Lipschitz Lifelong Reinforcement Learning

Erwan Lecarpentier[1, 2], **David Abel**[3], **Kavosh Asadi**[3, 4*], **Yuu Jinnai**[3],
Emmanuel Rachelson[1], **Michael L. Littman**[3]
[1]ISAE-SUPAERO, Université de Toulouse, France
[2]ONERA, The French Aerospace Lab, Toulouse, France
[3]Brown University, Providence, Rhode Island, USA
[4]Amazon Web Service, Palo Alto, California, USA
erwanlecarpentier@mailbox.org

Abstract

We consider the problem of knowledge transfer when an agent is facing a series of Reinforcement Learning (RL) tasks. We introduce a novel metric between Markov Decision Processes and establish that close MDPs have close optimal value functions. Formally, the optimal value functions are Lipschitz continuous with respect to the tasks space. These theoretical results lead us to a value-transfer method for Lifelong RL, which we use to build a PAC-MDP algorithm with improved convergence rate. Further, we show the method to experience no negative transfer with high probability. We illustrate the benefits of the method in Lifelong RL experiments.

1 Introduction

Lifelong Reinforcement Learning (RL) is an online problem where an agent faces a series of RL tasks, drawn sequentially. Transferring knowledge from prior experience to speed up the resolution of new tasks is a key question in that setting (Lazaric 2012; Taylor and Stone 2009). We elaborate on the intuitive idea that *similar* tasks should allow a large amount of transfer. An agent able to compute online a similarity measure between source tasks and the current target task could be able to perform transfer accordingly. By measuring the amount of inter-task similarity, we design a novel method for value transfer, practically deployable in the online Lifelong RL setting. Specifically, we introduce a metric between MDPs and prove that the optimal Q-value function is Lipschitz continuous with respect to the MDP space. This property makes it possible to compute a provable upper bound on the optimal Q-value function of an unknown target task, given the learned optimal Q-value function of a source task. Knowing this upper bound accelerates the convergence of an RMax-like algorithm (Brafman and Tennenholtz 2002), relying on an optimistic estimate of the optimal Q-value function. Overall, the proposed transfer method consists of computing online the distance between source and target tasks, deducing the upper bound on the optimal Q value function of the source task and using this bound to accelerate learning. Importantly, the method exhibits no negative transfer, *i.e.*, it cannot cause performance degradation, as the computed upper bound provably does not underestimate the optimal Q-value function.

Our contributions are as follows. First, we study theoretically the Lipschitz continuity of the optimal Q-value function in the task space by introducing a metric between MDPs (Section 3). Then, we use this continuity property to propose a value-transfer method based on a local distance between MDPs (Section 4). Full knowledge of both MDPs is not required and the transfer is non-negative, which makes the method applicable online and safe. In Section 4.3, we build a PAC-MDP algorithm called *Lipschitz RMax*, applying this transfer method in the online Lifelong RL setting. We provide sample and computational complexity bounds and showcase the algorithm in Lifelong RL experiments (Section 5).

2 Background and Related Work

Reinforcement Learning (RL) (Sutton and Barto 2018) is a framework for sequential decision making. The problem is typically modeled as a Markov Decision Process (MDP) (Puterman 2014) consisting of a 4-tuple $\langle \mathcal{S}, \mathcal{A}, R, T \rangle$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ an action space, R_s^a is the expected reward of taking action a in state s and $T_{ss'}^a$ is the transition probability of reaching state s' when taking action a in state s. Without loss of generality, we assume $R_s^a \in [0, 1]$. Given a discount factor $\gamma \in [0, 1)$, the expected cumulative return $\sum_t \gamma^t R_{s_t}^{a_t}$ obtained along a trajectory starting with state s and action a using policy π in MDP M is denoted by $Q_M^\pi(s, a)$ and called the Q-function. The optimal Q-function Q_M^* is the highest attainable expected return from s, a and $V_M^*(s) = \max_{a \in \mathcal{A}} Q_M^*(s, a)$ is the optimal value function in s. Notice that $R_s^a \leq 1$ implies $Q_M^*(s, a) \leq \frac{1}{1-\gamma}$ for all $s, a \in \mathcal{S} \times \mathcal{A}$. This maximum upper bound is used by the RMax algorithm as an optimistic initialization of the learned Q function. A key point to reduce the sample complexity of this algorithm is to benefit from a tighter upper bound, which is the purpose of our transfer method.

Lifelong RL (Silver, Yang, and Li 2013; Brunskill and Li 2014) is the problem of experiencing online a series of MDPs drawn from an unknown distribution. Each time an MDP is sampled, a classical RL problem takes place where the agent is able to interact with the environment to maximize its expected return. In this setting, it is reasonable to think that knowledge gained on previous MDPs could be re-used

[*] Kavosh Asadi finished working on this project before joining Amazon.

to improve the performance in new MDPs. In this paper, we provide a novel method for such transfer by characterizing the way the optimal Q-function can evolve across tasks. As commonly done (Wilson et al. 2007; Brunskill and Li 2014; Abel et al. 2018), we restrict the scope of the study to the case where sampled MDPs share the same state-action space $\mathcal{S} \times \mathcal{A}$. For brevity, we will refer indifferently to MDPs, models or tasks, and write them $M = \langle R, T \rangle$.

Using a metric between MDPs has the appealing characteristic of quantifying the amount of similarity between tasks, which intuitively should be linked to the amount of transfer achievable. Song et al. (2016) define a metric based on the bi-simulation metric introduced by Ferns, Panangaden, and Precup (2004) and the Wasserstein metric (Villani 2008). value transfer is performed between states with low bi-simulation distances. However, this metric requires knowing both MDPs completely and is thus unusable in the Lifelong RL setting where we expect to perform transfer before having learned the current MDP. Further, the transfer technique they propose does allow negative transfer (see Appendix, Section 1). Carroll and Seppi (2005) also define a value-transfer method based on a measure of similarity between tasks. However, this measure is not computable online and thus not applicable to the Lifelong RL setting. Mahmud et al. (2013) and Brunskill and Li (2013) propose MDP clustering methods; respectively using a metric quantifying the regret of running the optimal policy of one MDP in the other MDP and the $\mathcal{L}_1$ norm between the MDP models. An advantage of clustering is to prune the set of possible source tasks. They use their approach for policy transfer, which differs from the value-transfer method proposed in this paper. Ammar et al. (2014) learn the model of a source MDP and view the prediction error on a target MDP as a dissimilarity measure in the task space. Their method makes use of samples from both tasks and is not readily applicable to the online setting considered in this paper. Lazaric, Restelli, and Bonarini (2008) provide a practical method for sample transfer, computing a similarity metric reflecting the probability of the models to be identical. Their approach is applicable in a batch RL setting as opposed to the online setting considered in this paper. The approach developed by Sorg and Singh (2009) is very similar to ours in the sense that they prove bounds on the optimal Q-function for new tasks, assuming that both MDPs are known and that a soft homomorphism exists between the state spaces. Brunskill and Li (2013) also provide a method that can be used for Q-function bounding in multi-task RL.

Abel et al. (2018) present the MaxQInit algorithm, providing transferable bounds on the Q-function with high probability while preserving PAC-MDP guarantees (Strehl, Li, and Littman 2009). Given a set of solved tasks, they derive the probability that the maximum over the Q-values of previous MDPs is an upper bound on the current task's optimal Q-function. This approach results in a method for non-negative transfer with high probability once enough tasks have been sampled. The method developed by Abel et al. (2018) is similar to ours in two fundamental points: first, a theoretical upper bounds on optimal Q-values across the MDP space is built; secondly, this provable upper bound is used to transfer knowledge between MDPs by replacing the maximum $\frac{1}{1-\gamma}$ bound in an RMax-like algorithm, providing PAC guarantees. The difference between the two approaches is illustrated in Figure 1, where the MaxQInit bound is the one developed by Abel et al. (2018), and the LRMax bound is the one we present in this paper. On this figure, the essence of the LRMax bound is noticeable. It stems from the fact that the optimal Q value function is locally Lipschitz continuous in the MDP space w.r.t. a specific pseudometric. Confirming the intuition, close MDPs w.r.t. this metric have close optimal Q values. It should be noticed that no bound is uniformly better than the other as intuited by Figure 1. Hence, combining all the bounds results in a tighter upper bound as we will illustrate in experiments (Section 5). We first carry out the theoretical characterization of the Lipschitz continuity properties in the following section. Then, we build on this result to propose a practical transfer method for the online Lifelong RL setting.

