property, finding a derivative which is unbounded, as shown in Figure 2 right.*

Note that, in this example, we are able to find a non-LC state value function even in the apparently simple case of $\mathcal{A} = \{0\}$, where $L_\pi = 0$. Therefore, this example also shows that the dynamics of the system alone is enough to make the state value function irregular. Furthermore, the same example can be adapted to prove that, for a generic Lipschitz MDP and a pair of Lipschitz policies, a bound like the one of Theorem 3 cannot be obtained in general.

[3] $\operatorname{clip}(x, a, b)$ is the *clipping* function, i.e., $\max\{\min\{x, b\}, a\}$.

Theorem 4. *There exist an (L_p, L_r)-LC MDP and an L_π-LC policy π such that for every finite constant $C > 0$, (even depending on L_p, L_π, and L_r), there exists an L_π-LC policy π' such that:*

$$J^{\pi} - J^{\pi'} \geqslant C \mathop{\mathbb{E}}_{s \sim d^{\pi}} [\mathcal{W}(\pi(\cdot|s), \pi'(\cdot|s))].$$

The proof is reported in Appendix. If we set $\pi = \pi_E$ as the expert policy and $\pi' = \pi_I$ as an imitator policy, Theorem 4 shows that, even if the MDP and the policies are LC, we cannot, in general, upper bound the performance difference $J^{\pi_E} - J^{\pi_I}$ with the expected Wasserstein distance $\mathbb{E}_{s\sim d^{\pi}}[\mathcal{W}(\pi_E(\cdot|s), \pi_I(\cdot|s))]$. This is in line with the fact that, without additional assumptions, e.g., when $\gamma L_p(L_\pi + 1) < 1$ does not hold, Theorem 4 is vacuous. Therefore, these bounds cannot be improved in the framework of Lipschitz continuity, however, a weaker notion of regularity can be used to generalize the previous theorems.

4 Hölder Continuity Is All We Need

In this section, we propose an approach for overcoming the limitations of the Lipschitz continuity, discussed in the previous section. In Section 4.1, we show that the state-action value function Q^π is always Hölder continuous, provided that the MDP and the policy are LC. Then, in Section 4.2, we apply these findings to BC, deriving a bound on the performance difference $J^{\pi_E} - J^{\pi_I}$ in terms of the Wasserstein distance that holds for *every* LC MDP and policy.

4.1 The Hölder Continuity of the Value Function

The first step to improve the result of (Rachelson and Lagoudakis 2010) is to observe that, like in Example 2, even when the value function is not Lipschitz continuous, it keeps being continuous. This observation is not, in principle, accounted for by the previous analysis, which provides no result when $\gamma L_p(1 + L_\pi) > 1$. This suggests that employing a notion of regularity that is stronger than continuity but weaker than Lipschitz continuity, as Hölder continuity, might lead to an improvement of the analysis. Indeed, we are able to prove the following generalization.

Theorem 5 (Hölder-continuity of the Q-function)**.** *Let $\mathcal{M}$ be an (L_p, L_r)-LC MDP, let π be an L_π-LC policy, and let*

$$0 < \alpha < \overline{\alpha} := \min\left\{1, \frac{-\log\gamma}{\log(L_p(1 + L_\pi))}\right\}.$$

If the state space $\mathcal{S}$ and the action space $\mathcal{A}$ admit finite diameter[4] $\operatorname{diam}(\mathcal{S})$ and $\operatorname{diam}(\mathcal{A})$, respectively, then the state-action value function Q^π is $(\alpha, L_{Q^\pi,\alpha})-$HC with a Hölder constant bounded by:

$$L_{Q^\pi,\alpha} := \frac{L_r\left(\operatorname{diam}(\mathcal{S}) + \operatorname{diam}(\mathcal{A})\right)^{1-\alpha}}{1 - \gamma(L_p(1 + L_\pi))^\alpha}.$$

The proof is reported in Appendix.Furthermore, we can easily obtain the Hölder constant of the state value function V^π.

