Putting the Spotlight on the Initial State Distribution

Aditya Makkar, Aarshvi Gajjar, Eugene Vinitsky
{adityamakkar,aarshvi,eugenevinitsky}@nyu.edu

New York University

Abstract

The initial state distribution in reinforcement learning (RL) is often treated as a technical detail, overshadowed by the focus on policy optimization and value function approximation. This paper challenges this perspective by providing a rigorous analysis and intuition of how the initial state distribution affects the objective function of RL. We derive performance difference lemmas that quantify how changes in the initial distribution propagate through the learning objective, revealing bounds that scale with $\frac{1}{1-\gamma}$ in the infinite horizon setting and $(T+1)$ in the finite horizon case, where γ is the discount factor and T is the horizon length. These lemmas and an illustrative example, demonstrate that seemingly minor changes in where an agent begins can lead to dramatically different outcomes—even when following the same policy. These results have immediate implications for practical RL deployments where the training and testing distributions often differ, and provide an alternate theoretical perspective for recent advances in reverse curriculum learning and local planning algorithms.

1 Introduction

Reinforcement learning (RL) provides a powerful framework for an agent learning through trial and error in sequential interactions with an unknown environment to maximize rewards (Foster & Rakhlin, 2023; Szepesvári, 2025). This interaction is formalized using the language of Markov decision processes (MDPs) which, among other things, specifies an initial state distribution (Bertsekas & Shreve, 1978; Puterman, 1994; Hernandez-Lerma & Lasserre, 2012). In theory, the standard optimality criterion in RL seeks policies that are optimal for *all* possible initial state distributions simultaneously. In practice, however, this stringent criterion is rarely adopted; policies are usually optimized for a fixed, predetermined initial state distribution.

When training RL agents, the assumption of a fixed initial state distribution is often too limiting or simply undesirable (Florensa et al., 2017; Tavakoli et al., 2019; Ecoffet et al., 2021; Yin et al., 2023; Mhammedi et al., 2024; Krishnamurthy et al., 2025). Consider training a robot to navigate through a building. Standard reinforcement learning formulations assume the robot will always start from the same distribution of locations—perhaps always at the main entrance. But what happens when the robot is deployed and starts from the storage room, or the third floor, or anywhere else? This seemingly straightforward scenario reveals a fundamental brittleness in many reinforcement learning formulations, highlighting their vulnerability when faced with situations that differ from the original training assumptions.

1.1 The Importance of Initial State Distribution

To understand why the initial state distribution matters, consider the following example:

Example 1.1 (The Bottleneck MDP)**.** Consider an MDP whose state space has two regions connected by a narrow bottleneck (see Figure 1). Region A contains small positive rewards, while

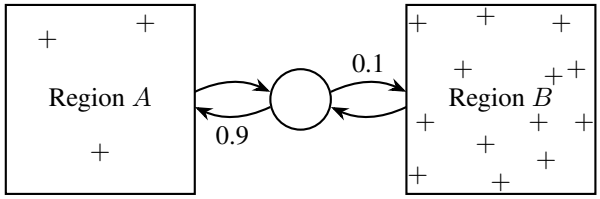

Figure 1: Bottleneck MDP. The state space is partitioned into three parts: two regions A and B, and a single bottleneck state connecting them. All the actions at the bottleneck state take the agent to a state in A with probability 0.9.

Region B contains large positive rewards. The bottleneck is a single state that connects them, and there's a high probability (0.9) of "slipping back" when trying to cross from A to B. There are no rewards in trying to cross the bottleneck.

Suppose there are two initial state distributions: ρ_A, supported on a subset of region A, and ρ_B, supported on a subset of region B. In the infinite horizon case, for a large enough discount factor, the optimal policy for states in region A will attempt to cross over to region B. In the finite horizon case, if the horizon is long enough, the policy will match the infinite horizon case. However, if the horizon is short, the optimal policy under ρ_A will avoid crossing over to region B, because there are no rewards in trying to cross the bottleneck. This example foreshadows our theoretical results: using $J(\rho, \pi^*)$ to denote the objective function for initial state distribution ρ and optimal policy π^*, the performance difference $J(\rho_A, \pi^*) - J(\rho_B, \pi^*)$ can be very large; these quantities will be defined formally in Section 2.