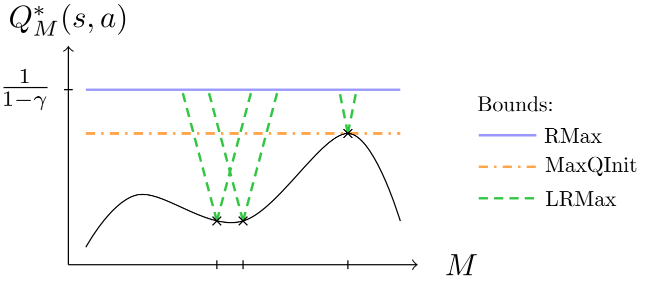


Figure 1: The optimal Q-value function represented for a particular s, a pair across the MDP space. The RMax, MaxQInit and LRMax bounds are represented for three sampled MDPs.

3 Lipschitz Continuity of Q-Functions

The intuition we build on is that similar MDPs should have similar optimal Q-functions. Formally, this insight can be translated into a continuity property of the optimal Q-function over the MDP space $\mathcal{M}$. The remainder of this section mathematically formalizes this intuition that will be used in the next section to derive a practical method for value transfer. To that end, we introduce a local pseudometric characterizing the distance between the models of two MDPs at a particular state-action pair. A reminder and a detailed discussion on the metrics used herein can be found in the Appendix, Section 2.

Definition 1. *Given two tasks $M = \langle R, T \rangle$, $\bar{M} = \langle \bar{R}, \bar{T} \rangle$, and a function $f : \mathcal{S} \to \mathbb{R}^+$, we define the* pseudometric between models *at $(s, a) \in \mathcal{S} \times \mathcal{A}$ w.r.t. f as:*

$$D^f_{sa}(M, \bar{M}) \triangleq |R^a_s - \bar{R}^a_s| + \sum_{s' \in \mathcal{S}} f(s') |T^a_{ss'} - \bar{T}^a_{ss'}|. \quad (1)$$

This pseudometric is relative to a positive function f. We implicitly cast this definition in the context of discrete state spaces. The extension to continuous spaces is straightforward but beyond the scope of this paper. For the sake of clarity in the remainder of this study, we introduce

$$D_{sa}(M \| \bar{M}) \triangleq D^{\gamma V^*_{\bar{M}}}_{sa}(M, \bar{M}),$$

corresponding to the pseudometric between models with the particular choice of $f = \gamma V^*_{\bar{M}}$. From this definition stems the following pseudo-Lipschitz continuity result.

Proposition 1 (Local pseudo-Lipschitz continuity)**.** *For two MDPs $M, \bar{M}$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\left|Q^*_M(s,a) - Q^*_{\bar{M}}(s,a)\right| \leq \Delta_{sa}(M, \bar{M}), \qquad (2)$$

with the local MDP pseudometric $\Delta_{sa}(M,\bar{M}) \triangleq \min\left\{d_{sa}(M\|\bar{M}), d_{sa}(\bar{M}\|M)\right\}$*, and the* local MDP dissimilarity $d_{sa}(M\|\bar{M})$ *is the unique solution to the following fixed-point equation for d_{sa}:*

$$d_{sa} = D_{sa}(M\|\bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} d_{s'a'}, \forall s, a. \qquad (3)$$

All the proofs of the paper can be found in the Appendix. This result establishes that the distance between the optimal Q-functions of two MDPs at $(s,a) \in \mathcal{S} \times \mathcal{A}$ is controlled by a local dissimilarity between the MDPs. The latter follows a fixed-point equation (Equation 3), which can be solved by Dynamic Programming (DP) (Bellman 1957). Note that, although the local MDP dissimilarity $d_{sa}(M\|\bar{M})$ is asymmetric, $\Delta_{sa}(M,\bar{M})$ *is* a pseudometric, hence the name *pseudo-Lipschitz continuity*. Notice that the policies in Equation 2 are the optimal ones for the two MDPs and thus are different. Proposition 1 is a mathematical result stemming from Definition 1 and should be distinguished from other frameworks of the literature that *assume* the continuity of the reward and transition models w.r.t. $\mathcal{S} \times \mathcal{A}$ (Rachelson and Lagoudakis 2010; Pirotta, Restelli, and Bascetta 2015; Asadi, Misra, and Littman 2018).This result establishes that the optimal Q-functions of two close MDPs, in the sense of Equation 1, are themselves close to each other. Hence, given $Q^*_{\bar{M}}$, the function

$$s, a \mapsto Q^*_{\bar{M}}(s,a) + \Delta_{sa}(M, \bar{M}) \qquad (4)$$

can be used as an upper bound on Q^*_M with M an unknown MDP. This is the idea on which we construct a computable and transferable upper bound in Section 4. In Figure 1, the upper bound of Equation 4 is represented by the LRMax bound. Noticeably, we provide a global pseudo-Lipschitz continuity property, along with similar results for the optimal value function V^*_M and the value function of a fixed policy. As these results do not directly serve the purpose of this article, we report them in the Appendix, Section 4.

4 Transfer Using the Lipschitz Continuity

A purpose of value transfer, when interacting online with a new MDP, is to initialize the value function and drive the exploration to accelerate learning. We aim to exploit value transfer in a method guaranteeing three conditions:

C1. the resulting algorithm is PAC-MDP;

C2. the transfer accelerates learning;

C3. the transfer is non-negative.

To achieve these conditions, we first present a transferable upper bound on Q^*_M in Section 4.1. This upper bound stems from the Lipschitz continuity result of Proposition 1. Then, we propose a practical way to *compute* this upper bound in Section 4.2. Precisely, we propose a surrogate bound that can be calculated online in the Lifelong RL setting, without having explored the source and target tasks completely. Finally, we implement the method in an algorithm described in Section 4.3, and demonstrate formally that it meets conditions C1, C2 and C3. Improvements are discussed in Section 4.4.

4.1 A Transferable Upper Bound on Q^*_M

From Proposition 1, one can naturally define a local upper bound on the optimal Q-function of an MDP given the optimal Q-function of another MDP.

Definition 2. *Given two tasks M and $\bar{M}$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, the* Lipschitz upper bound on Q^*_M induced by $Q^*_{\bar{M}}$ *is defined as $U_{\bar{M}}(s,a) \geq Q^*_M(s,a)$ with:*

$$U_{\bar{M}}(s,a) \triangleq Q^*_{\bar{M}}(s,a) + \Delta_{sa}(M, \bar{M}). \qquad (5)$$

The *optimism in the face of uncertainty* principle leads to considering that the long-term expected return from any state is the $\frac{1}{1-\gamma}$ maximum return, unless proven otherwise. Particularly, the RMax algorithm (Brafman and Tennenholtz 2002), explores an MDP so as to shrink this upper bound. RMax is a model-based, online RL algorithm with PAC-MDP guarantees (Strehl, Li, and Littman 2009), meaning that convergence to a near-optimal policy is guaranteed in a polynomial number of missteps with high probability. It relies on an optimistic model initialization that yields an optimistic upper bound U on the optimal Q-function, then acts greedily w.r.t. U. By default, it takes the maximum value $U(s,a) = \frac{1}{1-\gamma}$, but any tighter upper bound is admissible. Thus, shrinking U with Equation 5 is expected to improve the learning speed or sample complexity for new tasks in Lifelong RL.

In RMax, during the resolution of a task M, $\mathcal{S} \times \mathcal{A}$ is split into a subset of known state-action pairs K and its complement K^c of unknown pairs. A state-action pair is known if the number of collected reward and transition samples allows estimating an ϵ-accurate model in $\mathcal{L}_1$-norm with probability higher than $1-\delta$. We refer to ϵ and δ as the *RMax precision parameters*. This results in a threshold n_{known} on the number of visits $n(s,a)$ to a pair s,a that are necessary to reach this precision. Given the experience of a set of m MDPs $\bar{\mathcal{M}} = \{\bar{M}_1, \ldots, \bar{M}_m\}$, we define the total bound as the minimum over all the induced Lipschitz bounds.