Proposition 6 (Hölder-continuity of the V-function)**.** *Let π be an L_π-LC policy. If the state-action value function Q^π is $(\alpha, L_{Q^\pi,\alpha})$-HC, then the corresponding state value function V^π is $(\alpha, L_{V^\pi,\alpha})$-HC with:*

$$L_{V^\pi,\alpha} := L_{Q^\pi,\alpha}(L_\pi + 1)^\alpha.$$

These result represent a generalization of those of (Rachelson and Lagoudakis 2010), which are obtained by setting $\alpha = 1$.

Moreover, this Theorem 5 implies that the value functions of an LC MDP and policy is always continuous, since any HC function is also continuous, regardless of its constants, as it seemed from the previous example. Coming back to Example 2, we can perform further analyses.

Example 2 (continued). *We can use Theorem 5 to provide an upper bound on the value function even if the Lipschitz continuity does not hold. The critical exponent is given by:*

$$\overline{\alpha} = -\frac{\log\gamma}{\log(L_p(1 + L_\pi))} \approx 0.72.$$

For every value of $\alpha < \overline{\alpha}$, the state value function V^π is $(\alpha, L_{V^\pi,\alpha})-$HC. As we can see in Figure 2 right, for small α, the bound provided by $L_{V^\pi,\alpha}|s|^\alpha$ is tight for $s \to 1$.

4.2 A More General Bound Based on Wasserstein Distance

Similarly to what we have done in Section 3, to a result of regularity, we are able to associate a result about the loss of BC, bounding the difference in performance between two policies with their Wasserstein distance. Indeed, thank to Theorem 5, we can prove the following bound.

Theorem 7 (Optimal Error Rate for BC)**.** *Let π_E be the expert policy and π_I be the imitator policy. If that state-action value function Q^{π_I} of the imitator policy π_I is $(\alpha, L_{Q^{\pi_I},\alpha})$-HC, then it holds that:*

$$J^{\pi_E} - J^{\pi_I} \leqslant \frac{L_{Q^{\pi_I},\alpha}}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi_E}} \left[\mathcal{W}\left(\pi_E(\cdot|s), \pi_I(\cdot|s)\right)^\alpha\right].$$

Furthermore, if the MDP $\mathcal{M}$ is (L_p, L_r)-LC and the imitator policy π_I is L_{π_I}-LC, the bound is tight for what concerns the exponent α that cannot be improved above the critical value $\overline{\alpha}$ of Theorem 5.

The proof is reported in Appendix.As expected, a low value of α leads to a looser bound. Unfortunately, this bound, despite being tight in the exponent, is difficult to manage in practice. Indeed, in order to minimize the right-hand side, $\mathbb{E}_{s\sim d^{\pi_E}}\left[\mathcal{W}(\pi_E(\cdot|s), \pi_I(\cdot|s))^\alpha\right]$, one should know the value α in advance. However, $\alpha \leqslant \overline{\alpha}$ depends on the Lipschitz constants of the environment and of the policy, which are usually unknown. Therefore, no imitation learning algorithm can be trained to minimize this error explicitly. Fortunately, we can see that, weakening this result, we can obtain a more practical guarantee. Since $0 < \alpha < 1$, we can apply Jensen's inequality to obtain:

$$J^{\pi_E} - J^{\pi_I} \leqslant \frac{L_{Q^{\pi_I},\alpha}}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi_E}} \left[\mathcal{W}\left(\pi_E(\cdot|s), \pi_I(\cdot|s)\right)\right]^\alpha. \quad (4)$$

In this formulation, we minimize the expected Wasserstein distance only, and the knowledge of α is not needed, but its value impacts the kind of guarantee we can provide.

[4] The *diameter* of a metric space $(\mathcal{X}, d_{\mathcal{X}})$ is defined as: $\operatorname{diam}(\mathcal{X}) = \sup_{x,x'\in\mathcal{X}} d_{\mathcal{X}}(x, x')$.