This example is not a pathological edge case—it reflects common structures in real-world problems: navigation tasks have dangerous areas to avoid, and manipulation tasks have irreversible actions. In each case, where you start can fundamentally change the problem.

1.2 Theoretical Gaps and Practical Implications

The seminal work of Kakade & Langford (2002) touched upon the importance of the initial state distribution, showing how policies could be improved while providing performance guarantees. Their performance difference lemma, which we extend in this work, quantifies how changing the policy affects performance. However, the dual question—how changing the initial distribution affects performance—remained relatively unexplored. The importance of adapting initial state distributions shows up in many sub-areas of RL.

Local Planning Environment Access Model. In the local planning environment access model, the agent can start an episode at any state it has visited previously. This allows the agent to explore faster (Ecoffet et al., 2021; Tavakoli et al., 2019; Yin et al., 2023). This framework and its variations provably give better statistical guarantees (Mhammedi et al., 2024; Krishnamurthy et al., 2025).

Curriculum Learning and Automatic Reset Distributions. The curriculum learning literature (Bengio et al., 2009; Florensa et al., 2017) implicitly recognizes that the choice of initial states affects learning efficiency. (Florensa et al., 2017) focus on goal oriented tasks, or tasks with sparse rewards by training a robot to start from increasingly distant start states. They study the problem of finding the optimal start state distribution to maximize learning speed.

Offline RL. Offline-RL guarantees such as those in (Jiang & Xie, 2025) are stated using concentrability coefficients like $C^\pi_\rho = \sup_{x,a} \frac{\psi^\pi_\rho(x,a)}{\psi^{\pi_d}_{\rho_d}(x,a)}$, where ψ^π_ρ is the discounted state-action occupancy measure. Therefore, any change in initial state distributions shifts these coefficients and directly affects the offline-RL performance bounds.

2 Preliminaries and Notation

All spaces are assumed to be endowed with a σ-algebra such that they are standard Borel, i.e., there exists a metric on the space which makes it a complete separable metric space in such a way that the endowed σ-algebra is then the Borel σ-algebra with respect to the topology induced by the metric (Preston, 2008). In particular, any countable space must be equipped with the discrete σ-algebra for it to be standard Borel. For a space X, its Borel σ-algebra is denoted as $\mathcal{B}(\mathsf{X})$. Given two measurable spaces $(\mathsf{X}, \mathscr{X})$ and $(\mathsf{Y}, \mathscr{Y})$, a map $K\colon \mathsf{X} \times \mathscr{Y} \to [0,1]$ is called a stochastic kernel from $(\mathsf{X}, \mathscr{X})$ to $(\mathsf{Y}, \mathscr{Y})$—or, briefly, from X to Y—if the map $x \mapsto K(x,B)$ is $\mathscr{X}$-measurable for each $B \in \mathscr{Y}$, and the map $B \mapsto K(x,B)$ is a probability measure on $(\mathsf{Y}, \mathscr{Y})$ for each $x \in \mathsf{X}$. We use $\mathcal{M}_1(\mathsf{X})$ to denote the set of probability measures on the space X. We will often use the (functional-analytic) notation $\mathbb{P}f$ instead of $\int f \,\mathrm{d}\mathbb{P}$ to denote the integral of f with respect to the measure $\mathbb{P}$. We refer the reader to the book by Kallenberg (2021) for more details on probability theory.

A **Markov decision process** (Bertsekas & Shreve, 1978; Puterman, 1994; Hernandez-Lerma & Lasserre, 2012) consists of the following: (i) a state space X and a finite action space A, (ii) a stochastic kernel $K\colon (\mathsf{X} \times \mathsf{A}) \times \mathcal{B}(\mathsf{X}) \to [0,1]$, which specifies that with probability $K((x,a),B)$ the state transitions to some state in $B \in \mathcal{B}(\mathsf{X})$ on taking action $a \in \mathsf{A}$ in state $x \in \mathsf{X}$, (iii) a measurable reward function $r\colon \mathsf{X} \times \mathsf{A} \to [-r_{\max}, r_{\max}]$, for $r_{\max} > 0$, and (iv) an initial state distribution $\rho \in \mathcal{M}_1(\mathsf{X})$.