Proposition 2. *Given a partially known task $M = \langle R, T \rangle$, the set of known state-action pairs K, and the set of Lipschitz bounds on Q^*_M induced by previous tasks $\left\{U_{\bar{M}_1}, \ldots, U_{\bar{M}_m}\right\}$, the function Q defined below is an upper bound on Q^*_M for all $s, a \in \mathcal{S} \times \mathcal{A}$.*

$$Q(s,a) \triangleq \begin{cases} R^a_s + \gamma \sum\limits_{s' \in \mathcal{S}} T^a_{ss'} \max\limits_{a' \in \mathcal{A}} Q(s', a') \\ \hfill \textit{if } (s,a) \in K, \\ U(s,a) \textit{ otherwise,} \end{cases} \qquad (6)$$

with $U(s,a) = \min\left\{\frac{1}{1-\gamma}, U_{\bar{M}_1}(s,a), \ldots, U_{\bar{M}_m}(s,a)\right\}$.

Commonly in RMax, Equation 6 is solved to a precision ϵ_Q via Value Iteration. This yields a function Q that is a valid heuristic (bound on Q^*_M) for the exploration of MDP M.

4.2 A Computable Upper Bound on Q^*_M

The key issue addressed in this section is how to actually compute $U(s,a)$, particularly when both source and target tasks are partially explored. Consider two tasks M and $\bar{M}$, on which vanilla RMax has been applied, yielding the respective sets of known state-action pairs K and $\bar{K}$, along with the learned models $\hat{M} = \langle \hat{T}, \hat{R} \rangle$ and $\hat{\bar{M}} = \langle \hat{\bar{T}}, \hat{\bar{R}} \rangle$, and the upper bounds Q and $\bar{Q}$ respectively on Q^*_M and $Q^*_{\bar{M}}$. Notice that, if $K = \emptyset$, then $Q(s,a) = \frac{1}{1-\gamma}$ for all s, a pairs. Conversely, if $K^c = \emptyset$, Q is an ϵ-accurate estimate of Q^*_M in $\mathcal{L}_1$-norm with high probability. Equation 6 allows the transfer of knowledge from $\bar{M}$ to M if $U_{\bar{M}}(s,a)$ can be computed. Unfortunately, the true model and optimal value functions, necessary to compute $U_{\bar{M}}$, are *partially* known (see Equation 5). Thus, we propose to compute a looser upper bound based on the learned models and value functions. First, we provide an upper bound $\hat{D}_{sa}(M \| \bar{M})$ on $D_{sa}(M \| \bar{M})$ (Definition 1).

Proposition 3. *Given two tasks M, $\bar{M}$ and respectively K, $\bar{K}$ the subsets of $\mathcal{S} \times \mathcal{A}$ where their models are known with accuracy ϵ in $\mathcal{L}_1$-norm with probability at least $1 - \delta$,*

$$\boldsymbol{Pr}\left(\hat{D}_{sa}(M \| \bar{M}) \geq D_{sa}(M \| \bar{M})\right) \geq 1 - \delta$$

with $\hat{D}_{sa}(M \| \bar{M})$ the upper bound on the pseudometric between models *defined below for $B = \epsilon\left(1 + \gamma \max_{s'} \bar{V}(s')\right)$.*

$$\hat{D}_{sa}(M \| \bar{M}) \triangleq \begin{cases} D^{\gamma \bar{V}}_{sa}(\hat{M}, \hat{\bar{M}}) + 2B & \textit{if } (s,a) \in K \cap \bar{K} \\ \max\limits_{\bar{\mu} \in \mathcal{M}} D^{\gamma \bar{V}}_{sa}(\hat{M}, \bar{\mu}) + B & \textit{if } (s,a) \in K \cap \bar{K}^c \\ \max\limits_{\mu \in \mathcal{M}} D^{\gamma \bar{V}}_{sa}(\mu, \hat{\bar{M}}) + B & \textit{if } (s,a) \in K^c \cap \bar{K} \\ \max\limits_{\mu, \bar{\mu} \in \mathcal{M}^2} D^{\gamma \bar{V}}_{sa}(\mu, \bar{\mu}) & \textit{if } (s,a) \in K^c \cap \bar{K}^c \end{cases} \tag{7}$$

Importantly, this upper bound $\hat{D}_{sa}(M \| \bar{M})$ can be calculated analytically (see Appendix, Section 7). This makes $\hat{D}_{sa}(M \| \bar{M})$ usable in the online Lifelong RL setting, where already explored tasks may be partially learned, and little knowledge has been gathered on the current task. The magnitude of the B term is controlled by ϵ. In the case where no information is available on the maximum value of $\bar{V}$, we have that $B = \frac{\epsilon}{1-\gamma}$. ϵ measures the accuracy with which the tasks are known: the smaller ϵ, the tighter the B bound. Note that $\bar{V}$ is used as an upper bound on the true $V^*_{\bar{M}}$. In many cases, $\max_{s'} V^*_{\bar{M}}(s') \leq \frac{1}{1-\gamma}$; *e.g.* for stochastic shortest path problems, which feature rewards only upon reaching terminal states, we have that $\max_{s'} V^*_{\bar{M}}(s') = 1$ and thus $B = (1 + \gamma)\epsilon$ is a tighter bound for transfer. Combining $\hat{D}_{sa}(M \| \bar{M})$ and Equation 3, one can derive an upper bound $\hat{d}_{sa}(M \| \bar{M})$ on $d_{sa}(M \| \bar{M})$, detailed in Proposition 4.

Proposition 4. *Given two tasks $M, \bar{M} \in \mathcal{M}$, K the set of state-action pairs where (R, T) is known with accuracy ϵ in $\mathcal{L}_1$-norm with probability at least $1 - \delta$. If $\gamma(1 + \epsilon) < 1$, the solution $\hat{d}_{sa}(M \| \bar{M})$ of the following fixed-point equation on $\hat{d}_{sa}$ (for all $s, a \in \mathcal{S} \times \mathcal{A}$) is an upper bound on $d_{sa}(M \| \bar{M})$ with probability at least $1 - \delta$:*

$$\hat{d}_{sa} = \hat{D}_{sa}(M \| \bar{M}) + \begin{cases} \gamma \left(\sum\limits_{s' \in \mathcal{S}} \hat{T}^a_{ss'} \max\limits_{a' \in \mathcal{A}} \hat{d}_{s'a'} + \epsilon \max\limits_{s', a' \in \mathcal{S} \times \mathcal{A}} \hat{d}_{s'a'} \right) \textit{ if } s, a \in K, \\ \gamma \max\limits_{s', a' \in \mathcal{S} \times \mathcal{A}} \hat{d}_{s'a'} \textit{ otherwise.} \end{cases} \tag{8}$$

Similarly as in Proposition 3, the condition $\gamma(1 + \epsilon) < 1$ illustrates the fact that for a large return horizon (large γ), a high accuracy (small ϵ) is needed for the bound to be computable. Eventually, a computable upper bound on Q^*_M given $\bar{M}$ with high probability is given by

$$\hat{U}_{\bar{M}}(s,a) = \bar{Q}(s,a) + \min\left\{\hat{d}_{sa}(M \| \bar{M}), \hat{d}_{sa}(\bar{M} \| M)\right\}. \tag{9}$$

The associated upper bound on $U(s,a)$ (Equation 6) given the set of previous tasks $\bar{\mathcal{M}} = \{\bar{M}_i\}_{i=1}^m$ is defined by

$$\hat{U}(s,a) = \min\left\{\tfrac{1}{1-\gamma}, \hat{U}_{\bar{M}_1}(s,a), \ldots, \hat{U}_{\bar{M}_m}(s,a)\right\}. \tag{10}$$

This upper bound can be used to transfer knowledge from a partially solved source task to a target task. If $\hat{U}(s,a) \leq \frac{1}{1-\gamma}$ on a subset of $\mathcal{S} \times \mathcal{A}$, then the convergence rate can be improved. As complete knowledge of both tasks is not needed to compute the upper bound, it can be applied online in the Lifelong RL setting. In the next section, we explicit an algorithm that leverages this value-transfer method.