Remark 1. *If we perform BC in a (L_p, L_r)-LC MDP and with a L_{π_I}-LC imitator policy π_I, the best possible performance guarantee (from Equation 4) is given by:*

$$J^{\pi_E} - J^{\pi_I} \leqslant \mathcal{O}(\varepsilon^{\alpha}),$$

where ε is the square root of the imitation MSE, i.e., $\varepsilon^2 = \mathbb{E}_{s\sim d^{\pi_E}}[\|\pi_E(s) - \pi_I(s)\|_2^2]$ as defined in Example 1, and $\alpha < \overline{\alpha} = -\frac{\log \gamma}{\log(L_p(1+L_{\pi_I}))}$, the critical exponent.

Therefore, a very low value of α, corresponding to lack of regularity, can badly influence the possibility of learning a good imitator policy.

5 Noise Injection

BC may struggle when the regularity assumptions are lacking. However, in practice, using a *noisy* expert policy may significantly help the learning process (Laskey et al. 2017b). This empirical benefit is justified by the intuition that noise helps in exploring the neighborhood of the expert trajectories. In this section, we formulate this empirical evidence in a mathematically rigorous way. Indeed, we show how to break the barrier enforced by Theorem 7, whose result is obtained by a deterministic expert. Clearly, these advantages come with the price that a noisy expert might experience a loss in expected return compared to the deterministic one.

5.1 Noise Injection: A Mathematical Formulation

The simplest form of noise injection is realized by adding to the expert's action $a_{E,t}$ a noise component η_t. In particular, assuming that the action space is real, i.e., $\mathcal{A} \subseteq \mathbb{R}^n$, we have:

$$\forall t \in \mathbb{N}: \qquad \begin{cases} a_{t,E} \sim \pi_E(\cdot|s_t) \\ \eta_t \overset{\text{iid}}{\sim} \mathcal{L} \\ a_t = a_{t,E} + \eta_t \end{cases}, \tag{5}$$

where $\{\eta_t\}_{t\in\mathbb{N}}$ is a noise sequence whose components are independent between each other and from the sequences of states and actions, and identically distributed by law $\mathcal{L} \in \mathcal{P}(\mathcal{A})$. If $\mathcal{L}$ admits a density function, we can express the density function of the played action a_t as the convolution of the expert policy density function π_E and the density function ℓ of the noise law $\mathcal{L}$. Note that the formalization in Equation (5) encompasses distributions that do not correspond to the intuitive idea of *noise* (e.g., when $\mathcal{L}$ is a discrete law). To obtain a meaningful result, we enforce the following assumption.

Assumption 3. *The law of the noise $\mathcal{L}$ admits a density function w.r.t. a reference measure $\ell : \mathbb{R}^n \to \mathbb{R}_{\geqslant 0}$ and is TV-LC (see Definition 1) with constant L_ℓ.*

Under this assumption, denoting with $\pi_{E,\ell}$ the policy with noise injection, i.e., $a_t \sim \pi_{E,\ell}(\cdot|s_t)$, we have that:

$$\pi_{E,\ell}(a|s) = \int_{\mathbb{R}^n} \pi_E(a'|s)\ell(a-a')\mathrm{d}a', \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

This represents the convolution of the policy density function π_E and the noise density function ℓ. In other words, this shows that the action taken by the expert policy $a_{E,t}$ is averaged over the noise probability distribution.

Assumption 3 covers the most common types of noise, like the Gaussian or the uniform ones. In fact, we can prove that every univariate unimodal distribution satisfies Definition 1.Considering multivariate Gaussian noise, we directly derive the L_ℓ constant.