The agent specifies a policy $\pi = (\pi_t)_{t \geq 0}$, where $\pi_t\colon ((\mathsf{X} \times \mathsf{A})^t \times \mathsf{X}) \times \mathcal{B}(\mathsf{A}) \to [0,1]$ is a stochastic kernel, which specifies that at time step $t \in \{0,1,\ldots\}$, having observed the history $h_t = (x_0, a_0, x_1, a_1, \ldots, x_t) \in (\mathsf{X} \times \mathsf{A})^t \times \mathsf{X}$, the agent chooses an action from the set $A \in \mathcal{B}(\mathsf{A})$ with probability $\pi_t(h_t, A)$. We denote the set of all policies by Π.

Let Ω be the infinite Cartesian product $(\mathsf{X} \times \mathsf{A})^{\mathbb{N}}$, and $\mathscr{F}$ be the product σ-algebra on it. Given an MDP, with the notation as above, and a policy $\pi \in \Pi$, Ionescu-Tulcea theorem (Neveu, 1965; Hernandez-Lerma & Lasserre, 2012) tells us that there exists a unique probability measure $\mathbb{P}_\rho^\pi$ on $(\Omega, \mathscr{F})$ such that we can define the random variables $(X_t)_{t \geq 0}$ and $(A_t)_{t \geq 0}$ using the projection maps $\Omega \ni \omega = (x_0, a_0, x_1, a_1, \ldots) \mapsto X_t(\omega) = x_t \in \mathsf{X}$, and $\Omega \ni \omega = (x_0, a_0, x_1, a_1, \ldots) \mapsto A_t(\omega) = a_t \in \mathsf{A}$, that satisfy the following transition rules:

$$\mathbb{P}_\rho^\pi(X_0 \in B) = \rho(B),\ \forall\, B \in \mathcal{B}(\mathsf{X}), \quad \mathbb{P}_\rho^\pi(A_t \in C \,|\, H_t) = \pi_t(H_t, C),\ \forall\, C \in \mathcal{B}(\mathsf{A}),\ \text{and}$$
$$\mathbb{P}_\rho^\pi(X_{t+1} \in B \,|\, H_t, A_t) = K((X_t, A_t), B), \quad \forall\, B \in \mathcal{B}(\mathsf{X}),$$

where $H_t = (X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t)$ is the history random variable. If $\rho = \delta_x$, for $x \in \mathsf{X}$, is a Dirac measure, then we use the notation $\mathbb{P}_x^\pi$ instead of the cumbersome $\mathbb{P}_{\delta_x}^\pi$.

Infinite Horizon Setting In the setting of infinite horizon total discounted rewards, there is a discount factor $\gamma \in [0,1)$, and the goal is to find a policy $\pi \in \Pi$ that maximizes the expected sum of discounted rewards $J(\rho, \pi) = \mathbb{P}_\rho^\pi[\sum_{t=0}^\infty \gamma^t r(X_t, A_t)]$. We have explicitly shown the dependence of J on the initial state distribution ρ and the policy π; of course, J depends on other parameters of the MDP, like the transition kernel K, but we keep that dependence implicit. A policy $\pi^* \in \Pi$ satisfying

$$J(\mu, \pi^*) = \sup_{\pi \in \Pi} J(\mu, \pi),\ \forall\, \mu \in \mathcal{M}_1(\mathsf{X}), \tag{1}$$

is said to be optimal. It is well known (Blackwell, 1965; Bertsekas & Shreve, 1978; Puterman, 1994) that under this setting there exists a stationary policy that is optimal. A policy $\pi \in \Pi$ is stationary if there exists a stochastic kernel φ from X to A such that $\pi_t = \varphi$. Value functions are defined as

$$V^\pi(x) = J(\delta_x, \pi), \quad Q^\pi(x,a) = r(x,a) + \gamma \int_{\mathsf{X}} V^\pi(y) K((x,a), \mathrm{d}y), \quad \forall\, (x,a) \in \mathsf{X} \times \mathsf{A}.$$