4.3 Lipschitz RMax Algorithm

In Lifelong RL, MDPs are encountered sequentially. Applying RMax to task M yields the set of known state-action pairs K, the learned models $\hat{T}$ and $\hat{R}$, and the upper bound Q on Q^*_M. Saving this information when the task changes allows computing the upper bound of Equation 10 for the new target task, and using it to shrink the optimistic heuristic of RMax. This computation effectively transfers value functions between tasks based on task similarity. As the new task is explored online, the task similarity is progressively assessed with better confidence, refining the values of $\hat{D}_{sa}(M \| \bar{M})$, $\hat{d}_{sa}(M \| \bar{M})$ and eventually $\hat{U}$, allowing for more efficient transfer where the task similarity is appraised. The resulting algorithm, Lipschitz RMax (LRMax), is presented in Algorithm 1. To avoid ambiguities with $\bar{\mathcal{M}}$, we use $\hat{\mathcal{M}}$ to store learned features ($\hat{T}$, $\hat{R}$, K, Q) about previous MDPs. In a nutshell, the behavior of LRMax is precisely that of RMax, but with a tighter admissible heuristic $\hat{U}$ that becomes better as the new task is explored (while this heuristic remains constant in vanilla RMax). LRMax is PAC-MDP (Condition C1) as stated in Propositions 5 and 6 below. With $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, the sample complexity of vanilla RMax is $\tilde{\mathcal{O}}(S^2 A / (\epsilon^3 (1-\gamma)^3))$, which is improved by LRMax in Proposition 5 and meets Condition C2. Finally, $\hat{U}$ is a provable upper bound with high probability on Q^*_M, which avoids negative transfer and meets Condition C3.

Proposition 5 (Sample complexity (Strehl, Li, and Littman 2009))**.** *With probability $1 - \delta$, the greedy policy w.r.t. Q computed by LRMax achieves an ϵ-optimal return in MDP M after*

$$\tilde{\mathcal{O}}\left(\frac{S \left|\{s, a \in \mathcal{S} \times \mathcal{A} \mid \hat{U}(s,a) \geq V^*_M(s) - \epsilon\}\right|}{\epsilon^3 (1-\gamma)^3}\right)$$

samples (when logarithmic factors are ignored), with $\hat{U}$ defined in Equation 10 a non-static, decreasing quantity, upper bounded by $\frac{1}{1-\gamma}$.

Algorithm 1: Lipschitz RMax algorithm

```
Initialize $\hat{\mathcal{M}} = \emptyset$.
for each newly sampled MDP $M$ do
    Initialize $Q(s,a) = \frac{1}{1-\gamma}, \forall s, a$, and $K = \emptyset$
    Initialize $\hat{T}$ and $\hat{R}$ (RMax initialization)
    $Q \leftarrow$ UpdateQ$(\hat{\mathcal{M}}, \hat{T}, \hat{R})$
    for $t \in [1,$ max number of steps$]$ do
        $s =$ current state, $a = \arg\max_{a'} Q(s, a')$
        Observe reward $r$ and next state $s'$
        $n(s,a) \leftarrow n(s,a) + 1$
        if $n(s,a) < n_{known}$ then
            Store $(s, a, r, s')$
        if $n(s,a) = n_{known}$ then
            Update $K$ and $(\hat{T}^a_{ss'}, \hat{R}^a_s)$ (learned model)
            $Q \leftarrow$ UpdateQ$(\hat{\mathcal{M}}, \hat{T}, \hat{R})$
    Save $\hat{M} = \left(\hat{T}, \hat{R}, K, Q\right)$ in $\hat{\mathcal{M}}$
Function UpdateQ$(\hat{\mathcal{M}}, \hat{T}, \hat{R})$:
for $\bar{M} \in \hat{\mathcal{M}}$ do
    Compute $\hat{D}_{sa}(M \| \bar{M})$, $\hat{D}_{sa}(\bar{M} \| M)$ (Eq. 7)
    Compute $\hat{d}_{sa}(M \| \bar{M})$, $\hat{d}_{sa}(\bar{M} \| M)$ (DP on Eq. 8)
    Compute $\hat{U}_{\bar{M}}$ (Eq. 9)
Compute $\hat{U}$ (Eq. 10)
Compute and return $Q$ (DP on Eq. 6 using $\hat{U}$)
```

Proposition 5 shows that the sample complexity of LRMax is no worse than that of RMax. Consequently, in the worst case, LRMax performs as badly as learning from scratch, which is to say that the transfer method is not negative as it cannot degrade the performance.

Proposition 6 (Computational complexity). *The total computational complexity of LRMax (Algorithm 1) is*

$$\tilde{\mathcal{O}} \left(\tau + \frac{S^3 A^2 N}{(1-\gamma)} \ln \left(\frac{1}{\epsilon_Q (1-\gamma)} \right) \right)$$

with τ the number of interaction steps, ϵ_Q the precision of value iteration and N the number of source tasks.

4.4 Refining the LRMax Bounds

LRMax relies on bounds on the local MDP dissimilarity (Equation 8). The quality of the Lipschitz bound on $Q^*_{\bar{M}}$ can be improved according to the quality of those estimates. We discuss two methods to provide finer estimates.

Refining with prior knowledge. First, from the definition of $D_{sa}(M \| \bar{M})$, it is easy to show that this pseudometric between models is always upper bounded by $\frac{1+\gamma}{1-\gamma}$. However, in practice, the tasks experienced in a Lifelong RL experiment might not cover the full span of possible MDPs $\mathcal{M}$ and may systematically be closer to each other than $\frac{1+\gamma}{1-\gamma}$. For instance, the distance between two games in the Arcade Learning Environment (ALE) (Bellemare et al. 2013), is smaller than the maximum distance between any two MDPs defined on the common state-action space of the ALE (extended discussion in Appendix, Section 11). Let us note $\tilde{\mathcal{M}} \subset \mathcal{M}$ the set of possible MDPs for a particular Lifelong RL experiment. Let $D_{\max}(s,a) \triangleq \max_{M, \bar{M} \in \tilde{\mathcal{M}}^2} \left(D_{sa}(M \| \bar{M}) \right)$ be the *maximum model pseudo-distance* at a particular s, a pair on the subset $\tilde{\mathcal{M}}$. *Prior knowledge* might indicate a smaller upper bound for $D_{\max}(s,a)$ than $\frac{1+\gamma}{1-\gamma}$. We will note such an upper bound $D_{\max}$, considered valid for all s, a pairs, *i.e.*, such that $D_{\max} \geq \max_{s,a,M,\bar{M} \in \mathcal{S} \times \mathcal{A} \times \tilde{\mathcal{M}}^2} \left(D_{sa}(M \| \bar{M}) \right)$. In a Lifelong RL experiment, $D_{\max}$ can be seen as a rough estimate of the maximum model discrepancy an agent may encounter. Figure 2 illustrates the relative importance of $D_{\max}$ *vs.* $\frac{1+\gamma}{1-\gamma}$. Solving Equation 8 boils down to accumulating $\hat{D}_{sa}(M \| \bar{M})$ values in $\hat{d}_{sa}(M \| \bar{M})$. Hence, reducing a $\hat{D}_{sa}(M \| \bar{M})$ estimate in a single s, a pair actually reduces $\hat{d}_{sa}(M \| \bar{M})$ in *all* s, a pairs. Thus, replacing $\hat{D}_{sa}(M \| \bar{M})$ in Equation 8 by $\min\{D_{\max}, \hat{D}_{sa}(M \| \bar{M})\}$, provides a smaller upper bound $\hat{d}_{sa}(M \| \bar{M})$ on $d_{sa}(M \| \bar{M})$, and thus a smaller $\hat{U}$ which allows transfer if it is less than $\frac{1}{1-\gamma}$. Consequently, the knowledge of such a bound $D_{\max}$ can make a difference between successful and unsuccessful transfer, even if its value is of little importance. Conversely, setting a value for $D_{\max}$ quantifies the distance between MDPs where transfer is efficient.