Example 3. *Suppose the noise is sampled from a zero-mean Gaussian distribution $\mathcal{N}(0,\Sigma)$ with covariance matrix Σ, the previous integral writes, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$:*

$$\pi_{E,\ell}(a|s) = \int_{\mathbb{R}^n} \pi_E(a'|s) \underbrace{\frac{e^{-\frac{1}{2}(a-a')^T\Sigma^{-1}(a-a')}}{(2\pi)^{n/2}\det(\Sigma)^{1/2}}}_{\ell(a-a')} \mathrm{d}a',$$

where we recognise the Gaussian n-variate density ℓ. Assumption 3 is verified since, for $h \in \mathbb{R}^n$:

$$\begin{aligned} \mathrm{TV}(\mathcal{N}(h,\Sigma),\mathcal{N}(0,\Sigma)) &\leqslant \sqrt{\frac{1}{2}\mathrm{KL}(\mathcal{N}(h,\Sigma),\mathcal{N}(0,\Sigma))} \\ &= \frac{1}{2}\|h\|_{\Sigma^{-1}} \leqslant \frac{1}{2\sqrt{s_{\min}(\Sigma)}}\|h\|_2, \end{aligned}$$

where we used Pinsker's inequality and $s_{\min}(\cdot)$ denoted the minimum singular value of a matrix. In particular, if Σ is diagonal as $\sigma^2 I$, we have that $L_\ell = 1/(2\sigma)$.

It is worth noting that, in the diagonal covariance case, L_ℓ is proportional to σ^{-1}. This suggests that, the smaller the impact of the noise $\mathcal{L}$, i.e., the smaller the standard deviation σ, the higher the constant L_ℓ. Indeed, as σ decreases, the regularization effect of the noise becomes less relevant (in the limit $\sigma \to 0$, noise injection vanishes).

5.2 A Bound Based on Wasserstein Distance for Noise Injection

We are now able to prove a performance guarantee for BC with noise injection. The idea is based on a simple yet interesting fact. We can use the noise to smooth a bounded function, as in Proposition 1. Applying this approach to the state-action value function, leads to the following result.

Theorem 8. *Let π_E be the expert policy and π_I be the imitator policy. Let us suppose that we have injected a noise of density function ℓ, satisfying Assumption 3 to obtain a noisy expert $\pi_{E,\ell}$ and a noisy imitator $\pi_{I,\ell}$. If $|Q^{\pi_I}(s,a)| \leqslant Q_{\max}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, it holds that:*

$$J^{\pi_{E,\ell}} - J^{\pi_{I,\ell}} \leqslant \frac{2L_\ell Q_{\max}}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi_{E,\ell}}}[\mathcal{W}(\pi_E(\cdot|s), \pi_I(\cdot|s))].$$

The proof is reported in Appendix.Some observations are in order. First, note the similarity with Theorem 3, with the only difference being the substitution of L_{Q^π} with $2L_\ell Q_{\max}$. Second, we require no smoothness assumption (e.g., Lipschitz continuity) on the environment or on the policy. Yet, if in the previous result of Theorem 3 the constant L_{Q^π} could easily become infinite, now, the constant $2L_\ell Q_{\max}$ can be easily bounded by $\frac{2L_\ell R_{\max}}{(1-\gamma)^2}$, since $Q_{\max} \leqslant \frac{R_{\max}}{1-\gamma}$. From an intuitive perspective, the need for smoothness in the environment is replaced with an assumption on the density function of the noise. Lastly, we note that on