Finite Horizon Setting In the setting of finite horizon total rewards, there is a horizon parameter $T \in \mathbb{N}$, and the goal is to find a policy $\pi \in \Pi$ that maximizes the expected sum of total rewards $J(\rho, \pi) = \mathbb{P}_\rho^\pi\left[\sum_{t=0}^T r(X_t, A_t)\right]$. It should be clear from context whether J denotes the objective

function for the infinite horizon or the finite horizon setting. A policy π^* satisfying

$$J(\mu, \pi^*) = \sup_{\pi \in \Pi} J(\mu, \pi), \ \forall \, \mu \in \mathcal{M}_1(\mathsf{X}), \tag{2}$$

is said to be optimal. It is well known (Bellman, 1957; Bertsekas & Shreve, 1978; Puterman, 1994) that under this setting there exists a Markov policy that is optimal; a policy $\pi \in \Pi$ is Markov if each π_t is a stochastic kernel from X to A. In the finite horizon setting, value functions are defined for each time step $t \in [T] = \{0, 1, \ldots, T\}$ as

$$V_t^\pi(x) = \mathbb{P}_\rho^\pi \left[\sum_{s=t}^{T} r(X_s, A_s) \,\middle|\, X_t = x \right], \quad Q_t^\pi(x, a) = \mathbb{P}_\rho^\pi \left[\sum_{s=t}^{T} r(X_s, A_s) \,\middle|\, X_t = x, A_t = a \right].$$

3 Problem Dependent Bounds

We begin by highlighting that (1) and (2) define optimality in terms of all possible initial state distributions. However, typical reinforcement learning problems optimize performance with respect to a specific initial distribution. This is important because a policy $\pi^\dagger$ that satisfies $J(\rho, \pi^\dagger) = \sup_{\pi \in \Pi} J(\rho, \pi)$ for a fixed ρ may not be optimal. This may not matter if during the test time the initial state is sampled from ρ, but this is rarely the case.

3.1 Infinite Horizon Setting

To gain an understanding of how the initial state distribution affects the objective function, we analyze the difference $J(\rho, \pi) - J(\mu, \pi)$, keeping the policy π fixed but varying the initial distributions ρ and μ. We aim to bound this difference in terms of the distance between ρ and μ, and some property of the MDP. To this end, we define the state-action occupancy measure:

$$\psi_\rho^\pi(B, a) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\rho^\pi(X_t \in B, A_t = a), \quad \forall \, (B, a) \in \mathcal{B}(\mathsf{X}) \times \mathsf{A}.$$

It is easy to show that $\psi_\rho^\pi \in \mathcal{M}_1(\mathsf{X} \times \mathsf{A})$. Intuitively, $\psi_\rho^\pi(B, a)$ can be viewed as the probability that the agent occupies a state in B and takes action a from that state, where future occurrences are discounted.

Lemma 3.1 (Performance difference lemma (infinite horizon))**.** *Under the infinite horizon setting, the performance difference $J(\rho, \pi) - J(\mu, \pi)$ for a fixed policy π can be expressed*

(a) in terms of occupancy measures:

$$J(\rho, \pi) - J(\mu, \pi) = \frac{1}{1 - \gamma} (\psi_\rho^\pi - \psi_\mu^\pi) r,$$

where $\psi_\rho^\pi - \psi_\mu^\pi$ is a signed measure, and $(\psi_\rho^\pi - \psi_\mu^\pi) r$ denotes the integral of r with respect to this signed measure, and

(b) as a function of the distance between ρ and μ:

$$J(\rho, \pi) - J(\mu, \pi) \leq \frac{r_{max}}{1 - \gamma} |\rho - \mu|(\mathsf{X}) = \frac{2 r_{max}}{1 - \gamma} d_{TV}(\rho, \mu),$$

where $|\rho - \mu|$ denotes the total variation measure of the signed measure $\rho - \mu$, and d_{TV} denotes the total variation metric.

Proof. (a) We begin by writing the objective function J in terms of occupancy measure. Observe that, for each $n \geq 0$, the partial sum $\sum_{t=0}^{n} \gamma^t r(X_t, A_t)$ is dominated by $\sum_{t=0}^{n} \gamma^t r_{\max}$, which in turn converges to $\frac{r_{\max}}{1-\gamma}$ whose integral with respect to $\mathbb{P}_\rho^\pi$ is finite. We can now use Lebesgue's dominated convergence theorem to get

$$J(\rho, \pi) = \mathbb{P}_\rho^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \right] = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\rho^\pi r(X_t, A_t).$$

