On the Occupancy Measure of Non-Markovian Policies in Continuous MDPs

Romain Laroche* [1] **Rémi Tachet des Combes*** [1]

Abstract

The state-action occupancy measure of a policy is the expected (discounted or undiscounted) number of times a state-action couple is visited in a trajectory. For decades, RL books have been reporting the occupancy equivalence between Markovian and non-Markovian policies in countable state-action spaces under mild conditions. This equivalence states that the occupancy of any non-Markovian policy can be equivalently obtained by a Markovian policy, *i.e.* a memoryless probability distribution, conditioned only on its current state. While expected, for technical reasons, the translation of this result to continuous state space has resisted until now. Our main contribution is to fill this gap and to provide a general measure-theoretic treatment of the problem, permitting, in particular, its extension to continuous MDPs. Furthermore, we show that when the occupancy is infinite, we may encounter some non-trivial cases where the result does not hold anymore.

1. Introduction

Reinforcement learning (RL) is a popular and powerful theoretical framework for computational decision-making (Sutton & Barto, 2018), with many impressive accomplishments (Mnih et al., 2015; Silver et al., 2017). A central object of study in the field is the *Markovian policy*, in which an agent's actions are chosen from a *memoryless* probability distribution, *i.e.*, are conditioned only on its current state. The family of Markovian policies is broad enough to be interesting, yet simple enough to be amenable to analysis. For example, every MDP admits an optimal Markovian policy (Sutton & Barto, 2018), and it is possible to guarantee monotonic improvement when moving between Markovian policies (Kakade & Langford, 2002; Laroche & Tachet des Combes, 2021).

*Equal contribution [1]Unemployed, work done while at Microsoft Research Montréal, Canada. Correspondence to: Romain Laroche <romain.laroche@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

However, RL settings and algorithms often also involve *non-Markovian policies*, which may choose different probability distributions of actions in the same state depending on additional context. For example, non-Markovian policies are encountered in *Offline RL* (Levine et al., 2020), where the agent is not given the opportunity to interact with the environment at all, but instead must learn from a dataset of trajectories collected by an arbitrary set of policies. Other examples are algorithms that use replay buffers (Mnih et al., 2015), update online (Mnih et al., 2015), and/or allow sub-policy-switching, such as in the Semi MDP framework, options, or hierarchical policies (Barto & Mahadevan, 2003; Nachum et al., 2018; Stolle & Precup, 2002; Sutton et al., 1999). Formal analysis of these settings is possible, but somewhat involved. A typical approach is to prove a result under the restrictive assumption that trajectories are collected with a Markovian policy (Simao et al., 2020; Peng et al., 2019).

Nevertheless, a certain form of equivalence exists between Markovian policies and collections of non-Markovian policies. Indeed, the *occupancy measure*[1] (Szepesvári, 2022) of any non-Markovian policy whose total occupancy is finite can be equivalently obtained by a Markovian policy:

Theorem 1 (Restricted version of Theorem 4)**.** *Let m be an MDP with countable state and action spaces, a discount factor $\gamma < 1$, and let π be a policy. Then, there exists a Markovian policy $\tilde{\pi}$ that has the same occupancy measure in m as π.*

Variations of Theorem 1 are given in Borkar (1988); Feinberg & Shwartz (2002); Altman (1999); Ziebart (2010); Puterman (2014). Its most general version can be found in Altman (1999) where it is proved for a countable state space, a compact action space, and under the assumption that the policy is *absorbing* (*i.e.* that its total occupancy is finite), which subsumes the discounted case. Feinberg & Sonin (1996) is also of particular interest as it gives counter-examples showing that the absorbing condition is necessary for the equality to hold. When the condition is not verified but π is only *transient*, *i.e.* its occupancy measure on each state is finite[2], Altman (1999) also proves that the con-

[1]Also called state-action visits (Sutton & Barto, 2018), or distribution (Silver et al., 2014; Cheng et al., 2020).

[2]*Transience* is a necessary condition for the construction of $\tilde{\pi}$.

structed Markovian policy $\tilde{\pi}$ lower bounds π in terms of occupancy measures. While Theorem 1 has been featured in most theoretical RL books for several decades, its extension to continuous state-action spaces has resisted the test of time.

In this paper, we solve this decades long problem and provide a generalization of Theorem 1 to any state space (notably the continuous ones), including the standard RL environments not covered by the existing theory such as MuJoCo (Todorov et al., 2012) and DeepMind Control (Tunyasuvunakool et al., 2020). The extension leverages standard measure-theoretic concepts, necessary to handle arbitrary measurable states spaces. In particular, the *absorbing* and *transient* conditions on the policy from existing works naturally become finiteness and σ-finiteness of its occupancy measure.

The paper is organized as follows. First, we provide the background and notations used in the paper (Section 1.1). Next, we formally introduce the concept of occupancy measure (Section 1.2). Then, we give a minimal example illustrating the occupancy measure equivalence and the construction of the Markovian policy (Section 1.3). Section 2.1 details and discusses our various results. In Section 3, we broadly motivate the impact of the equivalence by enumerating fields of research where it could be potentially useful. Section 4 details the proofs of our main results. Finally, Section 5 concludes the paper.

1.1. Background and Notations

In this section, we introduce the definitions and notations required to state our result in its most general form. Capitalized letters denote random variables. Sets are denoted by calligraphic letters, and subsets by lower-case greek letters, except for μ, γ, and π, which are well-established notations for measures, discount factors, and policies respectively. If not mentioned otherwise, any set $\mathcal{X}$ is equipped with a σ-algebra $\Sigma_{\mathcal{X}}$ and a measure $\mu_{\mathcal{X}}$. We will write $\mathcal{P}(\Sigma_{\mathcal{X}})$ for the set of probability measures over $\Sigma_{\mathcal{X}}$.

Typically, for a countable set $\mathcal{X}$, $\Sigma_{\mathcal{X}}$ is the set of all subsets of $\mathcal{X}$, written $2^{\mathcal{X}}$ (also called its powerset), and $\mu_{\mathcal{X}}$ is defined as the counting measure, *i.e.*, $\mu_{\mathcal{X}}(\xi)$ is the number of elements in ξ for any $\xi \in \Sigma_{\mathcal{X}}$. Typically, for $\mathcal{X} \subset \mathbb{R}^n$, $\Sigma_{\mathcal{X}}$ is the Lebesgue σ-algebra and $\mu_{\mathcal{X}}(\xi)$ the Lebesgue measure.

A Markov Decision Process (MDP) is a tuple $m = \langle \mathcal{S}, \mathcal{A}, p_0, p, r, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $p_0(\cdot) \in \mathcal{P}(\Sigma_{\mathcal{S}} \times \{\emptyset, \{s_f\}\})$ denotes the initial state distribution, $p(\cdot|s,a) \in \mathcal{P}(\Sigma_{\mathcal{S}} \times \{\emptyset, \{s_f\}\})$ is the transition kernel, $s_f \notin \mathcal{S}$ denotes the final state where episode termination happens, $r(s,a) \in [-r_\top, r_\top]$ is the bounded stochastic reward function, and $\gamma \in [0,1]$ denotes the discount factor.

Definition 1 (Policy)**.** A policy π represents any function mapping its trajectory history $h_t = \langle s_0, a_0, r_0 \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t \rangle$ to a distribution over actions: $\pi(\cdot|h_t) \in \mathcal{P}(\Sigma_{\mathcal{A}})$. Let Π denote the space of policies.

This functional definition considers the policy as a black box: its inner workings do not matter as long as they do not manifest in the environment. It is a compact, necessary and sufficient, way of describing any policy[3]: either two policies differ for some (accessible) history and they are distinguishable, or they do not differ anywhere and they are indistinguishable. Thus, this is a fully general definition[4]: any behavior can be implemented by acting according to a member of Π.

Definition 2 (Markovian policy)**.** Policy π is said to be Markovian if its action probabilities only depend on the current state s_t: $\pi(\cdot|h_t) = \pi(\cdot|s_t) \in \mathcal{P}(\Sigma_{\mathcal{A}})$. Otherwise, policy π is non-Markovian. We let Π_{M} denote the space of Markovian policies. We let Π_{DM} denote the space of deterministic Markovian policies, *i.e.*, the set of Markovian policies such that $\pi(\cdot|s)$ is a Dirac distribution in some action a_s for any state $s \in \mathcal{S}$.

Because of the Markovian property of the MDP environment, Markovian policies are often a sufficiently broad set to solve the RL problem. In particular, there always exists an optimal policy that is deterministic Markovian, and the Markovian policy space happens to be convenient to navigate smoothly between deterministic Markovian policies.

Next, we shall also need various basic measure theory concepts that we recall here (Feller, 1968).

Definition 3 (σ-finiteness)**.** A measure μ on a measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$ is finite if $\mu(\mathcal{X}) < +\infty$. It is σ-finite if there exists a sequence $(\xi_n)_{n\in\mathbb{N}} \in \Sigma_{\mathcal{X}}^{\mathbb{N}}$ such that $\mathcal{X} = \cup_{n=0}^{\infty} \xi_n$ and $\mu(\xi_n) < +\infty$ for all $n \in \mathbb{N}$.

Intuitively, σ-finiteness states that the space can be decomposed into a countable union of finite measurable sets. For instance, the Lebesgue measure on $\mathbb{R}$ is σ-finite.