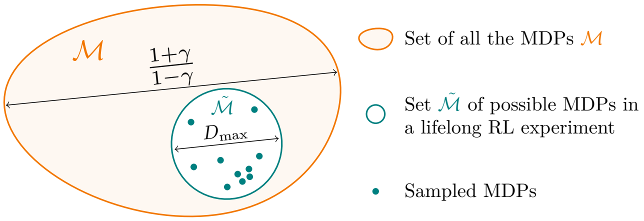

$\mathcal{M}$
$\frac{1+\gamma}{1-\gamma}$
$\tilde{\mathcal{M}}$
$D_{\max}$
Set of all the MDPs $\mathcal{M}$
Set $\tilde{\mathcal{M}}$ of possible MDPs in a lifelong RL experiment
Sampled MDPs

Figure 2: Illustration of the prior knowledge on the maximum pseudo-distance between models for a particular s, a pair.

Refining by learning the maximum distance. The value of $D_{\max}(s,a)$ can be estimated online for each s, a pair, discarding the hypothesis of available prior knowledge. We propose to use an empirical estimate of the maximum model distance at s, a: $\hat{D}_{\max}(s,a) \triangleq \max_{M, \bar{M} \in \hat{\mathcal{M}}^2} \{\hat{D}_{sa}(M \| \bar{M})\}$, with $\hat{\mathcal{M}}$ the set of explored tasks. The pitfall of this approach is that, with few explored tasks, $\hat{D}_{\max}(s,a)$ could underestimate $D_{\max}(s,a)$. Proposition 7 provides a lower bound on the probability that $\hat{D}_{\max}(s,a) + \epsilon$ does not underestimate $D_{\max}(s,a)$, depending on the number of sampled tasks.

Proposition 7. *Consider an algorithm producing ϵ-accurate model estimates $\hat{D}_{sa}(M \| \bar{M})$ for a subset K of $\mathcal{S} \times \mathcal{A}$ after interacting with any two MDPs $M, \bar{M} \in \mathcal{M}$. Assume $\hat{D}_{sa}(M \| \bar{M}) \geq D_{sa}(M \| \bar{M})$ for any $s, a \notin K$. For all $s, a \in \mathcal{S} \times \mathcal{A}$, $\delta \in (0, 1]$, after sampling m tasks, if m is large enough to verify $2(1 - p_{\min})^m - (1 - 2p_{\min})^m \leq \delta$,*

$$\boldsymbol{Pr} \left(\hat{D}_{\max}(s,a) + \epsilon \geq D_{\max}(s,a) \right) \geq 1 - \delta \, .$$

This result indicates when $\hat{D}_{\max}(s,a) + \epsilon$ upper bounds $D_{\max}(s,a)$ with high probability. In such a case, $\hat{D}_{sa}(M \| \bar{M})$ of Equation 8 can be replaced by $\min\{\hat{D}_{\max}(s,a) + \epsilon, \hat{D}_{sa}(M \| \bar{M})\}$ to tighten the bound on $d_{sa}(M \| \bar{M})$. Assuming a lower bound $p_{\min}$ on the sampling probability of a task implies that $\mathcal{M}$ is finite and is seen as a non-adversarial task sampling rule (Abel et al. 2018).

5 Experiments

The experiments reported here[1] illustrate how the Lipschitz bound (Equation 9) provides a tighter upper bound on Q^*, improving the sample complexity of LRMax compared to RMax, and making the transfer of inter-task knowledge effective. Graphs are displayed with 95% confidence intervals. For information in line with the Machine Learning Reproducibility Check-list (Pineau 2019) see the Appendix, Section 16.

We evaluate different variants of LRMax in a Lifelong RL experiment. The RMax algorithm will be used as a no-transfer baseline. LRMax(x) denotes Algorithm 1 with prior $D_{\max} = x$. MaxQInit denotes the MAXQINIT algorithm from Abel et al. (2018), consisting in a state-of-the art PAC-MDP algorithm. Both LRMax and MaxQInit algorithms achieve value transfer by providing a tighter upper bound on Q^* than $\frac{1}{1-\gamma}$. Computing both upper bounds and taking the minimum results in combining the two approaches. We include such a combination in our study with the LRMaxQInit algorithm. Similarly, the latter algorithm benefiting from prior knowledge $D_{\max} = x$ is denoted by LRMaxQInit(x). For the sake of comparison, we only compare algorithms with the same features, namely, tabular, online, PAC-MDP methods, presenting non-negative transfer.

The environment used in all experiments is a variant of the "tight" task used by Abel et al. (2018). It is an 11×11 grid-world, the initial state is in the centre, actions are the cardinal moves (Appendix, Section 12). The reward is always zero except for the three goal cells in the upper-right corner. Each sampled task has its own reward values, drawn from $[0.8, 1]$ for each of the three goal cells and its own probability of slipping (performing a different action than the one selected), picked in $[0, 0.1]$. Hence, tasks have different reward and transition functions. Notice the distinction in applicability between MaxQInit, that requires the set of MDPs to be finite, and LRMax, that can be used with any set of MDPs. For the comparison between both to be possible, we drew tasks from a finite set of 5 MDPs. We sample 15 tasks sequentially among this set, each run for 2000 episodes of length 10. The operation is repeated 10 times to narrow the confidence intervals. We set $n_{known} = 10$, $\delta = 0.05$, and $\epsilon = 0.01$ (discussion in Appendix, Section 15). Other Lifelong RL experiments are reported in Appendix, Section 13.

The results are reported in Figure 3. Figure 3a displays the discounted return for each task, averaged across episodes. Similarly, Figure 3b displays the discounted return for each episode, averaged across tasks (same color code as Figure 3a). Figure 3c displays the discounted return for five specific instances, detailed below. To avoid inter-task disparities, all the aforementioned discounted returns are displayed relative to an estimator of the optimal expected return for each task. For readability, Figures 3b and 3c display a moving average over 100 episodes. Figure 3d reports the benefits of various values of $D_{\max}$ on the algorithmic properties.