Now for each $t \geq 0$, let $\mathbb{P}^\pi_{\rho,t} = \mathbb{P}^\pi_\rho \circ \phi_t^{-1}$ denote the pushforward of the measure $\mathbb{P}^\pi_\rho$ with respect to the projection map $\Omega \ni \omega = (x_0, a_0, x_1, a_1, \ldots) \mapsto \phi_t(\omega) = (x_t, a_t) \in \mathsf{X} \times \mathsf{A}$. Note that $\phi_t = (X_t, A_t)$. We can now write

$$\mathbb{P}^\pi_\rho r(X_t, A_t) = \int_{\mathsf{X}\times\mathsf{A}} r(x,a) \mathbb{P}^\pi_{\rho,t}(\mathrm{d}x, \mathrm{d}a).$$

This notation allows us to write the state-action occupancy measure as

$$\psi^\pi_\rho = (1-\gamma) \sum_{t=0}^\infty \gamma^t \mathbb{P}^\pi_{\rho,t}. \tag{3}$$

Now write the infinite sum as

$$\sum_{t=0}^\infty \gamma^t \mathbb{P}^\pi_\rho r(X_t, A_t) = \sum_{t=0}^\infty \gamma^t \int_{\mathsf{X}\times\mathsf{A}} r(x,a) \mathbb{P}^\pi_{\rho,t}(\mathrm{d}x, \mathrm{d}a).$$

We can now employ Fubini's theorem to interchange the sum and the integral

$$\sum_{t=0}^\infty \gamma^t \int_{\mathsf{X}\times\mathsf{A}} r(x,a) \mathbb{P}^\pi_{\rho,t}(\mathrm{d}x, \mathrm{d}a) = \int_{\mathsf{X}\times\mathsf{A}} r(x,a) \sum_{t=0}^\infty \gamma^t \mathbb{P}^\pi_{\rho,t}(\mathrm{d}x, \mathrm{d}a).$$

Finally, use (3) to get

$$J(\rho, \pi) = \frac{1}{1-\gamma} \int_{\mathsf{X}\times\mathsf{A}} r(x,a) \psi^\pi_\rho(\mathrm{d}x, \mathrm{d}a).$$

The difference calculation now follows directly:

$$J(\rho,\pi) - J(\mu,\pi) = \frac{1}{1-\gamma} \int_{\mathsf{X}\times\mathsf{A}} r(x,a)(\psi^\pi_\rho - \psi^\pi_\mu)(\mathrm{d}x, \mathrm{d}a) = \frac{1}{1-\gamma}(\psi^\pi_\rho - \psi^\pi_\mu) r.$$

(b) We begin by writing the objective function J as an expectation of the value function with respect to the initial state distribution: $J(\rho,\pi) = \rho V^\pi$, which follows directly from their definitions. We can therefore write

$$J(\rho,\pi) - J(\mu,\pi) = \rho V^\pi - \mu V^\pi = (\rho - \mu) V^\pi,$$

where the last expression denotes the integral of V^π with respect to the signed measure $\rho - \mu$. Now note that V^π is bounded because for all $x \in \mathsf{X}$,

$$V^\pi(x) = \mathbb{P}^\pi_x \left[\sum_{t=0}^\infty \gamma^t r(X_t, A_t) \right] \leq \mathbb{P}^\pi_x \left[\sum_{t=0}^\infty \gamma^t r_{\max} \right] = \frac{r_{\max}}{1-\gamma}.$$

This allows us to write

$$(\rho - \mu) V^\pi \leq |\rho - \mu| |V^\pi| \leq \frac{r_{\max}}{1-\gamma} |\rho - \mu|(\mathsf{X}).$$

The first inequality above follows from the following argument: denote $\nu = \rho - \mu$, and let $\nu = \nu^+ - \nu^-$ be the Jordan decomposition of the signed measure ν into positive measures ν^+ and ν^- that are mutually singular. Then

$$\nu V^\pi \leq |\nu V^\pi| = \left| \nu^+ V^\pi - \nu^- V^\pi \right| \leq \left| \nu^+ V^\pi \right| + \left| \nu^- V^\pi \right| \leq \nu^+ |V^\pi| + \nu^- |V^\pi| = |\nu| |V^\pi|.$$