Definition 4 (Radon-Nikodym derivative)**.** Let μ and ν denote two σ-finite measures where ν is absolutely continuous with respect to μ (*i.e.*, $\mu(\xi) = 0 \implies \nu(\xi) = 0$). There exists a function $f : \mathcal{X} \to [0, +\infty]$ such that for all $\xi \in \Sigma_{\mathcal{X}}$,

$$\nu(\xi) = \int_{\xi} f(x)\mu(dx). \tag{1}$$

Any function f verifying 1 is called a Radon-Nikodym derivative and is denoted $\frac{d\nu}{d\mu}$. Two functions f_1, f_2

[3] Although, it is arguably an inefficient way of designing/implementing one.

[4] In order to have a well-defined occupancy measure, we must restrict ourselves to policies that reset their memory at the start of every trajectory. A thorough discussion is provided at the end of Section 2.1 regarding this.

that verify 1 are equal up to a μ-null set, *i.e.*, $\mu\left(\{x \text{ s.t. } f_1(x) \neq f_2(x)\}\right) = 0$.

1.2. Occupancy Measures

We now have the machinery required to introduce our main object of study: the occupancy measure.

Definition 5 (Occupancy)**.** Given measurable subsets of the state and action spaces, $\sigma \in \Sigma_{\mathcal{S}}, \alpha \in \Sigma_{\mathcal{A}}$, the occupancy $\mu^\pi_\gamma(\sigma, \alpha)$ of a policy $\pi \in \Pi$ in an MDP $m = \langle \mathcal{S}, \mathcal{A}, p_0, p, r, \gamma \rangle$ is the expected discounted number of visits of a state-action pair $(s, a) \in \sigma \times \alpha$ occurring during a trajectory:

$$\mu^\pi_\gamma(\sigma, \alpha) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}\left(S_t \in \sigma\right) \times \mathbb{1}\left(A_t \in \alpha\right) \,\middle|\, \begin{array}{l} S_0 \sim p_0(\cdot), A_t \sim \pi(\cdot|H_t), \\ S_{t+1} \sim p(\cdot|S_t, A_t) \end{array}\right]. \tag{2}$$

We will use the conventions that $\mu^\pi_\gamma(\sigma) \doteq \mu^\pi_\gamma(\sigma, \mathcal{A})$, and with finite state-action sets $\mu^\pi_\gamma(s, a) \doteq \mu^\pi_\gamma(\{s\}, \{a\})$. Note that Definition 5 holds both for discrete and continuous state and action spaces.

We start by establishing that the occupancy as defined in Eq. (2) is well-defined and is a measure for any policy π and any MDP m, this will allow us to leverage standard results from measure theory.

Theorem 2 (**Occupancy is a measure**)**.** *Let $\pi \in \Pi$ be any policy as defined in 1, then, μ^π_γ is well-defined on $\mathbb{R}^+ \cup \{+\infty\}$ and is a measure.*

We defer the proof of this result to Appendix B. The second interesting property of μ^π_γ concerns the discounted return ρ^π_γ of π:

$$\rho^\pi_\gamma := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \,\middle|\, \begin{array}{l} S_0 \sim p_0(\cdot), A_t \sim \pi(\cdot|H_t), \\ R_t \sim r(S_t, A_t), S_{t+1} \sim p(\cdot|S_t, A_t) \end{array}\right]. \tag{3}$$

Lemma 3. *If ρ^π_γ exists, then it is uniquely characterized by μ^π_γ: $\rho^\pi_\gamma = \int_{\mathcal{S}} \int_{\mathcal{A}} \mathbb{E}\left[r(s, a)\right] \mu^\pi_\gamma(ds, da)$.*

We defer the proof of this result to Appendix C. A general equivalence in terms of value is harder to make as there is no clear definition of a state marginalized value function for non Markovian policies.

1.3. Illustrative Example

We consider the MDP m represented in Figure 1, with a single state $\mathcal{S} = \{s\}$ two actions $\mathcal{A} = \{a_1, a_2\}$, and such that $p(s|s, a_1) = 1$ and a_2 terminates the episode. For a fixed $n \geq 1$, we consider the following deterministic policy: $\pi(a_1|s, t < n) = 1$ and $\pi(a_2|s, t = n) = 1$. All its trajectories are therefore the same: it performs a_1 n times, then a_2 which terminates the trajectory. Fundamentally, π is non-Markovian since its actions depends on the timestep t. It is direct to observe that the occupancy measure, corresponding to the expected discounted state and state-action visits (denoted $\mu^\pi_\gamma(s)$ and $\mu^\pi_\gamma(s, a)$ respectively), verifies:

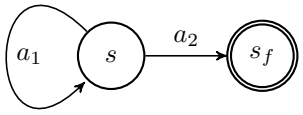

Figure 1: Minimal MDP such that a_1 loops and a_2 is terminal.

$$\mu^\pi_\gamma(s) = \sum_{t=0}^{n} \gamma^t \quad \mu^\pi_\gamma(s, a_1) = \sum_{t=0}^{n-1} \gamma^t \quad \mu^\pi_\gamma(s, a_2) = \gamma^n.$$

We define a Markovian policy $\tilde{\pi}$, equal to the expected behavior of π in s over all possible trajectories:

$$\tilde{\pi}(a_1|s) \doteq \frac{\mu^\pi_\gamma(s, a_1)}{\mu^\pi_\gamma(s)} = \frac{\sum_{t=0}^{n-1} \gamma^t}{\sum_{t=0}^{n} \gamma^t} \tag{4}$$

$$\tilde{\pi}(a_2|s) \doteq \frac{\mu^\pi_\gamma(s, a_2)}{\mu^\pi_\gamma(s)} = \frac{\gamma^n}{\sum_{t=0}^{n} \gamma^t}. \tag{5}$$

Given the MDP structure, we have:

$$\mu^{\tilde{\pi}}_\gamma(s, a_1) = \tilde{\pi}(a_1|s) + \tilde{\pi}(a_1|s)\, \gamma \mu^{\tilde{\pi}}_\gamma(s, a_1). \tag{6}$$

Therefore:

$$\begin{aligned} \mu^{\tilde{\pi}}_\gamma(s, a_1) &= \frac{\tilde{\pi}(a_1|s)}{1 - \tilde{\pi}(a_1|s)\gamma} = \frac{\sum_{t=0}^{n-1} \gamma^t}{\sum_{t=0}^{n} \gamma^t - \gamma \sum_{t=0}^{n-1} \gamma^t} \\ &= \sum_{t=0}^{n-1} \gamma^t = \mu^\pi_\gamma(s, a_1). \end{aligned} \tag{7}$$

Similarly, $\mu^{\tilde{\pi}}_\gamma(s, a_2) = \tilde{\pi}(a_2|s) + \tilde{\pi}(a_1|s)\ \gamma \mu^{\tilde{\pi}}_\gamma(s, a_2)$, which gives:

$$\begin{aligned} \mu^{\tilde{\pi}}_\gamma(s, a_2) &= \frac{\tilde{\pi}(a_2|s)}{1 - \tilde{\pi}(a_1|s)\gamma} = \frac{\gamma^n}{\sum_{t=0}^{n} \gamma^t - \gamma \sum_{t=0}^{n-1} \gamma^t} \\ &= \gamma^n = \mu^\pi_\gamma(s, a_2). \end{aligned} \tag{8}$$

We see that $\tilde{\pi}$ has exactly the same state-action visits as π. Our main theorem states that, in any MDP and under

mild assumptions, such a policy always exists. We wish to emphasize, however, that their trajectories distributions differ: all trajectories generated with π have the same length, $n+1$, while the length of trajectories generated with $\tilde{\pi}$ follows a geometric law. It is also worth noticing that $\tilde{\pi}$ depends on the choice of the discount factor, but that, for any discount factor $\gamma < 1$, there will exist a Markovian policy equivalent to π in terms of state-action occupancy.

2. Theory

We start by enunciating the main theorem, derive its corollaries, and explicate the occupancy equivalence relation in Subsection 2.1. Then, we discuss the necessity of the σ-finiteness in Subsection 2.2. Finally, we conclude the theory with a series of additional remarks in Subsection 2.3.

2.1. Main Theorem and Occupancy Equivalence

Let us now state our main theorem and discuss its implications. Its proof can be found in Section 4.

Theorem 4 (State-action occupancy equivalence). *Let π be a policy with σ-finite occupancy measure μ^π_γ. For any measurable $\alpha \subseteq \mathcal{A}$, we define $\tilde{\pi}$ as the Radon-Nikodym derivative:*

$$\tilde{\pi}(\alpha|s) := \frac{d\mu^\pi_\gamma(\cdot,\alpha)}{d\mu^\pi_\gamma(\cdot)}(s), \tag{9}$$

where $\mu^\pi_\gamma(\cdot,\alpha)$ and $\mu^\pi_\gamma(\cdot)$ are seen as measures on $\mathcal{S}$. The following statements hold.