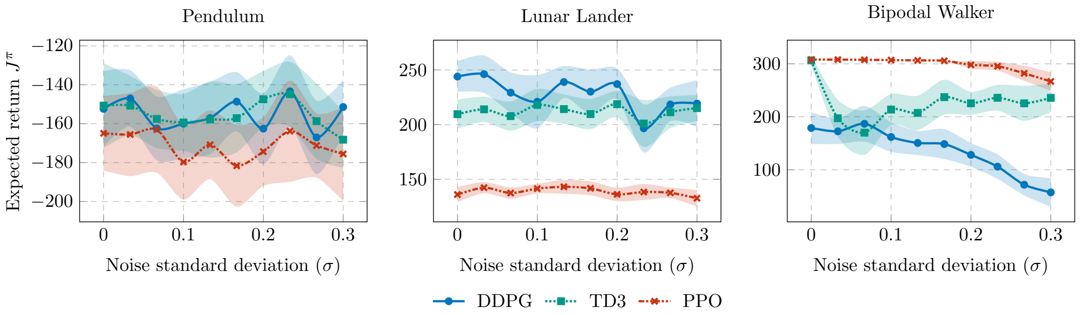

Figure 3: The performance of the expert J^π as a function of the standard deviation of the noise σ. The performance is measured on 40 episodes int environment repeated for 20 different random seeds (nuance represents the 95% non-parametric c.i.).

the right-hand side of the formula, the error is measured by the Wasserstein distance of the non-noisy policies. This is advisable since it implies that the intrinsic error due to the noise does not affect the bound besides the γ-discounted visitation. We show in Appendix that this quantity is always smaller than its counterpart involving the noisy policies.

Remark 2. *If we perform BC injecting a noise η_t of density function ℓ and satisfying Assumption 3, we have the following performance guarantee:*

$$J^{\pi_{E,\ell}} - J^{\pi_{I,\ell}} \leqslant \mathcal{O}(\varepsilon),$$

where ε is the MSE of the imitation policy as in Remark 1.

In comparison with Remark 1 for standard BC, we can appreciate that, here, the exponent α disappeared. Indeed, we have a performance bound that decreases linearly in the MSE. In many cases, when the environment is not intrinsically very smooth, or the expert policy is irregular, the α parameter can be very small, slowing down the convergence significantly. Instead, a liner decay is a relevant improvement of the v speed. Furthermore, as already noted, no assumption of regularity is required in Theorem 8, so that the last result has a much wider range of applications.

6 Practical Considerations

In the previous sections, we have seen that the use of noise injection allows having a much better performance guarantee than standard BC (see Remarks 1 and 2). Still, in practice, what matters is to have an imitator policy that is good itself rather than an imitator that is simply good in mimicking a given policy. Therefore, if with the noise injection we negatively affect the performance of the expert, i.e., if $J^{\pi_{E,\ell}} \ll J^{\pi_E}$, the results given about noise injection could become useless. On the contrary, we argue that adding noise to the expert's action to a certain extent, does not particularly affect performance. In Figure 3, we show the results of testing this statement on some of the most common continuous-actions environments of the `OpenAI gym` (Brockman et al. 2016) library. In this simulation,[5] we first train an expert policy with DDPG (Lillicrap et al. 2015), TD3 (Fujimoto, Hoof, and Meger 2018) and PPO (Schulman et al. 2017) in the following `OpenAI gym` environments:

- `Pendulum-v0`: this environment has a continuous action space $[-2, 2]$. The objective is to apply torque on a pendulum to swing it into an upright position. The whole system is very regular, as it is governed by simple differential equations, and is also deterministic, except for the initial position of the pendulum, which is random.
- `LunarLanderContinuous-v2`: this environment has a continuous action space $[-1, 1]^2$. Here, we have to make a rocket land safely in a landing pad. The dynamics is quite complex, and stochasticity is present to simulate the effect of the wind.
- `BipedalWalker-v3`: this environment has a continuous action space $[-1, 1]^4$. Here we have to make a bipedal robot walk. The dynamics is even more complex, but the whole system is deterministic.

Then, we evaluated the performance of these experts with noise injection with Gaussian noise with different standard deviations. As we can see in figure 3, even when the noise increases until it is close to the radius of the action space, at least in seven cases out of nine, the performance does not suffer significant drops. Intuitively, this can be explained by the fact that we applied an i.i.d. zero-mean noise sequence that is independent of the state and the action. Thus, its effect does not accumulate over the horizon.