We finally need the standard result that $|\rho - \mu|(\mathsf{X}) = 2 d_{\mathrm{TV}}(\rho, \mu)$. Suppose $\mathsf{X} = P \cup N$ is a Hahn decomposition for ν, i.e., $P \cap N = \varnothing$, $\nu(E) \geq 0$ for all measurable $E \subseteq P$, and $\nu(F) \leq 0$ for all measurable $F \subseteq N$. Then using the fact that $\nu^+(\mathsf{X}) - \nu^-(\mathsf{X}) = \nu(\mathsf{X}) = \rho(\mathsf{X}) - \mu(\mathsf{X}) = 1 - 1 = 0$ at step $\diamond$, we get

$$\begin{aligned} d_{\mathrm{TV}}(\rho,\mu) &= \sup_{B \in \mathcal{B}(\mathsf{X})} |\rho(B) - \mu(B)| = |\rho(P) - \mu(P)| = \nu^+(P) + \nu^-(P) \\ &= \nu^+(P) = \nu^+(\mathsf{X}) \overset{\diamond}{=} \frac{\nu^+(\mathsf{X}) + \nu^-(\mathsf{X})}{2} = \frac{1}{2} |\nu|(\mathsf{X}). \end{aligned}$$

□

Interpreting our results Part (a) tells us that if the state-action occupancy measures for the two initial state distributions are close, then we should expect the objective functions to also be close; in MDPs with ergodic $(X_t)_{t\geq 0}$, for example, this can happen if the mixing time is small because the agent quickly "forgets" about the initial state.

Part (b) gives us an upper bound on the difference $J(\rho,\pi) - J(\mu,\pi)$ as a function of the distance between ρ and μ. This inequality is tight. To see this, consider the case where the supports of ρ and μ are disjoint. Then the total variation measure $|\rho - \mu|$ equals $\rho + \mu$, and thus $|\rho - \mu|(\mathsf{X}) = 2$. This gives $J(\rho,\pi) - J(\mu,\pi) \leq \frac{2r_{\max}}{1-\gamma}$. This inequality is an equality if (i) the state space consists of two disconnected components $\mathsf{X} = \mathsf{X}_+ \cup \mathsf{X}_-$ such that all the positive rewards are in X_+ and all the negative rewards are in X_-, (ii) we let ρ be supported on X_+ and μ be supported on X_-, and (iii) we let π be the policy that is optimal on X_+ and is pessimal on X_-. Then we can make the reward function such that $J(\rho,\pi) = \frac{r_{\max}}{1-\gamma}$ and $J(\mu,\pi) = -\frac{r_{\max}}{1-\gamma}$.

On the other hand, the more ρ and μ put measure on the same parts of the state space, the smaller the total variation measure $|\rho - \mu|$ and the total variation distance $d_{\mathrm{TV}}(\rho,\mu)$ will be, and thus the smaller the difference $J(\rho,\pi) - J(\mu,\pi)$ will be.

Contrast this lemma with the classical performance difference lemma (Kakade & Langford, 2002; Agarwal et al., 2022; Foster & Rakhlin, 2023), where the difference $J(\rho,\pi) - J(\rho,\eta)$ keeps the initial state distribution fixed and varies the policy, and is expressed as

$$J(\rho,\pi) - J(\rho,\eta) = \frac{1}{1-\gamma}\psi_\rho^\pi A^\eta,$$

where $A^\eta(x,a) = Q^\eta(x,a) - V^\eta(x)$ denotes the advantage function.

3.2 Finite Horizon Setting

The analysis for the finite horizon setting follows similarly as in the infinite horizon setting, except the analysis here is simpler due to finite sums. We define the state-action occupancy measure as follows:

$$\psi_{\rho,t}^\pi(B,a) = \mathbb{P}_\rho^\pi(X_t \in B, A_t = a), \quad \forall\, (B,a) \in \mathcal{B}(\mathsf{X}) \times \mathsf{A},\, t \in [T],$$

$$\psi_\rho^\pi(B,a) = \frac{1}{T+1}\sum_{t=0}^{T} \psi_{\rho,t}^\pi(B,a), \quad \forall\, (B,a) \in \mathcal{B}(\mathsf{X}) \times \mathsf{A}.$$

Lemma 3.2 (Performance difference lemma (finite horizon))**.** *Under the finite horizon setting, we can write*