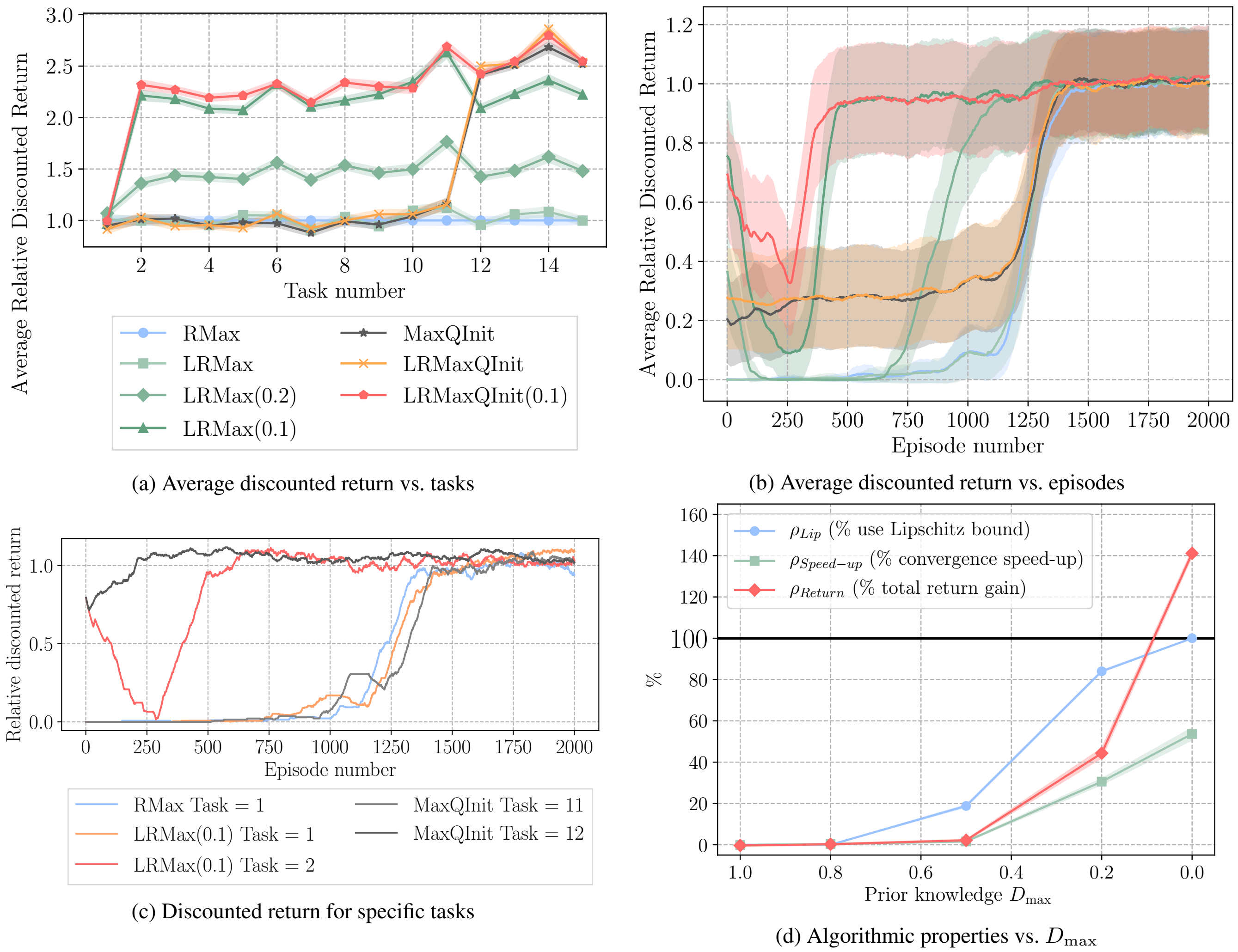
In Figure 3a, we first observe that LRMax benefits from the transfer method, as the average discounted return increases as more tasks are experienced. Moreover, this advantage appears as early as the second task. In contrast, MaxQInit requires to wait for task 12 before benefiting from transfer. As suggested in Section 4.4, increasing amounts of prior knowledge allow the LRMax transfer method to be more efficient: a smaller known upper bound $D_{\max}$ on $\hat{D}_{sa}(M \| \bar{M})$ accelerates convergence. Combining both approaches in the LRMaxQInit algorithm outperforms all other methods. Episode-wise, we observe in Figure 3b that the LRMax transfer method allows for faster convergence, *i.e.*, lower sample complexity. Interestingly, LRMax exhibits three stages in the learning process. 1) The first episodes are characterized by a direct exploitation of the transferred knowledge, causing these episodes to yield high payoff. This behavior is a consequence of the combined facts that the Lipschitz bound (Equation 9) is larger on promising regions of $\mathcal{S} \times \mathcal{A}$ seen on previous tasks and the fact that LRMax acts greedily w.r.t. that bound. 2) This high performance regime is followed by the exploration of unknown regions of $\mathcal{S} \times \mathcal{A}$, in our case yielding low returns. Indeed, as promising regions are explored first, the bound becomes tighter for the corresponding state-action pairs, enough for the Lipschitz bound of unknown pairs to become larger, thus driving the exploration towards low payoff regions. Such regions are then identified and never revisited. 3) Eventually, LRMax stops exploring and converges to the optimal policy. Importantly, in all experiments, LRMax never experiences negative transfer, as supported by the provability of the Lipschitz upper bound with high probability. LRMax is at least as efficient as the no-transfer RMax baseline.

Figure 3c displays the collected returns of RMax, LRMax(0.1), and MaxQInit for specific tasks. We observe that LRMax benefits from transfer as early as Task 2, where the previous 3-stage behavior is visible. MaxQInit takes until task 12 to leverage the transfer method. However, the bound it provides is tight enough that it does not have to explore.

In Figure 3d, we display the following quantities for various values of $D_{\max}$: ρ_{Lip}, the fraction of the time the Lipschitz bound was tighter than the RMax bound $\frac{1}{1-\gamma}$; $\rho_{Speed\text{-}up}$, is the relative gain of time steps before convergence when comparing LRMax to RMax. This quantity is estimated based on the last updates of the empirical model $\bar{M}$; ρ_{Return}, is the relative total return gain on 2000 episodes of LRMax w.r.t. RMax. First, we observe an increase of ρ_{Lip} as $D_{\max}$ becomes tighter. This means that the Lipschitz bound of Equation 9 becomes effectively smaller than $\frac{1}{1-\gamma}$. This phenomenon leads to faster convergence, indicated by $\rho_{Speed\text{-}up}$. Eventually, this increased convergence rate allows for a net total return gain, as can be seen with the increase of ρ_{Return}.

Overall, in this analysis, we have showed that LRMax benefits from an enhanced sample complexity thanks to the value-transfer method. The knowledge of a prior $D_{\max}$ increases

[1] Code available at https://github.com/SuReLI/llrl



(a) Average discounted return vs. tasks

(b) Average discounted return vs. episodes

(c) Discounted return for specific tasks

(d) Algorithmic properties vs. $D_{\max}$

Figure 3: Experimental results. LRMax benefits from an enhanced sample complexity thanks to the value-transfer method.

this benefit. The method is comparable to the MaxQInit method and has some advantages such as the early fitness for use and the applicability to infinite sets of tasks. Moreover, the transfer is non-negative while preserving the PAC-MDP guarantees of the algorithm. Additionally, we show in Appendix, Section 14 that, when provided with any prior $D_{\max}$, LRMax increasingly stops using it during exploration, confirming the claim of Section 4.4 that providing $D_{\max}$ enables transfer even if its value is of little importance.

6 Conclusion

We have studied theoretically the Lipschitz continuity property of the optimal Q-function in the MDP space w.r.t. a new metric. We proved a local Lipschitz continuity result, establishing that the optimal Q-functions of two close MDPs are themselves close to each other. We then proposed a value-transfer method using this continuity property with the Lipschitz RMax algorithm, practically implementing this approach in the Lifelong RL setting. The algorithm preserves PAC-MDP guarantees, accelerates learning in subsequent tasks and exhibits no negative transfer. Improvements of the algorithm were discussed in the form of prior knowledge on the maximum distance between models and online estimation of this distance. As a non-negative, similarity-based, PAC-MDP transfer method, the LRMax algorithm is the first method of the literature combining those three appealing features. We showcased the algorithm in Lifelong RL experiments and demonstrated empirically its ability to accelerate learning while not experiencing any negative transfer. Notably, our approach can directly extend other PAC-MDP algorithms (Szita and Szepesvári 2010; Rao and Whiteson 2012; Pazis, Parr, and How 2016; Dann, Lattimore, and Brunskill 2017) to the Lifelong setting. In hindsight, we believe this contribution provides a sound basis to non-negative value transfer via MDP similarity, a study that was lacking in the literature. Key insights for the practitioner lie both in the theoretical analysis and in the practical derivation of a transfer scheme achieving non-negative transfer with PAC guarantees. Further, designing scalable methods conveying the same intuition could be a promising research direction.

Acknowledgments

We would like to thank Dennis Wilson for fruitful discussions and comments on the paper. This research was supported by the Occitanie region, France; ISAE-SUPAERO (Institut Supérieur de l'Aéronautique et de l'Espace); fondation ISAE-SUPAERO; École Doctorale Systèmes; and ONERA, the French Aerospace Lab.

References

Abel, D.; Jinnai, Y.; Guo, S. Y.; Konidaris, G.; and Littman, M. L. 2018. Policy and Value Transfer in Lifelong Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 20–29.

Ammar, H. B.; Eaton, E.; Taylor, M. E.; Mocanu, D. C.; Driessens, K.; Weiss, G.; and Tuyls, K. 2014. An Automated Measure of MDP Similarity for Transfer in Reinforcement Learning. In *Workshops at the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*.

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

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research* 47: 253–279.

Bellman, R. 1957. *Dynamic Programming*. Princeton, USA: Princeton University Press.

Brafman, R. I.; and Tennenholtz, M. 2002. R-max - a General Polynomial Time Algorithm for Near-Optimal Reinforcement Learning. *Journal of Machine Learning Research* 3(Oct): 213–231.

Brunskill, E.; and Li, L. 2013. Sample Complexity of Multitask Reinforcement Learning. In *Proceedings of the 29th conference on Uncertainty in Artificial Intelligence (UAI 2013)*.

Brunskill, E.; and Li, L. 2014. PAC-inspired Option Discovery in Lifelong Reinforcement Learning. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, 316–324.