7 Conclusions

In this paper, we have addressed BC for continuous-action environments from a theoretical perspective. We have shown that the existing theoretical guarantees on BC are not suitable when dealing with continuous actions. Thus, we have derived a first bound for the performance guarantees, under the assumption that the imitator value function is Lipschitz continuous. Since this latter assumption is demanding (i.e., it is not guaranteed even when the underlying MDP and policy are LC), we have relaxed it by studying the continuity properties of the value function. As a result of independent interest, we have proved that the value function is always Hölder continuous, under the milder assumption that the underlying MDP and policy are LC. Then, we have applied these findings to obtain a general bound for the performance gap of BC, which we have proved to be tight. Finally, we have formalized noise injection and we have shown the advantages of this practice when applied to BC.

[5] Details can be found in Appendix.

References

Arora, S.; and Doshi, P. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artif. Intell.*, 297: 103500.

Asadi, K.; Misra, D.; and Littman, M. L. 2018. Lipschitz Continuity in Model-based Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 264–273.

Asfour, T.; Azad, P.; Gyarfas, F.; and Dillmann, R. 2008. Imitation Learning of Dual-Arm Manipulation Tasks in Humanoid Robots. *Int. J. Humanoid Robotics*, 5(2): 183–202.

Bain, M.; and Sammut, C. 1995. A Framework for Behavioural Cloning. In *Machine Intelligence 15, Intelligent Agents*, 103–129.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. OpenAI Gym. gym:arXiv:1606.01540.

Fujimoto, S.; Hoof, H.; and Meger, D. 2018. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, 1587–1596. PMLR.

Geng, T.; Lee, M.; and Hülse, M. 2011. Transferring human grasping synergies to a robot. *Mechatronics*, 21(1): 272–284.

Ho, J.; and Ermon, S. 2016. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 4565–4573.

Jeon, H. J.; Milli, S.; and Dragan, A. D. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*.

Kakade, S. M.; and Langford, J. 2002. Approximately Optimal Approximate Reinforcement Learning. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML)*, 267–274.

Laskey, M.; Lee, J.; Fox, R.; Dragan, A. D.; and Goldberg, K. 2017a. DART: Noise Injection for Robust Imitation Learning. In *1st Annual Conference on Robot Learning (CoRL)*, 143–156.

Laskey, M.; Lee, J.; Fox, R.; Dragan, A. D.; and Goldberg, K. 2017b. DART: Noise Injection for Robust Imitation Learning. In *1st Annual Conference on Robot Learning (CoRL)*, 143–156.

Likmeta, A.; Metelli, A. M.; Ramponi, G.; Tirinzoni, A.; Giuliani, M.; and Restelli, M. 2021. Dealing with multiple experts and non-stationarity in inverse reinforcement learning: an application to real-life problems. *Mach. Learn.*, 110(9): 2541–2576.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Munos, R.; and Szepesvári, C. 2008. Finite-Time Bounds for Fitted Value Iteration. *Journal of Machine Learning Research*, 9(5).

Osa, T.; Pajarinen, J.; Neumann, G.; Bagnell, J. A.; Abbeel, P.; and Peters, J. 2018. An Algorithmic Perspective on Imitation Learning. *Found. Trends Robotics*, 7(1-2): 1–179.

Pirotta, M.; Restelli, M.; and Bascetta, L. 2015. Policy gradient in Lipschitz Markov Decision Processes. *Mach. Learn.*, 100(2-3): 255–283.

Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rachelson, E.; and Lagoudakis, M. G. 2010. On the locality of action domination in sequential decision making. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*.

Rényi, A.; et al. 1961. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1. Berkeley, California, USA.

Rozo, L. D.; Jiménez, P.; and Torras, C. 2013. A robot learning from demonstration framework to perform force-based manipulation tasks. *Intell. Serv. Robotics*, 6(1): 33–51.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Villani, C. 2009. *Optimal transport: old and new*, volume 338. Springer.

Xu, T.; Li, Z.; and Yu, Y. 2020. Error Bounds of Imitating Policies and Environments. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*.

Zanzotto, F. M. 2019. Viewpoint: Human-in-the-loop Artificial Intelligence. *J. Artif. Intell. Res.*, 64: 243–252.