(a) the difference in terms of occupancy measure:

$$J(\rho,\pi) - J(\mu,\pi) = (T+1)(\psi_\rho^\pi - \psi_\mu^\pi)r, \text{ and}$$

(b) the difference expressed as an upper bound:

$$|J(\rho,\pi) - J(\mu,\pi)| \leq (T+1)r_{max}|\rho - \mu|(\mathsf{X}) = 2(T+1)r_{max}d_{TV}(\rho,\mu).$$

Proof. (a) Just like in the infinite horizon case, we write the objective function J as an integral of the reward function with respect to the occupancy measure, which then immediately gives us the desired equation.

$$J(\rho,\pi) = \mathbb{P}_\rho^\pi\left[\sum_{t=0}^{T} r(X_t,A_t)\right] = \sum_{t=0}^{T}\mathbb{P}_\rho^\pi r(X_t,A_t) = \sum_{t=0}^{T}\int_{\mathsf{X}\times\mathsf{A}} r(x,a)\mathbb{P}_{\rho,t}^\pi(\mathrm{d}x,\mathrm{d}a)$$

$$= \int_{\mathsf{X}\times\mathsf{A}} r(x,a)\sum_{t=0}^{T}\mathbb{P}_{\rho,t}^\pi(\mathrm{d}x,\mathrm{d}a) = (T+1)\int_{\mathsf{X}\times\mathsf{A}} r(x,a)\psi_\rho^\pi(\mathrm{d}x,\mathrm{d}a) = (T+1)\psi_\rho^\pi r.$$

(b) Just like in the infinite horizon case, we begin by writing the objective function J as $J(\rho,\pi) = \rho V_0^\pi$. This gives $J(\rho,\pi) - J(\mu,\pi) = (\rho-\mu)V_0^\pi$. Now since $V_0^\pi(x) \le (T+1)r_{\max}$, we get

$$(\rho-\mu)V_0^\pi \le |\rho-\mu|\,|V_0^\pi| \le (T+1)r_{\max}|\rho-\mu|(\mathsf{X}).$$

□

Here, the difference $J(\rho,\pi) - J(\mu,\pi)$ scales with with the horizon length T, while in the infinite horizon case the difference scales with the effective horizon $\frac{1}{1-\gamma}$.

4 Discussion and Conclusion

In this paper we focused on the effect of initial state distribution on the objective function for the settings of infinite horizon with discounted rewards and finite horizon. We provided performance difference lemmas that quantify the impact of changing the initial state distributions while keeping the policy fixed. These lemmas contrast with the classical performance difference lemma that varies the policy but keeps the distribution fixed. The analysis in this paper could provide RL researchers with insights to design algorithms for robust policy learning under uncertain initial distributions and adaptive schemes that adjust to different test-time distributions.

A natural next step is to extend our analysis to the average reward setting, which models long-term behavior more faithfully than the discounted reward setting. Here the aim is to find a policy that maximizes:

$$J(\rho,\pi) = \lim_{N\to\infty} \frac{1}{N}\mathbb{P}_\rho^\pi\left[\sum_{t=0}^{N} r(X_t,A_t)\right],$$

where the limit is replaced by either $\limsup$ or $\liminf$ if it does not exist (Arapostathis et al., 1993).

Another promising direction is to bound the difference $J(\rho,\pi) - J(\mu,\pi)$ in terms of the "hardness" of the MDP, in the spirit of Maillard et al. (2014). One can quantify the hardness in several ways. Given an initial state distribution ρ and a policy π, define

$$\mathfrak{h}_\rho(V^\pi) = \int_{\mathsf{X}\times\mathsf{A}} \left[\int_{\mathsf{X}} \left(V^\pi(y) - \int_{\mathsf{X}} V^\pi(z)K((x,a),\mathrm{d}z)\right)^2 K((x,a),\mathrm{d}y)\right]^{\frac{1}{2}} \psi_\rho^\pi(\mathrm{d}x,\mathrm{d}a),$$

and the worst-case analogue

$$\mathfrak{h}(V^\pi) = \sup_{x\in\mathsf{X}} \left[\int_{\mathsf{X}} \left(V^\pi(y) - \int_{\mathsf{X}} V^\pi(z)K^\pi(x,\mathrm{d}z)\right)^2 K^\pi(x,\mathrm{d}y)\right]^{\frac{1}{2}},$$