Carroll, J. L.; and Seppi, K. 2005. Task Similarity Measures for Transfer in Reinforcement Learning Task Libraries. In *Proceedings of the 5th International Joint Conference on Neural Networks (IJCNN 2005)*, volume 2, 803–808. IEEE.

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 5713–5723.

Ferns, N.; Panangaden, P.; and Precup, D. 2004. Metrics for Finite Markov Decision Processes. In *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence (UAI 2004)*, 162–169. AUAI Press.

Lazaric, A. 2012. Transfer in Reinforcement Learning: a Framework and a Survey. In *Reinforcement Learning*, 143–173. Springer.

Lazaric, A.; Restelli, M.; and Bonarini, A. 2008. Transfer of Samples in Batch Reinforcement Learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, 544–551.

Mahmud, M. M.; Hawasly, M.; Rosman, B.; and Ramamoorthy, S. 2013. Clustering Markov Decision Processes for Continual Transfer. *Computing Research Repository (arXiv/CoRR)* URL https://arxiv.org/abs/1311.3959.

Pazis, J.; Parr, R. E.; and How, J. P. 2016. Improving PAC Exploration using the Median of Means. In *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*, 3898–3906.

Pineau, J. 2019. Machine Learning Reproducibility Checklist. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf. Version 1.2, March 27, 2019, last accessed on August 27, 2020.

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

Puterman, M. L. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Rachelson, E.; and Lagoudakis, M. G. 2010. On the Locality of Action Domination in Sequential Decision Making. In *Proceedings of the 11th International Symposium on Artificial Intelligence and Mathematics (ISAIM 2010)*.

Rao, K.; and Whiteson, S. 2012. V-MAX: Tempered Optimism for Better PAC Reinforcement Learning. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, 375–382.

Silver, D. L.; Yang, Q.; and Li, L. 2013. Lifelong Machine Learning Systems: Beyond Learning Algorithms. In *AAAI Spring Symposium: Lifelong Machine Learning*, volume 13, 05.

Song, J.; Gao, Y.; Wang, H.; and An, B. 2016. Measuring the Distance Between Finite Markov Decision Processes. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*, 468–476.

Sorg, J.; and Singh, S. 2009. Transfer via Soft Homomorphisms. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, 741–748. International Foundation for Autonomous Agents and Multiagent Systems.

Strehl, A. L.; Li, L.; and Littman, M. L. 2009. Reinforcement Learning in Finite MDPs: PAC Analysis. *Journal of Machine Learning Research* 10(Nov): 2413–2444.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT press, Cambridge.

Szita, I.; and Szepesvári, C. 2010. Model-Based Reinforcement Learning with Nearly Tight Exploration Complexity Bounds. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, 1031–1038.

Taylor, M. E.; and Stone, P. 2009. Transfer Learning for Reinforcement Learning Domains: A Survey. *Journal of Machine Learning Research* 10(Jul): 1633–1685.

Lipschitz Lifelong Reinforcement Learning
Appendix

Erwan Lecarpentier[1, 2], **David Abel**[3], **Kavosh Asadi**[3, 4*], **Yuu Jinnai**[3],
Emmanuel Rachelson[1], **Michael L. Littman**[3]
[1]ISAE-SUPAERO, Université de Toulouse, France
[2]ONERA, The French Aerospace Lab, Toulouse, France
[3]Brown University, Providence, Rhode Island, USA
[4]Amazon Web Service, Palo Alto, California, USA
erwanlecarpentier@mailbox.org

1 Negative transfer

In the lifelong RL setting, it is reasonable to think that knowledge gained on previous MDPs could be re-used to improve the performance in new MDPs. Such a practice, known as knowledge transfer, sometimes does cause the opposite effect, *i.e.*, decreases the performance. In such a case, we talk about *negative transfer*. Several attempt to formally define negative transfer have been done, but researchers hardly agree on a single definition, as *performance* can be defined in various ways. For instance, it can be characterized by the speed of convergence, the area under the learning curve, the final score of the learned policy or classifier, and many other things. Defining negative transfer is out of the scope of this paper, but let us give an example of why this phenomenon can be problematic.

In their paper, Song et al. (2016) propose a transfer methods based on the metric between MDPs they introduce, stemming from the bi-simulation metric introduced by Ferns, Panangaden, and Precup (2004). In their method, a bi-simulation metric is computed between each pair of states belonging respectively to the source and target MDPs. Roughly, this metric tells how *different* are the transition and reward models corresponding to the states pairs, for the action maximizing th distance. More precisely, if we note (T, R) and $(\bar{T}, \bar{R})$ the models of two MDPs, and $(s, s') \in \mathcal{S}$ a state pair, the distance d between s and s' is defined by

$$d(s, s') = \max_{a \in \mathcal{A}} \left(\left| R_s^a - \bar{R}_{s'}^a \right| + c \, W_1 \left(T_{s \cdot}^a, \bar{T}_{s' \cdot}^a \right) \right) , \tag{11}$$

where $c \in \mathbb{R}$ is a positive constant and W_1 is the 1-Wasserstein metric. For each state of the target model, the closest counterpart state (with the smallest bi-simulation distance) of the source MDP is identified and its learned Q-values are used to initialize the Q-function of the target MDP. In their experiments, Song et al. (2016) run a standard Q-Learning algorithm (Watkins and Dayan 1992) with an ϵ-greedy exploration strategy thereafter.

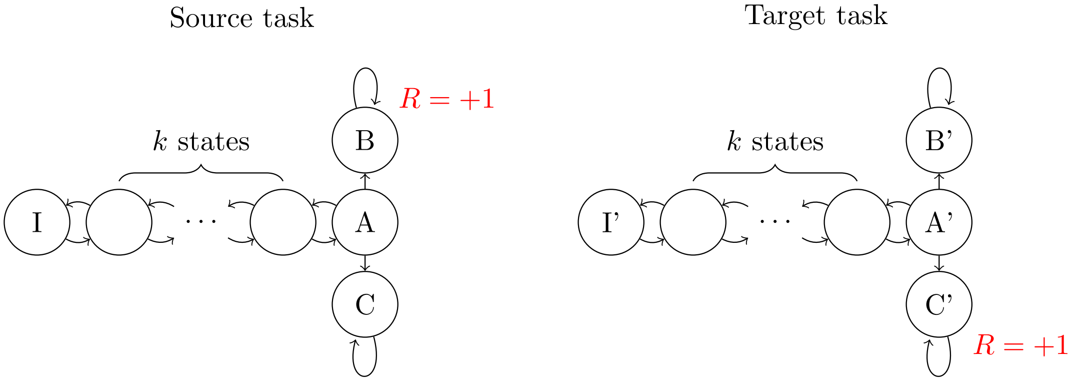


Figure 1: The T-shaped MDP transfer task.

Let us now consider applying this method to a similar task to the T-shaped MDP transfer task proposed by Taylor and Stone (2009). The source and target tasks are respectively described on the left and right sides of Figure 1. In each task, the states are

[*] Kavosh Asadi finished working on this project before joining Amazon.

represented by the circles and the arrows between them correspond to the available actions that allow to move from one state to the other. The initial state of both tasks is the left state I for the source task and I′ for the target task. Between the states I and A in the source task (respectively I′ to A′ in the target task) are k states, k being a parameter increasing the distance to travel from I to A (respectively I′ to A′). The tasks are deterministic and the reward is zero everywhere except for the state B in the source task and C′ in the target task where a reward of $+1$ is received. Consequently, the optimal policy in the source task is to go to the state A and then to the state B. In the target task, the same applies except that a transition to state C should be applied in place of state B′ when the agent is in state A′.

Regardless of the parameters used in the bi-simulation metric of Equation 11, the direct state transfer method from Song et al. (2016) maps the following states together as they share the exact same model:

$$\begin{aligned} \text{I} &\longleftrightarrow \text{I}' \\ \text{k states} &\longleftrightarrow \text{k states} \\ \text{A} &\longleftrightarrow \text{A}'\,. \end{aligned}$$

Hence, during learning, the Q-function of the target task is initialized with the values of the Q-function of the source task. Therefore, the behavior derived with the Q-Learning algorithm is the optimal policy of the source task, but in the target task. Depending on the value of the learning rate of the algorithm, the time to favor action DOWN in state A′ instead of action UP can be arbitrarily long. Also, depending on the value of ϵ, the exploration of state C′ due to the ϵ-greedy strategy can be arbitrarily unlikely. Finally, the time needed for one of those two events to occur increases proportionally to the value of k, which can be arbitrarily large.