- *$\tilde{\pi}(\alpha|s)$ exists and is a probability measure on $\mathcal{A}$ for any $s \in \mathcal{S}$, i.e., a Markovian policy.*
- *$\tilde{\pi}$ admits a σ-finite occupancy measure and $\mu^{\tilde{\pi}}_\gamma \leq \mu^\pi_\gamma$.*
- *Moreover, if μ^π_γ is finite (i.e. $\mu^\pi_\gamma(\mathcal{S}) < \infty$), then $\mu^{\tilde{\pi}}_\gamma = \mu^\pi_\gamma$.*

Remark: Note that $\tilde{\pi}$ is uniquely defined up to a $\mu^\pi_\gamma(\cdot)$-null set. Further characterizing it is unimportant as a set of states with null occupancy measure will almost surely never be visited.

In two very generic settings, the Radon-Nikodym derivative $\tilde{\pi}$ can be explicitly characterized.

Corollary 5. *When $\mathcal{S}$ and $\mathcal{A}$ are finite, we let $\mu^\pi_\gamma(s,a)$ denote the σ-finite state-action occupancy measure of $(s,a) \in \mathcal{S}\times\mathcal{A}$ under π. Then, $\tilde{\pi}(a|s) := \frac{\mu^\pi_\gamma(s,a)}{\mu^\pi_\gamma(s)}$, and $\mu^{\tilde{\pi}}_\gamma(s,a) = \mu^\pi_\gamma(s,a)$.*

Corollary 5 covers for instance the illustration provided in Section 1.3. Importantly, the finiteness of $\mathcal{S}$ and the σ-finiteness of μ^π_γ imply $\mu^\pi_\gamma(\mathcal{S}) < \infty$, and thus the equality of occupancy measures.

Corollary 6. *When $\mathcal{S}$ and/or $\mathcal{A}$ are continuous, let $d^\pi_\gamma(s,a)$ denote the state-action occupancy density of the pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ and assume its existence. Then, $\tilde{\pi}(\alpha|s) := \int_\alpha \frac{d^\pi_\gamma(s,a)}{d^\pi_\gamma(s)}da$, and by abuse of notation $\tilde{\pi}(a|s) := \frac{d^\pi_\gamma(s,a)}{d^\pi_\gamma(s)}$. Furthermore, $d^{\tilde{\pi}}_\gamma(s,a) \leq d^\pi_\gamma(s,a)$ with equality if $\int_\mathcal{S} d^\pi_\gamma(s)ds < \infty$.*

Corollary 6 covers exclusively-continuous state and action spaces such as certain robotic manipulation tasks and some MuJoCo environments (Todorov et al., 2012). We note that Theorem 4 is more general than Corollaries 5 and 6 combined as the state-action visitation density in infinite state-action space is not always defined (there may be Dirac points). We now discuss the theorem in details.

Policy performance: The first implication of Theorem 4, combined with Lemma 3, is that for any policy π with finite occupancy measure, there exists a Markovian policy with the same performance (proof in Appendix C).

Corollary 7. *Under suitable existence assumptions, $\rho^\pi_\gamma = \rho^{\tilde{\pi}}_\gamma$.*

Idempotence: Equation (9) may be interpreted as an operator over policies: $\tilde{\pi} = \mathcal{R}\pi$ up to a $\mu^\pi_\gamma(\cdot)$-null set. Proposition 8 proves that this operator is idempotent: $\mathcal{R}\mathcal{R}\pi = \mathcal{R}\pi$ (still up to a $\mu^\pi_\gamma(\cdot)$-null set). In other words, $\mathcal{R}$ is a projection from the policy space Π to the Markovian policy space Π_M. This can also be seen through the lens of the following pseudo-metric on Π: $d(\pi_1,\pi_2) = TV(\mu^{\pi_1}_\gamma, \mu^{\pi_2}_\gamma)$ (where TV denotes the total variation). Furthermore, one can define the occupancy equivalence relation: $\pi_1 \overset{\text{occ}}{\sim} \pi_2$ if policies π_1 and π_2 are occupancy equivalent, meaning that $\mu^{\pi_1}_\gamma = \mu^{\pi_2}_\gamma$ and implying that $\mathcal{R}\pi_1 = \mathcal{R}\pi_2$ up to a $\big(\mu^{\pi_1}_\gamma(\cdot) + \mu^{\pi_2}_\gamma(\cdot)\big)$-null set (proof in Appendix D).

Proposition 8. *If π is Markovian with a σ-finite occupancy measure, then $\tilde{\pi} = \pi$, where equality is up to a $\mu^\pi_\gamma(\cdot)$-null set.*

Universality: Our result is universal: it applies to any MDP m, any policy π, and any discount factor γ, as long as the occupancy measure of π is σ-finite. The σ-finiteness condition is not an artifact of the proof technique, as illustrated in the following subsection.

2.2. Necessity of the σ-finiteness of μ^π_γ

Proposition 9. *If π has a σ-infinite occupancy measure, $\tilde{\pi}$ may be undetermined and there may not be any Markovian policy with the same occupancy measure as π.*

We detail next, two counter-examples proving the Proposition 9.

Example 1 (Example where $\tilde{\pi}$ is undetermined)**.** We use the minimal undiscounted ($\gamma = 1$) MDP in Figure 2 with

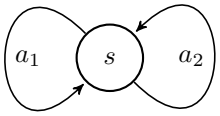

Figure 2: Minimal MDP such that $p(s|s, a_1) = 1$ and $p(s|s, a_2) = 1$.

a single state $\mathcal{S} = \{s\}$ and two actions $\mathcal{A} = \{a_1, a_2\}$ that loop: $p(s|s, a_1) = 1$ and $p(s|s, a_2) = 1$. We consider the non-Markovian policy π that chooses its action as follows: $\forall i \in \mathbb{N}$, if $t \in [2^i, 2^{i+1})$, then $\pi(a_1) = \frac{(-1)^i+1}{2}$, *i.e.* during each epoch i, the policy plays deterministically a_1 if i is even otherwise a_2 if i is odd.

Then, the ratio $\tilde{\pi} \doteq \frac{\mu_1^\pi(s,a_1)}{\mu_1^\pi(s)}$ is undetermined as the limit of the action selection ratio does not converge as t goes to infinity. We could argue that any non-determinisitic Markovian policy would admit the same occupancy measure as π: $\mu_1^\pi(s, a_1) = \mu_1^\pi(s, a_2) = \infty$, but the counter-example in the next subsection shows that this is not always possible.

Example 2 (Example where no Markovian policy reproduces the occupancy)**.** We use the minimal undiscounted ($\gamma = 1$) MDP in Figure 1 with a single state $\mathcal{S} = \{s\}$ and two actions $\mathcal{A} = \{a_1, a_2\}$, such that $p(s|s, a_1) = 1$ and a_2 is terminal. We consider the non-Markovian policy π that chooses its action uniformly at timestep $t = 0$ and deterministically a_1 for $t \geq 1$. Its occupancy measure is therefore:

$$\mu_1^\pi(s, a_1) = \infty \qquad \mu_1^\pi(s, a_2) = \frac{1}{2}. \tag{10}$$

The set of Markovian policies $\pi_\theta \in \Pi_{\text{M}}$ may be parametrized with a single parameter $\pi_\theta(a_1|s) \doteq \theta \in [0, 1]$ and $\pi_\theta(a_2|s) \doteq 1 - \theta \in [0, 1]$. A Markovian policy π_θ admits the following occupancy measure:

$$\text{if } \theta < 1, \qquad \mu_1^{\pi_\theta}(s, a_1) = \frac{1}{1-\theta} \qquad \mu_1^{\pi_\theta}(s, a_2) = 1, \tag{11}$$

$$\text{if } \theta = 1, \qquad \mu_1^{\pi_\theta}(s, a_1) = \infty \qquad \mu_1^{\pi_\theta}(s, a_2) = 0, \tag{12}$$

none of which match the occupancy of π.

We additionally refer to Feinberg & Sonin (1996) for an example where the inequality between measures is strict if $\mu_\gamma^\pi(\mathcal{S}) = +\infty$ even when $\mu_\gamma^\pi(\{s\}) < +\infty$ for all $s \in \mathcal{S}$. We also note that in Altman (1999); Feinberg & Sonin (1996), the notion of *transience* is introduced, and means that the occupancy measure of any state is finite. This is equivalent to the occupancy measure being sigma-finite. Indeed, sigma-finiteness means that the state space is the countable union of a set of finite-measured measurable sets. In the discrete case, it is implied by the finiteness of the measure of each singleton (aka transience), as the state space is the countable union of the singletons it contains.

It is remarkable that in discrete undiscounted MDPs, transience can be characterized as states being visited a finite number of times. In other words, for each state, there is a timestep after which it is never visited again in the trajectory. However, Example 3 hereafter shows that the same characterization of transience does not imply sigma-finiteness in continuous MDPs: in a given trajectory, a state is not visited more than once, however, the constructed policy's occupancy measure is not sigma-finite.

Finiteness of the occupancy measure: Equality in Theorem 4 applies only if the policy's occupancy measure is finite. It is trivially verified if $\gamma < 1$. We give now a tighter sufficient condition.

Proposition 10. *If all deterministic Markovian policies have finite occupancy measures, any policy π admits a finite occupancy measure and Theorem 4 applies.*

Proof of this proposition can be found in Appendix E. When the state space is countable (as assumed in previous works studying these types of policy equivalences), finiteness can be relaxed into σ-finiteness in the above proposition. However, as soon as the state space is continuous, this relaxation fails. In other words, with continuous MDPs, even if all deterministic Markovian policies are σ-finite, there may exist non-Markovian policies for which constructing an equivalent Markovian policy using Equation (9) cannot be done. We give below such a counter-example.