where $K^\pi(x,B) = \int_{\mathsf{A}} K((x,a),B)\pi(x,\mathrm{d}a)$, $B\in\mathcal{B}(\mathsf{X})$ defines a stochastic kernel from X to X. Here $\mathfrak{h}_\rho(V^\pi)$ measures the average (with respect to the occupancy measure) standard deviation of the value of the next state. Intuitively speaking, $\mathfrak{h}_\rho(V^\pi)$ is large if, on average, the agent is uncertain about the next state's value, and thus the MDP is hard. As for $\mathfrak{h}(V^\pi)$, the quantity inside the square brackets measures the variance of the value of the next state if the agent is in state $x\in\mathsf{X}$ and takes action according to policy π. $\mathfrak{h}_\rho$ models hardness using the average uncertainty, while $\mathfrak{h}$ models the hardness using the worst-case uncertainty.

References

Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. 2022. URL https://rltheorybook.github.io/.

Aristotle Arapostathis, Vivek S. Borkar, Emmanuel Fernández-Gaucherand, Mrinal K. Ghosh, and Steven I. Marcus. Discrete-time controlled markov processes with average cost criterion: A survey. *SIAM Journal on Control and Optimization*, 31(2):282–344, March 1993. DOI: 10.1137/0331018.

Richard Bellman. *Dynamic Programming*. A RAND Corporation Research Study. Princeton University Press, Princeton, NJ, 1957. ISBN 0-691-07951-X. Sixth printing, 1972.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 41–48, Montreal, Quebec, Canada, 2009. ACM.

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, Inc., USA, 1978. ISBN 0120932601.

David Blackwell. Discounted dynamic programming. *The Annals of Mathematical Statistics*, 36(1): 226–235, February 1965. DOI: 10.1214/aoms/1177700291.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. First return, then explore. *Nature*, 590:580–586, 2021. DOI: 10.1038/s41586-020-03157-9.

Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Proceedings of the Conference on Robot Learning (CoRL)*, pp. 482–495. PMLR, November 2017.

Dylan J. Foster and Alexander Rakhlin. *Foundations of Reinforcement Learning and Interactive Decision Making*. 2023. URL https://arxiv.org/abs/2312.16730.

O. Hernandez-Lerma and J.B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Stochastic Modelling and Applied Probability. Springer New York, 2012. ISBN 9781461207290.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. Draft version; under review. Submitted to *Statistical Science*., 2025. URL https://nanjiang.cs.illinois.edu/files/STS_Special_Issue_Offline_RL.pdf.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 267–274, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer Cham, 3rd edition, 2021. DOI: 10.1007/978-3-030-61871-1.

Akshay Krishnamurthy, Gene Li, and Ayush Sekhari. The role of environment access in agnostic reinforcement learning, 2025. URL https://arxiv.org/abs/2504.05405.

Odalric-Ambrym Maillard, Timothy A. Mann, and Shie Mannor. How hard is my mdp?" the distribution-norm to the rescue". In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/7335f569596c706ccdf756fc8f812a94-Paper.pdf.

Zakaria Mhammedi, Dylan J. Foster, and Alexander Rakhlin. The power of resets in online reinforcement learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 12334–12407. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/16f8a0852b31bc9dc791ecf313247a57-Paper-Conference.pdf.

Jacques Neveu. *Mathematical Foundations of the Calculus of Probability*. Holden–Day, San Francisco, 1965. Translated by Amiel Feinstein; foreword by R. Fortet.

Chris Preston. *Some Notes on Standard Borel and Related Spaces*. 2008. URL https://arxiv.org/abs/0809.3066.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Csaba Szepesvári. Theoretical foundations of reinforcement learning – lecture notes, 2025. URL https://rltheory.github.io/. Course CMPUT 605/653.

Arash Tavakoli, Vitaly Levdik, Riashat Islam, Christopher M. Smith, and Petar Kormushev. Exploring restart distributions. In *Proceedings of the 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, Montréal, Canada, 2019.

Dong Yin, Sridhar Thiagarajan, Nevena Lazic, Nived Rajaraman, Botao Hao, and Csaba Szepesvári. Sample efficient deep reinforcement learning via local planning. *CoRR*, abs/2301.12579, 2023. DOI: 10.48550/ARXIV.2301.12579. URL https://doi.org/10.48550/arXiv.2301.12579.