Example 3 (All deterministic Markovian policies have σ-finite occupancy but there exists a σ-infinite occupancy policy)**.** We consider the deterministic continuous MDP m where $\mathcal{S} = [0, 2]$, $s_0 = 0$, $\mathcal{A} = (0, 1]$, $p(s + a|s, a) = 1$, $\gamma = 1$, and the trajectory terminates when $s + a > 2$.

We start by establishing that any deterministic policy π_d (Markovian or non-Markovian) has a σ-finite occupancy measure: since the environment and the policy are deterministic, every trajectory is the same. Since $\mathcal{A} = (0, 1]$ and $p(s + a|s, a) = 1$, the state s_t is strictly increasing with t. This implies that either:

1. the trajectory terminates and the occupancy measure is finite (and therefore σ-finite),

2. or the trajectory is upper bounded by 2 and by the monotone convergence theorem, it must converge to

some state s_∞ without ever reaching it (if $s_t = s_\infty$ for some t, then $s_{t+1} > s_\infty$, which is a contradiction).

Since case 1. proves our point, we focus on case 2. from now on. Still from the strictly increasing property, we infer that the occupancy measure of π_d is 1 for the states s_t on the deterministic trajectory and 0 everywhere else. We consider the following partition of $\mathcal{S}$:

$$\sigma_0 \doteq [s_\infty, 2] \qquad \forall i > 0, \quad \sigma_i \doteq [s_{i-1}, s_i). \tag{13}$$

By construction, $\mu_\gamma^{\pi_d}(\sigma_0) = 0, \forall i > 0, \mu_\gamma^{\pi_d}(\sigma_i) = 1$, and $\mathcal{S} = \bigcup_{i\in\mathbb{N}} \sigma_i$, which proves the σ-finiteness of $\mu_\gamma^{\pi_d}$.

Now, we construct a policy π that is σ-infinite.

$$\pi(\cdot|t=0) = \mathcal{U}([0,1]) \qquad \pi(\cdot|t>0) = \frac{1}{t} - \frac{1}{t+1} \tag{14}$$

Let A_0 denote the first action, which is the only stochastic one, then the state reached at time t is:

$$\begin{aligned} S_{t+1} &= S_t + \frac{1}{t} - \frac{1}{t+1} = A_0 + \sum_{t'=1}^{t} \frac{1}{t'} - \frac{1}{t'+1} \\ &= A_0 + 1 - \frac{1}{t+1}, \end{aligned} \tag{15}$$

which converges to $A_0 + 1$ as t tends to infinity. For any segment $[b, c] \subset [0, 1]$ with $b < c$, $\mathbb{P}(A_0 \in (b, c]) = c - b > 0$. Then, we look at the measure of $[b + 1, c + 1]$:

$$\begin{aligned} &\mu_\gamma^\pi([b+1, c+1]) \geq \\ &\mathbb{P}\left(A_0 \in (b,c]\right) \mathbb{E}\left[\sum_{t=0}^{\infty} \mathbb{1}(s_t \in [b+1, c+1]) \;\middle|\; A_0 \in (b,c]\right] \\ &= \mathbb{P}\left(A_0 \in (b,c]\right) \mathbb{E}\left[\sum_{t=\lceil \frac{1}{A_0 - b} \rceil}^{\infty} 1 \;\middle|\; A_0 \in (b,c]\right] = \infty, \end{aligned}$$

which concludes the proof that μ_γ^π is σ-infinite.

2.3. Additional Remarks

Trajectory distribution: As already noted in Section 1.3, the same occupancy measure does not imply the same trajectory distribution, which may have a specific role in non-bootstrapping algorithms. For instance, Decision Transformers (Chen et al., 2021; Furuta et al., 2022; Emmons et al., 2022) learn the return-conditional distribution of actions in each state, and then define a policy by sampling from the distribution of actions that receive high return in each state. It is therefore a trajectory-based algorithm as opposed to more classical RL bootstrapping approaches that are sample-based and can fully take advantage of Theorem 4. Nevertheless, it is possible to prove the following result, stating that even though trajectory distributions are not equal, any trajectory generated by π has a non-zero probability under $\bar{\pi}$ (proof in Appendix F).

Proposition 11. *For any $t \geq 0$, we let τ_t^π denote the trajectory distribution on $(\mathcal{S} \times \mathcal{A})^t$ induced by executing π t times in the environment, starting from p_0. Then, τ_t^π is absolutely continuous with respect to $\tau_t^{\bar{\pi}}$.*

Policies with inter-episode non-stationarity: The non-stationarity within a trajectory is already handled by the non-Markovian-ness as the trajectory history includes the timestamp information. While useful in practice, for instance to deal with learning algorithms, the non-stationarity across trajectories is trickier to handle because they have a trajectory dependent occupancy measure, and even worse, they may not have any "average occupancy measure". For instance, for a stateless MDP with 2 terminating actions, the non-stationary policy $\pi_t(a_1) = 1 - \pi_t(a_2) = 1$ when $t \in [2^{2k}, 2^{2k+1} - 1]$ and $\pi_t(a_1) = 1 - \pi_t(a_2) = 0$ when $t \in [2^{2k+1}, 2^{2k+2} - 1]$ for all integers k, will not admit any occupancy measure (nor an equivalent Markovian policy).

Thus, our occupancy measure theory does not concern policies that carry memory from one episode to another. Still, oftentimes datasets and replay buffers are collected with a learning algorithm, *i.e.*, a policy π that is updated across time. Let us consider the recording of the N policies $\{\pi_i\}_{i\in[N]} \in \Pi^N$ that were used in each individual episode:

$$\pi_i(\cdot \mid h) \doteq \pi\left(\cdot \mid h \cup h_{\tau_{i-1}} \cup \cdots \cup h_{\tau_1}\right), \tag{16}$$

where h is the current trajectory history and h_{τ_i} denotes the recorded history of the i^{th} trajectory. Then, Theorem 4 may be applied on the inter-episode-policy $\bar{\pi} \doteq \mathcal{U}(\{\pi_i\}_{i\in[N]})$ that uniformly samples at the start of each trajectory a policy among the N policies, allowing us to conclude that the occupancy in the dataset may still be reproduced by a single Markovian policy. Nevertheless, we stress again the fact that this occupancy is not connected to that of π, since the latter is undetermined.

Non-Markovian policies usefulness: The occupancy equivalence does not deny the algorithmic interest of non-Markovian policies, as their existence may be entailed by the problem setting (Offline RL or replay buffers), and as they may prove their usefulness by allowing to inject inductive bias, such as options, or by generating diverse behaviors with an ensemble of agents. On the contrary, we view Theorem 4 as a tool allowing to carry out theoretical grounding usually restricted to Markovian policies to non-Markovian policies. In the next section, we showcase various such applications.

3. Applications

Many RL domains or algorithms may benefit from Theorem 4 as they rely on the use of non-Markovian policies (generally a collection of Markovian policies). In most of these domains, the non-Markovian property of the policies

is a feature, not an issue: it allows one to break down and better compound some conflicting objectives, to induce diversity, and/or to design new policies from elementary ones. Theorem 4 may be a powerful tool for their respective convergence guarantees by proving that their non-Markovian policy admits a well-studied (i.e., Markovian) policy emulating its occupancy measure.

Non-Markovian policy induced by the problem setting: Behavioral Cloning (Urbancic, 1994; Torabi et al., 2018) and Imitation Learning (Ross et al., 2011; Ho & Ermon, 2016; Hussein et al., 2017) consist in training an agent to reproduce an expert behavior from demonstrations. Sometimes, the expert behavior collection is generated from several near-optimal policies. Moreover, some approaches involve interactive data collection processes in order to make sure that the agent can recover from its own errors. In both cases, the collected data does not come from a single Markovian policy and Theorem 4 may prove to be useful. In Imitation Learning, the goal is to "reproduce" a behavior policy. However, it is generally modeled as a Markovian policy. Our result proves that their approach is sound under the mild assumption that $\gamma < 1$.

Offline RL consists in training a policy on a fixed set of observations without access to the true environment (Lange et al., 2012; Levine et al., 2020). Similarly to Imitation Learning, most algorithms and analyses either implicitly or explicitly make the assumption that the behavior policy β that was used to collect data is unique and Markovian (Laroche et al., 2019; Fujimoto et al., 2019b; Buckman et al., 2020; Kumar et al., 2020; Thomas et al., 2015; Yu et al., 2021; Yin et al., 2021; Shi et al., 2022). However, this is generally not true: in healthcare for instance, patients are often followed by different doctors/health centers with different policies. Furthermore, typical benchmarks for offline reinforcement learning are constructed from the experience replay of DQN runs (Fujimoto et al., 2019a; Agarwal et al., 2020b), or via an amalgamation of expert policies (Fu et al., 2020).

We will further illustrate the impact of our result by looking at a particular algorithm family called SPIBB for Safe Policy Improvement with Baseline Bootstrapping (Laroche et al., 2019; Nadjahi et al., 2019). It consists in allowing policy change only when the change is sufficiently supported by the dataset. These algorithms rely on two components: a state-action uncertainty and an estimate of the behavior policy (Simao et al., 2020) Estimating a non-Markovian policy faces the curse of dimensionality, thus such policies are often not treated. Our theorem proves that the main theoretical results in Simao et al. (2020) actually carry over to non-Markovian behaviour policies. Indeed, since there exists a Markovian policy that yields the same expected performance and occupancy measure, one may simply consider that the data was received from the induced Markovian policy and obtain the same performance properties.

Non-Markovian policy induced by algorithmic family: Multi-objective algorithms often combine several policies to generate a behavior that matches the new objective trade-offs (Shelton, 2001; Vamplew et al., 2009). Ensemble RL (Wiering & Van Hasselt, 2008), algorithm selection for RL (Laroche & Féraud, 2018), diversity-induced exploration (Eysenbach et al., 2019), sets of policies based on Generalized Policy Improvement (Barreto et al., 2017; Alegre et al., 2022), and curriculum for RL (Czarnecki et al., 2018) all rely on training a family of RL agents. Some more theoretical papers focus on non-Markovian policies (Wu et al., 2004; Scherrer & Lesner, 2012). More and more, policy gradient algorithms utilize and maintain several policies. The PC-PG algorithm (Agarwal et al., 2020a) consists in improving the global convergence guarantees of policy gradient methods by implementing an initial state distribution that covers the whole state space. To do so, they learn a policy cover that is made of multiple Markovian policies. Jekyll & Hyde (Laroche & Tachet des Combes, 2021) is another actor-critic algorithm improving convergence guarantees by maintaining two Markovian policies: one dedicated to pure exploration and the other to pure exploitation. Finally, it is worth mentioning distributed agents which perform training updates over several behavioral policies (Mnih et al., 2016; Horgan et al., 2018; Schmitt et al., 2020).

4. Proof of Theorem 4

Proof of Theorem 4. Let us start with the first bullet point, concerning the existence of $\tilde{\pi}$ and the fact that it is indeed a Markovian policy.

Letting $\alpha \subseteq \mathcal{A}$ be a measurable set in $\mathcal{A}$. We see that for any $\sigma \subseteq \mathcal{S}$ measurable, we have $\mu^{\pi}_{\gamma}(\sigma, \alpha) \leq \mu^{\pi}_{\gamma}(\sigma)$. This directly implies that $\mu^{\pi}_{\gamma}(\cdot, \alpha)$ is absolutely continuous with respect to $\mu^{\pi}_{\gamma}(\cdot)$ (both seen as measures on $\mathcal{S}$). Since $\mu^{\pi}_{\gamma}(\cdot)$ is σ-finite, the Radon-Nikodym theorem states that $\mu^{\pi}_{\gamma}(\cdot, \alpha)$ admits a density with respect to $\mu^{\pi}_{\gamma}(\cdot)$, we let $\tilde{\pi}$ denote that Radon-Nikodym derivative.

We now prove that $\tilde{\pi}$ is a probability measure. The non-negativity is directly inherited from that of measure μ^{π}_{γ}. The null-empty set comes from the fact that $\mu^{\pi}_{\gamma}(\sigma, \emptyset) = 0$ for all measurable sets σ. The countable additivity is a consequence of the countable additivity of the Radon-Nykodym derivative and of the measure μ^{π}_{γ}. And finally, it is clear from its definition that $\tilde{\pi}(\mathcal{A}|s) = 1$ and that $\tilde{\pi}$ only depends on the current state, which makes it a Markovian policy.

Let us now move the second bullet point, which is our core result. Recalling that $\mu^{\pi}_{\gamma}(ds) = \int_{\mathcal{A}} \mu^{\pi}_{\gamma}(ds, da)$, and $\mu^{\tilde{\pi}}_{\gamma}(ds) = \int_{\mathcal{A}} \mu^{\tilde{\pi}}_{\gamma}(ds, da)$, we wish to prove that for all $\sigma \in$

$\Sigma_{\mathcal{S}}$: $\mu_\gamma^{\tilde{\pi}}(\sigma) \leq \mu_\gamma^{\pi}(\sigma)$, with equality if $\mu_\gamma^{\pi}(\mathcal{S}) < \infty$. Letting $\sigma \in \Sigma_{\mathcal{S}}$ denote a measurable set such that $\mu_\gamma^{\pi}(\sigma) < \infty$, we know from the conservation of mass of σ-finite occupancy measures (Proposition 12 in Appendix A) that:

$$\mu_\gamma^{\pi}(\sigma) = p_0(\sigma) + \gamma \int_{\mathcal{S}} \int_{\mathcal{A}} \mu_\gamma^{\pi}(ds_{-1}, da)\; p(\sigma | s_{-1}, a), \tag{17}$$

where $p(\sigma|s_{-1}, a)$ denotes the probability of transitioning to σ when taking action a in state s_{-1}. Now, by definition of the Radon-Nikodym derivative, we have $\mu_\gamma^{\pi}(ds, da) = \mu_\gamma^{\pi}(ds) \frac{d\mu_\gamma^{\pi}(\cdot \times da)}{d\mu_\gamma^{\pi}(\cdot \times \mathcal{A})}(s) = \mu_\gamma^{\pi}(ds)\tilde{\pi}(da|s)$. Using that property, we see that:

$$\mu_\gamma^{\pi}(\sigma) = p_0(\sigma) + \gamma \int_{\mathcal{S}} \int_{\mathcal{A}} \mu_\gamma^{\pi}(ds_{-1})\tilde{\pi}(da|s_{-1})\; p(\sigma|s_{-1}, a) \tag{18}$$

$$= p_0(\sigma) + \gamma \int_{\mathcal{S}} \mu_\gamma^{\pi}(ds_{-1})\; p^{\tilde{\pi}}(s_{-1}, \sigma), \tag{19}$$

where $p^{\tilde{\pi}}(s_{-1}, ds)$ is the Markov kernel on $\mathcal{S} \times \mathcal{S}$ obtained by composition of p and $\tilde{\pi}$, and $p^{\tilde{\pi}}(s_{-1}, \sigma)$ is the probability of transitioning to σ when acting according to $\tilde{\pi}$ in state s_{-1}. Applying the above conservation equality recursively t times gives:

$$\begin{aligned}\mu_\gamma^{\pi}(\sigma) = p_0(\sigma) &+ \gamma \int_{\mathcal{S}} p_0(ds_{-1}) p^{\tilde{\pi}}(s_{-1}, \sigma) + \cdots \\ &+ \gamma^t \int_{\mathcal{S}} p_0(ds_{-t})\; p_t^{\tilde{\pi}}(s_{-t}, \sigma) \\ &+ \gamma^{t+1} \int_{\mathcal{S}} \mu_\gamma^{\pi}(ds_{-t-1})\; p_{t+1}^{\tilde{\pi}}(s_{-t-1}, \sigma),\end{aligned} \tag{20}$$

where $p_t^{\tilde{\pi}}$ denotes the composition of $p^{\tilde{\pi}}$ with itself t times. The equality can easily be shown by induction and Fubini's theorem. Given the finiteness of $\mu_\gamma^{\pi}(\sigma)$ and the positivity of all the terms involved, there exists $l \geq 0$ such that $\gamma^{t+1} \int_{\mathcal{S}} \mu_\gamma^{\pi}(ds_{-t-1})\; p_{t+1}^{\tilde{\pi}}(s_{-t-1}, \sigma) \to_{t\to\infty} l$. We obtain:

$$\mu_\gamma^{\pi}(\sigma) = p_0(\sigma) + \sum_{t=1}^{\infty} \gamma^t \int_{\mathcal{S}} p_0(ds_{-t})\; p_t^{\tilde{\pi}}(s_{-t}, \sigma) \;+\; l. \tag{21}$$

Now, by the very definition of occupancy measures and Markov policies, we see that $p_0(\sigma) + \sum_{t=1}^{T} \gamma^t \int_{\mathcal{S}} p_0(ds_{-t})\; p_t^{\tilde{\pi}}(s_{-t}, \sigma)$ is the partial sum of $\mu_\gamma^{\tilde{\pi}}(\sigma)$ in Equation 2:[5]

$$\mu_\gamma^{\tilde{\pi}}(\sigma, \mathcal{A}) = \lim_{T\to+\infty} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \mathbb{1}\left(S_t \in \sigma\right) \times \mathbb{1}\left(A_t \in \mathcal{A}\right) \middle| \begin{array}{l} S_0 \sim p_0(\cdot), A_t \sim \tilde{\pi}(\cdot|S_t), \\ S_{t+1} \sim p(\cdot|S_t, A_t) \end{array}\right]. \tag{22}$$

[5]Note that in the general case, it is not the partial sum of $\mu_\gamma^{\pi}(\sigma)$ due to π's non-Markovian character.

The convergence of this partial sum is guaranteed by the finiteness of $\mu_\gamma^{\pi}(\sigma)$. In other words, we get:

$$\mu_\gamma^{\tilde{\pi}}(\sigma) = p_0(\sigma) + \sum_{t=1}^{\infty} \gamma^t \int_{\mathcal{S}} p_0(ds_{-t})\; p_t^{\tilde{\pi}}(s_{-t}, \sigma) \tag{23}$$

$$= \mu_\gamma^{\pi}(\sigma) \;-\; l \leq \mu_\gamma^{\pi}(\sigma) < +\infty. \tag{24}$$

Now, for $\sigma \in \Sigma_{\mathcal{S}}$ (possibly of infinite μ_γ^{π}-measure), the σ-finiteness of μ_γ^{π} implies there exists a sequence $(\sigma_n)_{n\in\mathbb{N}}$ of disjoint sets such that $\forall n \in \mathbb{N}, \mu_\gamma^{\pi}(\sigma_n) < +\infty$ and $\sigma = \cup_{n=0}^{\infty} \sigma_n$. We compute:

$$\mu_\gamma^{\tilde{\pi}}(\sigma) = \sum_{n=0}^{\infty} \mu_\gamma^{\tilde{\pi}}(\sigma_n) \leq \sum_{n=0}^{\infty} \mu_\gamma^{\pi}(\sigma_n) = \mu_\gamma^{\pi}(\sigma). \tag{25}$$

Combining this equality with the policy definition (9) gives the final result: $\mu_\gamma^{\tilde{\pi}}(ds, da) \leq \mu_\gamma^{\pi}(ds, da)$.

We are left with proving that $\lim_{t\to\infty} \gamma^{t+1} \int_{\mathcal{S}} \mu_\gamma^{\pi}(ds_{-t-1})\; p_{t+1}^{\tilde{\pi}}(s_{-t-1}, \sigma) = 0$ when $\mu_\gamma^{\pi}(\mathcal{S}) < +\infty$. It is obvious when $\gamma < 1$, since the integral term is bounded. The case $\gamma = 1$ is somewhat more involved (as is customary with undiscounted MDPs).

We start by noticing that $\int_{\mathcal{S}} \mu_\gamma^{\tilde{\pi}}(ds_{-t-1})\; p_{t+1}^{\tilde{\pi}}(s_{-t-1}, \sigma) \to_{t\to\infty} 0$. This stems directly from applying Eq. (21) to $\tilde{\pi}$ and then leveraging the first equality in (24) . Now, we let $L = \{s \in \mathcal{S} \mid p_t^{\tilde{\pi}}(s, \sigma) \to_{t\to\infty} 0\}$. From $\int_{L^C} \mu_\gamma^{\tilde{\pi}}(ds_{-t-1})\; p_{t+1}^{\tilde{\pi}}(s_{-t-1}, \sigma) \to_{t\to\infty} 0$ and by the definition of L, we infer that $\mu_\gamma^{\tilde{\pi}}(L^C) = 0$. It also stems directly from Proposition 11 that μ_γ^{π} is absolutely continuous with respect to $\mu_\gamma^{\tilde{\pi}}$, which implies that $\mu_\gamma^{\pi}(L^C) = 0$. Finally, by the dominated convergence theorem, applicable since $\mu_\gamma^{\pi}(\mathcal{S}) < +\infty$ and $p_{t+1}^{\tilde{\pi}} \leq 1$, we get:

$$\begin{aligned}&\int_{\mathcal{S}} \mu_\gamma^{\pi}(ds_{-t-1})\; p_{t+1}^{\tilde{\pi}}(s_{-t-1}, \sigma) \\ &\qquad = \int_{L} \mu_\gamma^{\pi}(ds_{-t-1})\; p_{t+1}^{\tilde{\pi}}(s_{-t-1}, \sigma) \to_{t\to\infty} 0,\end{aligned} \tag{26}$$

which concludes the proof. □

5. Conclusion

In this paper, we developed a general theory of the occupancy measure in MDPs, and extended to continuous state spaces the result stating that, for any non-Markovian policy admitting a finite occupancy, there exists a Markovian policy with the same occupancy. We also provided a variety of auxiliary results analyzing the equivalence and the conditions under which it is feasible.

References

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in Neural Information Processing Systems*, 33:13399–13412, 2020a.

Agarwal, R., Schuurmans, D., and Norouzi, M. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020b.

Alegre, L. N., Bazzan, A., and Silva, B. C. D. Optimistic linear support and successor features as a basis for optimal policy transfer. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 394–413. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/alegre22a.html.

Altman, E. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. Successor features for transfer in reinforcement learning. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Barto, A. G. and Mahadevan, S. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(1):41–77, 2003.

Borkar, V. A convex analytic approach to markov decision processes. *Probability Theory and Related Fields*, 1988. URL https://doi.org/10.1007/BF00353877.

Buckman, J., Gelada, C., and Bellemare, M. G. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=a7APmM4B9d.

Cheng, C.-A., Tachet des Combes, R., Boots, B., and Gordon, G. A reduction from reinforcement learning to no-regret online learning. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3514–3524. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/cheng20b.html.

Czarnecki, W., Jayakumar, S., Jaderberg, M., Hasenclever, L., Teh, Y. W., Heess, N., Osindero, S., and Pascanu, R. Mix & match agent curricula for reinforcement learning. In *International Conference on Machine Learning*, pp. 1087–1095. PMLR, 2018.

Emmons, S., Eysenbach, B., Kostrikov, I., and Levine, S. Rvs: What is essential for offline RL via supervised learning? In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=S874XAIpkR-.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. Diversity is all you need: Learning skills without a reward function. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

Feinberg, E. and Shwartz, A. (eds.). *Handbook of Markov Decision Processes - Methods and Applications*. Kluwer International Series, 2002.

Feinberg, E. A. and Sonin, I. Notes on equivalent stationary policies in markov decision processes with total rewards. *Math. Methods Oper. Res.*, 44(2):205–221, 1996. URL http://dblp.uni-trier.de/db/journals/mmor/mmor44.html#FeinbergS96.

Feller, W. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968. ISBN 0471257087. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20{&}path=ASIN/0471257087.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Fujimoto, S., Conti, E., Ghavamzadeh, M., and Pineau, J. Benchmarking batch deep reinforcement learning algorithms. *arXiv preprint arXiv:1910.01708*, 2019a.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019b.

Furuta, H., Matsuo, Y., and Gu, S. S. Generalized decision transformer for offline hindsight information matching. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=CAjxVodl_v.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Horgan, D., Quan, J., Budden, D., Barth-Maron, G., Hessel, M., van Hasselt, H., and Silver, D. Distributed prioritized experience replay. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1179–1191, 2020.

Lange, S., Gabel, T., and Riedmiller, M. *Batch Reinforcement Learning*, pp. 45–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_2.

Laroche, R. and Féraud, R. Reinforcement learning algorithm selection. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

Laroche, R. and Tachet des Combes, R. Dr Jekyll and Mr Hyde: The strange case of off-policy policy updates. In *Proceedings of the 34th Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2109.14727v1.

Laroche, R., Trichelair, P., and Tachet des Combes, R. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Nachum, O., Gu, S. S., Lee, H., and Levine, S. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Nadjahi, K., Laroche, R., and Tachet des Combes, R. Safe policy improvement with soft baseline bootstrapping. In *Proceedings of the 17th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2019.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

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

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, pp. 627–635, 2011.

Scherrer, B. and Lesner, B. On the use of non-stationary policies for stationary infinite-horizon markov decision processes. *Advances in Neural Information Processing Systems*, 25, 2012.

Schmitt, S., Hessel, M., and Simonyan, K. Off-policy actor-critic with shared experience replay. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8545–8554. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/schmitt20a.html.

Shelton, C. R. Importance sampling for reinforcement learning with multiple objectives. *MIT AI Technical Reports*, 2001.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. *arXiv preprint arXiv:2202.13890*, 2022.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. A. Deterministic policy gradient algorithms. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 387–395. JMLR.org, 2014. URL http://dblp.uni-trier.de/db/conf/icml/icml2014.html#SilverLHDWR14.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Simao, T. D., Laroche, R., and Tachet des Combes, R. Safe policy improvement with an estimated baseline policy. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, May 2020.

Stolle, M. and Precup, D. Learning options in reinforcement learning. In *International Symposium on abstraction, reformulation, and approximation*, pp. 212–223. Springer, 2002.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Sutton, R. S., Precup, D., and Singh, S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211, 1999.

Szepesvári, C. Theoretical foundations of reinforcement learning course, 2022. URL https://rltheory.github.io/lecture-notes/planning-in-mdps/lec2/.

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Torabi, F., Warnell, G., and Stone, P. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.

Tunyasuvunakool, S., Muldal, A., Doron, Y., Liu, S., Bohez, S., Merel, J., Erez, T., Lillicrap, T., Heess, N., and Tassa, Y. Dm control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2020.100022. URL https://www.sciencedirect.com/science/article/pii/S2665963820300099.

Urbancic, T. Reconstructing human skill with machine learning. In *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI)*, pp. 498–502. John Wiley, 1994.

Vamplew, P., Dazeley, R., Barker, E., and Kelarev, A. Constructing stochastic mixture policies for episodic multiobjective reinforcement learning tasks. In *Australasian joint conference on artificial intelligence*, pp. 340–349. Springer, 2009.

Wiering, M. A. and Van Hasselt, H. Ensemble algorithms in reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(4):930–936, 2008.

Wu, W., Arapostathis, A., and Kumar, R. On non-stationary policies and maximal invariant safe sets of controlled markov chains. In *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, volume 4, pp. 3696–3701. IEEE, 2004.

Yin, M., Bai, Y., and Wang, Y.-X. Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34, 2021.

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. COMBO: Conservative offline model-based policy optimization. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Ziebart, B. D. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, USA, 2010. AAI3438449.

A. Auxiliary Result

For the sake of completeness, we include the following very standard result.

Proposition 12 (Conservation of mass)**.** *For any policy π whose occupancy measure is σ-finite, we have:*

$$\mu_\gamma^\pi(ds) = p_0(ds) + \gamma \int_{\mathcal{S}} \int_{\mathcal{A}} \mu_\gamma^\pi(ds_{-1}, da)\; p(ds|s_{-1}, a). \tag{27}$$

Proof. Note that the equality is meant in the sense of measures, that is, for any $\sigma \in \Sigma_{\mathcal{S}}$:

$$\mu_\gamma^\pi(\sigma) = p_0(\sigma) + \gamma \int_{\mathcal{S}} \int_{\mathcal{A}} \mu_\gamma^\pi(ds_{-1}, da)\; p(\sigma|s_{-1}, a). \tag{28}$$

Let us start by proving that this equality holds for any σ such that $\mu_\gamma^\pi(\sigma) < +\infty$:

$$\mu_\gamma^\pi(\sigma, \alpha) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}\left(S_t \in \sigma\right) \times \mathbb{1}\left(A_t \in \alpha\right) \,\middle|\, \begin{array}{l} S_0 \sim p_0(\cdot), A_t \sim \pi(\cdot|H_t), \\ S_{t+1} \sim p(\cdot|S_t, A_t) \end{array}\right] \tag{29}$$

$$= \sum_{t=0}^{\infty} \gamma^t \int_\sigma \int_\alpha p_t(ds, da), \tag{30}$$

where the second line is a direct application of Fubini's theorem (valid since $\mu_\gamma^\pi(\sigma) < +\infty$), and $p_t(ds, da)$ denotes the measure on $\mathcal{S} \times \mathcal{A}$ induced by (S_t, A_t) when both evolve according to p_0, π and p. Note in particular that $p_0(ds, da) = p_0(ds)\pi(da|s)$ where $p_0(ds)$ is the initial state distribution in the MDP. Naturally, $p_t(ds)$ denotes the marginal of $p_t(ds, da)$ on $\mathcal{S}$. Focusing on $\alpha = \mathcal{A}$, we see that:

$$\mu_\gamma^\pi(\sigma) = \sum_{t=0}^{\infty} \gamma^t \int_\sigma p_t(ds) = p_0(\sigma) + \sum_{t=1}^{\infty} \gamma^t \int_\sigma p_t(ds) \tag{31}$$

$$= p_0(\sigma) + \sum_{t=1}^{\infty} \gamma^t \int_\sigma \int_{\mathcal{S}} \int_{\mathcal{A}} p_{t-1}(ds_{-1}, da)\; p(ds|s_{-1}, a) \tag{32}$$

$$= p_0(\sigma) + \gamma \int_\sigma \left[\sum_{t=1}^{\infty} \gamma^{t-1} \int_{\mathcal{S}} \int_{\mathcal{A}} p_{t-1}(ds_{-1}, da)\right]\; p(ds|s_{-1}, a) \tag{33}$$

$$= p_0(\sigma) + \gamma \int_\sigma \int_{\mathcal{S}} \int_{\mathcal{A}} \mu_\gamma^\pi(ds_{-1}, da)\; p(ds|s_{-1}, a), \tag{34}$$

where all integrals on σ are with respect to s, Fubini was applied again, and we used the equality $p_t(ds) = \int_{\mathcal{S}} \int_{\mathcal{A}} p_{t-1}(ds_{-1}, da)\; p(ds|s_{-1}, a)$ that stems directly from the definition of a Markov Decision Process.

Finally, for any $\sigma \in \Sigma_{\mathcal{S}}$, we know from the σ-finiteness of μ_γ^π that there exists a sequence $(\sigma_n)_{n\in\mathbb{N}}$ of disjoint measurable sets such that $\forall n \in \mathbb{N}, \mu_\gamma^\pi(\sigma_n) < +\infty$ and $\sigma = \cup_{n=0}^{\infty} \sigma_n$. Applying Eq. 28 to σ_n and summing over $n \in \mathbb{N}$ concludes the proof. □

B. Proof of Theorem 2

Theorem 2 (**Occupancy is a measure**)**.** *Let $\pi \in \Pi$ be any policy as defined in 1, then, μ_γ^π is well-defined on $\mathbb{R}^+ \cup \{+\infty\}$ and is a measure.*

Proof of Theorem 2. Fixing $\sigma \in \Sigma_{\mathcal{S}}$ and $\alpha \in \Sigma_{\mathcal{A}}$, we notice that the following sequence is increasing with T:

$$U_T \coloneqq \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \mathbb{1}\left(S_t \in \sigma\right) \times \mathbb{1}\left(A_t \in \alpha\right) \,\middle|\, \begin{array}{l} S_0 \sim p_0(\cdot), A_t \sim \pi(\cdot|H_t), \\ S_{t+1} \sim p(\cdot|S_t, A_t) \end{array}\right].$$

Therefore, the monotone convergence theorem guarantees that U_T converges on $\mathbb{R}^+ \cup \{+\infty\}$ when T tends to infinity and its limit is by construction the occupancy measure of π, which is therefore defined for every state set σ and action set α.

In order to establish that μ_γ^π is a measure over the algebra product $\Sigma_\mathcal{S} \times \Sigma_\mathcal{A}$, we need to check (i) its positivity, (ii) that $\mu_\gamma^\pi(\emptyset) = 0$, and (iii) its countable additivity with respect to disjoint sets. (i) has been established right before, (ii) is a direct consequence that $\mathbb{1}(S_t \in \emptyset) \times \mathbb{1}(A_t \in \emptyset) = 0$, and (iii) is a simple summation order change, justified by the positivity of all quantities involved. □

C. Characterization of Performance

Lemma 3. *If ρ_γ^π exists, then it is uniquely characterized by μ_γ^π: $\rho_\gamma^\pi = \int_\mathcal{S} \int_\mathcal{A} \mathbb{E}[r(s,a)]\, \mu_\gamma^\pi(ds, da)$.*

Proof. We recall the definition of the performance ρ_γ^π:

$$\rho_\gamma^\pi := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \,\middle|\, \begin{array}{l} S_0 \sim p_0(\cdot), A_t \sim \pi(\cdot|H_t), \\ R_t \sim r(S_t, A_t), S_{t+1} \sim p(\cdot|S_t, A_t) \end{array} \right]. \tag{35}$$

We start by noting that the existence of ρ_γ^π is not guaranteed in the case of $\gamma = 1$ and $\mu_\gamma^\pi(\mathcal{S}) = +\infty$ (see Example 4).

Assuming it does exist (or that any of those two conditions is not verified), the law of total expectation gives:

$$\rho_\gamma^\pi := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{E}[r(S_t, A_t)] \,\middle|\, \begin{array}{l} S_0 \sim p_0(\cdot), A_t \sim \pi(\cdot|H_t), \\ S_{t+1} \sim p(\cdot|S_t, A_t) \end{array} \right]. \tag{36}$$

Applying the law of total expectation a second time, and Fubini-Tonelli's theorem allows to conclude. □

Example 4 (Indeterminate performance)**.** We consider the MDP depicted in Figure 2 with $r(s, a_1) = 1$, $r(s, a_2) = -1$, $\pi(a_1|t = 0) = 1$ and $\pi(\cdot|t > 0) = 1 - \pi(\cdot|t - 1)$: the trajectory is deterministically looping over state s, performing alternatively a_1 and a_2. If $\gamma = 1$, $\rho_1^\pi = \lim_{T\to\infty} \sum_{t=0}^{T} R_t$, but $\sum_{t=0}^{T} R_t$ does not have any limit as T tends to infinity, since it equals 1 if T is even, and 0 otherwise.

Corollary 7. *Under suitable existence assumptions, $\rho_\gamma^\pi = \rho_\gamma^{\tilde{\pi}}$.*

Proof. Since $\mu_\gamma^\pi = \mu_\gamma^{\tilde{\pi}}$, we get from Lemma 3:

$$\rho_\gamma^\pi = \int_\mathcal{S} \int_\mathcal{A} \mathbb{E}[r(s,a)]\, \mu_\gamma^\pi(ds, da) = \int_\mathcal{S} \int_\mathcal{A} \mathbb{E}[r(s,a)]\, \mu_\gamma^{\tilde{\pi}}(ds, da) = \rho_\gamma^{\tilde{\pi}},$$

which concludes the proof. □

D. Idempotence and Projection

Proposition 8. *If π is Markovian with a σ-finite occupancy measure, then $\tilde{\pi} = \pi$, where equality is up to a $\mu_\gamma^\pi(\cdot)$-null set.*

Proof. Given the definition of $\tilde{\pi}(\alpha|s) = \frac{d\mu_\gamma^\pi(\cdot,\alpha)}{d\mu_\gamma^\pi(\cdot,\mathcal{A})}(s)$, to show that $\tilde{\pi} = \pi$, we simply need to prove that for any $\sigma \in \Sigma_\mathcal{S}$ and $\alpha \in \Sigma_\mathcal{A}$:

$$\mu_\gamma^\pi(\sigma, \alpha) = \int_\sigma \mu_\gamma^\pi(ds, \mathcal{A}) \pi(\alpha|s). \tag{37}$$

From the definition 2 of $\mu_\gamma^\pi(\sigma, \alpha)$, we have:

$$\mu_\gamma^\pi(\sigma, \alpha) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}(S_t \in \sigma) \times \mathbb{1}(A_t \in \alpha) \,\middle|\, \begin{array}{l} S_0 \sim p_0(\cdot), A_t \sim \pi(\cdot|H_t), \\ S_{t+1} \sim p(\cdot|S_t, A_t) \end{array} \right] \tag{38}$$

$$= \sum_{t=0}^{\infty} \gamma^t \int_\sigma \int_\alpha p_t(ds, da) \tag{39}$$

where we reused the notations from the proof of Proposition 12. Given the Markovian nature of π, we see that: $p_t(ds, da) = p_t(ds)\pi(da|s)$. Reintroducing this into 39, we get:

$$\mu_\gamma^\pi(\sigma, \alpha) = \sum_{t=0}^{\infty} \gamma^t \int_\sigma \int_\alpha p_t(ds)\pi(da|s) = \int_\sigma \sum_{t=0}^{\infty} \gamma^t p_t(ds)\pi(\alpha|s) = \int_\sigma \mu_\gamma^\pi(ds)\pi(\alpha|s), \tag{40}$$

which proves that π is indeed the Radon-Nikodym derivative $\frac{d\mu_\gamma^\pi(\cdot,\alpha)}{d\mu_\gamma^\pi(\cdot,\mathcal{A})}(s)$ and shows the equality to $\tilde{\pi}$ up to a $\mu_\gamma^\pi(\cdot, \mathcal{A})$-null set. □

E. Finiteness and σ-finiteness

The (σ-)finiteness of the occupancy measure is an interesting property as it will be required to avoid encountering indeterminate formulas. We note that $\gamma < 1$ suffices to guarantee that all Markovian policies have a finite occupancy measure for any policy. Below, we derive a more general characterisation of MDPs that admit only policies with finite occupancy measures.

Proposition 10. *If all deterministic Markovian policies have finite occupancy measures, any policy π admits a finite occupancy measure and Theorem 4 applies.*

Proof. We notice that the expected performance under reward $r(s, a) = 1\ \forall s, a$ is equal to the occupancy measure $\mu_\gamma^\pi(\mathcal{S})$ over the full state-action pair set. We use the well known theoretical result (Sutton & Barto, 1998) that there exists a deterministic Markovian policy that optimises any MDP. We let π^* denote one such optimal policy. If the occupancy measure is finite for any deterministic Markovian policies then it is for π^*, and we have for all π:

$$\mu_\gamma^\pi(\mathcal{S}) \leq \mu_\gamma^{\pi^*}(\mathcal{S}) < \infty, \tag{41}$$

which establishes the finiteness of the occupancy measure of π.

If we additionally consider a countable state space, the above reasoning can be extended to the σ-finite case. Let us assume that all deterministic Markovian policies are σ-finite, and consider a fixed $s \in \mathcal{S}$. We define the reward function $r(s, a) = 1\ \forall a$ and $r(s', a) = 0\ \forall s' \neq s, a$. There exists a deterministic Markovian policy π^* maximizing that reward function. Since π^* has a σ-finite occupancy measure, its performance is finite. We obtain for any policy π:

$$\mu_\gamma^\pi(\{s\}) \leq \mu_\gamma^{\pi^*}(\{s\}) < \infty. \tag{42}$$

Since this holds for any state s, and since the state space is countable, we conclude that μ_γ^π is σ-finite. □

F. Absolute Continuity of Finite Trajectory Distribution

Proposition 11. *For any $t \geq 0$, we let τ_t^π denote the trajectory distribution on $(\mathcal{S} \times \mathcal{A})^t$ induced by executing π t times in the environment, starting from p_0. Then, τ_t^π is absolutely continuous with respect to $\tau_t^{\tilde{\pi}}$.*

Proof. We prove this result by induction on $t \geq 0$. For $t = 0$, let us consider $(\sigma, \alpha) \in \Sigma_\mathcal{S} \times \Sigma_\mathcal{A}$ such that $\tau_0^{\tilde{\pi}}(\sigma, \alpha) = 0$. This implies that:

$$\int_\sigma \int_\alpha p_0(ds)\tilde{\pi}(da|s) = \int_\sigma p_0(ds) \frac{d\mu_\gamma^\pi(\cdot, \alpha)}{d\mu_\gamma^\pi(\cdot)}(s) = 0. \tag{43}$$

Letting $N_\alpha = \{s \in \mathcal{S} \mid \frac{d\mu_\gamma^\pi(\cdot,\alpha)}{d\mu_\gamma^\pi(\cdot)}(s) = 0\}$, we see that necessarily $p_0(N_\alpha \cap \sigma) = p_0(\sigma)$ (otherwise the above integrals would not be 0). Now, by definition of the Radon-Nikodym derivative, we have:

$$\mu_\gamma^\pi(N_\alpha, \alpha) = \int_{N_\alpha} \mu_\gamma^\pi(ds) \frac{d\mu_\gamma^\pi(\cdot, \alpha)}{d\mu_\gamma^\pi(\cdot)}(s) = 0. \tag{44}$$

Since $p_0(N_\alpha \cap \sigma) = p_0(\sigma)$, we know that: $\tau_0^\pi(\sigma, \alpha) = \tau_0^\pi(N_\alpha \cap \sigma, \alpha) \leq \mu_\gamma^\pi(N_\alpha, \alpha) = 0$, which concludes the base case.

Let us proceed to the induction step. We consider $(\sigma, \alpha) \in (\Sigma_{\mathcal{S}} \times \Sigma_{\mathcal{A}})^{t+1}$, with $\tau^{\tilde{\pi}}_{t+1}(\sigma, \alpha) = 0$, and aim to prove that $\tau^{\pi}_{t+1}(\sigma, \alpha) = 0$. We let $\sigma_{|t}$ and $\alpha_{|t}$ denote the first t components of σ and α, and σ_{t+1} and α_{t+1} their $t+1$-th. From the Markov property of the various objects involved, we have:

$$\tau^{\tilde{\pi}}_{t+1}(\sigma, \alpha) = \int_{\alpha_{t+1}} \int_{\sigma_{t+1}} \int_{\sigma_{|t}, \alpha_{|t}} \tilde{\pi}(da_{t+1} \mid s_{t+1}) p(ds_{t+1} \mid s_t, a_t) \tau^{\tilde{\pi}}_t(ds_{|t}, da_{|t}). \tag{45}$$

As far as π is concerned, we have:

$$\tau^{\pi}_{t+1}(\sigma, \alpha) = \int_{\alpha_{t+1}} \int_{\sigma_{t+1}} \int_{\sigma_{|t}, \alpha_{|t}} \pi(da_{t+1} \mid s_{|t+1}, a_{|t}) p(ds_{t+1} \mid s_t, a_t) \tau^{\pi}_t(ds_{|t}, da_{|t}). \tag{46}$$

Since $\tau^{\tilde{\pi}}_{t+1}(\sigma, \alpha) = 0$, three cases are possible (corresponding to the measure of the three integrals from right to left being null):

(i). $\tau^{\tilde{\pi}}_t(\sigma_{|t}, \alpha_{|t}) = 0$. In this case, the induction hypothesis implies $\tau^{\pi}_t(\sigma_{|t}, \alpha_{|t}) = 0$, and thus $\tau^{\pi}_{t+1}(\sigma, \alpha) = 0$.

(ii). $\int_{\sigma_{|t}, \alpha_{|t}} p(\sigma_{t+1} \mid s_t, a_t) \tau^{\tilde{\pi}}_t(ds_{|t}, da_{|t}) = 0$. We define $N_1 = \{(s, a) \mid p(\sigma_{t+1} \mid s, a) \neq 0\}$, and see that necessarily $\tau^{\tilde{\pi}}_t(N_1) = 0$, implying that $\tau^{\pi}_t(N_1) = 0$ by the induction hypothesis, and thus that $\tau^{\pi}_{t+1}(\sigma, \alpha) = 0$.

(iii). $\int_{\sigma_{t+1}} \tau^{\tilde{\pi}}_{t+1}(ds_{t+1}) \tilde{\pi}(\alpha_{t+1} | s_{t+1}) = 0$, where we overloaded the notations by letting $\tau^{\tilde{\pi}}_{t+1}(ds_{t+1})$ be the distribution of the $t+1$-th state in the trajectory when following $\tilde{\pi}$. Similarly as above, this implies that: $\tau^{\tilde{\pi}}_{t+1}(N_\alpha \cap \sigma_{t+1}) = \tau^{\tilde{\pi}}_{t+1}(\sigma_{t+1})$. In addition, using the same argument as in (ii), the induction hypothesis can be applied to get $\tau^{\pi}_{t+1}(\sigma_{|t} \times (N_\alpha \cap \sigma_{t+1}), \alpha) = \tau^{\pi}_{t+1}(\sigma, \alpha)$. Finally, $\tau^{\pi}_{t+1}(\sigma_{|t} \times (N_\alpha \cap \sigma_{t+1}), \alpha) \leq \mu^{\pi}_{\gamma}(N_\alpha, \alpha) = 0$.

This concludes the induction step, and with it the proof of the proposition. □