This case illustrates the difficulty facing any transfer method in the general context of lifelong RL. The method proposed by Song et al. (2016) can be highly efficient in some cases as they show in experiments, but the lack of theoretical guarantees makes negative transfer possible. Generally, using a similarity measure such that the bi-simulation metric helps to discourage using some source tasks when the computed similarity is too low. However, as we saw in the T-shaped MDP example, this rule is not absolute and the choice of the metric is important. The approach we develop in this paper aims at avoiding negative transfer by providing a conservative transferred knowledge that is simply of no use when the similarity between source and target tasks is too low. This is intuitive as we do not expect *any* task to provide transferable knowledge to *any* other task.

2 Discussion on metrics and related notions

A *metric* on a set X is a function $m : X \times X \to \mathbb{R}$ which has the following properties for any $x, y, z \in X$:

P1. $m(x, y) \geq 0$ (positivity),

P2. $m(x, y) = 0 \Leftrightarrow x = y$ (positive definiteness),

P3. $m(x, y) = m(y, x)$ (symmetry),

P4. $m(x, z) \leq m(x, y) + m(y, z)$ (triangle inequality).

If property P2 is not verified by m, but instead we have for any $x \in X$ that $m(x, x) = 0$, then m is called a *pseudo-metric*. If m only verifies P1, P2 and P4 then m is called a *quasi-metric*. If m only verifies P1 and P2 and if X is a set of probability measures, then m is called a *divergence*.

From this, the pseudo-metric between models of Definition 1 is indeed a pseudo-metric as it is relative to a positive function f that could be equal to zero and break property P2.

The local MDP dissimilarity between MDPs $d_{sa}(M\|\bar{M})$ of Proposition 1 does not respect properties P2 and P3, hence the name *dissimilarity*. The $\Delta_{sa}(M, \bar{M}) = \min\left\{d_{sa}(M\|\bar{M}), d_{sa}(\bar{M}\|M)\right\}$ quantity, however, regains property P3 and is hence a pseudo-metric. A noticeable consequence is that Proposition 1 is "in the spirit" of a Lipschitz continuity result but cannot be called as such, hence the name *pseudo-Lipschitz continuity*.

The same goes for the global dissimilarity $d(M\|\bar{M}) = \frac{1}{1-\gamma} \max_{s,a \in \mathcal{S}\times\mathcal{A}} \left(D_{sa}(M\|\bar{M})\right)$. However, using $\min\left\{d_M^{\bar{M}}, d_{\bar{M}}^M\right\}$ allows to regain property 3 and makes this quantity a pseudo-metric again between MDPs.

3 Proof of Proposition 1

Notation 1. *Given two sets X and Y, we note $\mathcal{F}(X, Y)$ the set of functions defined on the domain X with codomain Y.*

Lemma 1. *Given two MDPs $M, \bar{M} \in \mathcal{M}$, the following equation on $d \in \mathcal{F}(\mathcal{S}\times\mathcal{A}, \mathbb{R})$ is a fixed-point equation admitting a unique solution for any $(s, a) \in \mathcal{S}\times\mathcal{A}$:*

$$d_{sa} = D_{sa}(M\|\bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \max_{a' \in \mathcal{A}} d_{s'a'}\,.$$

We refer to this unique solution as $d_{sa}(M\|\bar{M})$.

Proof of Lemma 1. The proof follows closely that in (Puterman 2014) that proves that the Bellman operator over value functions is a contraction mapping. Let L be the functional operator that maps any function $d \in \mathcal{F}\left(\mathcal{S} \times \mathcal{A}, \mathbb{R}\right)$ to

$$\begin{array}{rlll} Ld: & \mathcal{S} \times \mathcal{A} & \rightarrow & \mathbb{R} \\ & s, a & \mapsto & D_{sa}(M \| \bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} d_{s'a'} \,. \end{array}$$

Then for f and g, two functions from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have that

$$\begin{aligned} Lf_{sa} - Lg_{sa} &= \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \left(\max_{a' \in \mathcal{A}} f_{s'a'} - \max_{a' \in \mathcal{A}} g_{s'a'} \right) \\ &\leq \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} \left(f_{s'a'} - g_{s'a'} \right) \\ &\leq \gamma \left\| f - g \right\|_\infty \,. \end{aligned}$$

Since this is true for any pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have that

$$\left\| Lf - Lg \right\|_\infty \leq \gamma \left\| f - g \right\|_\infty \,.$$

Since $\gamma < 1$, L is a contraction mapping in the metric space $\left(\mathcal{F}\left(\mathcal{S} \times \mathcal{A}, \mathbb{R}\right), \left\| \cdot \right\|_\infty\right)$. This metric space being complete and non-empty, it follows by direct application of the Banach fixed-point theorem that the equation $d = Ld$ admits a unique solution. □

Proof of Proposition 1. The proof is by induction. The value iteration sequence of iterates $\left(Q^n_M\right)_{n \in \mathbb{N}}$ of the optimal Q-function of any MDP $M \in \mathcal{M}$ is defined for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ by:

$$\begin{aligned} Q^0_M(s, a) &= 0 \,, \\ Q^{n+1}_M(s, a) &= R^a_s + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s', a'), \; \forall n \in \mathbb{N} \,. \end{aligned}$$

Consider two MDPs $M, \bar{M} \in \mathcal{M}$. It is obvious that $\left| Q^0_M(s, a) - Q^0_{\bar{M}}(s, a) \right| \leq d_{sa}(M \| \bar{M})$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Suppose the property $\left| Q^n_M(s, a) - Q^n_{\bar{M}}(s, a) \right| \leq d_{sa}(M \| \bar{M})$ true at rank $n \in \mathbb{N}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Consider now the rank $n + 1$ and a pair $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned} \left| Q^{n+1}_M(s, a) - Q^{n+1}_{\bar{M}}(s, a) \right| &= \left| R^a_s - \bar{R}^a_s + \gamma \sum_{s' \in \mathcal{S}} \left[T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s', a') - \bar{T}^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_{\bar{M}}(s', a') \right] \right| \\ &\leq \left| R^a_s - \bar{R}^a_s \right| + \gamma \sum_{s' \in \mathcal{S}} \left| T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s', a') - \bar{T}^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_{\bar{M}}(s', a') \right| \\ &\leq \left| R^a_s - \bar{R}^a_s \right| + \gamma \sum_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}} Q^n_{\bar{M}}(s', a') \left| T^a_{ss'} - \bar{T}^a_{ss'} \right| \\ &\quad + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \left| \max_{a' \in \mathcal{A}} Q^n_M(s', a') - \max_{a' \in \mathcal{A}} Q^n_{\bar{M}}(s', a') \right| \\ &\leq \left| R^a_s - \bar{R}^a_s \right| + \sum_{s' \in \mathcal{S}} \gamma V^*_{\bar{M}}(s') \left| T^a_{ss'} - \bar{T}^a_{ss'} \right| + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} \left| Q^n_M(s', a') - Q^n_{\bar{M}}(s', a') \right| \\ &\leq D_{sa}(M \| \bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a'} d_{s'a'}(M \| \bar{M}) \\ &\leq d_{sa}(M \| \bar{M}) \,, \end{aligned}$$

where we used Lemma 1 in the last inequality. Since Q^*_M and $Q^*_{\bar{M}}$ are respectively the limits of the sequences $\left(Q^n_M\right)_{n \in \mathbb{N}}$ and $\left(Q^n_{\bar{M}}\right)_{n \in \mathbb{N}}$, it results from passage to the limit that

$$\left| Q^*_M(s, a) - Q^*_{\bar{M}}(s, a) \right| \leq d_{sa}(M \| \bar{M}) \,.$$

By symmetry, we also have $\left| Q^*_M(s, a) - Q^*_{\bar{M}}(s, a) \right| \leq d_{sa}(M \| \bar{M})$ and we can take the minimum of the two valid upper bounds, yielding:

$$\left| Q^*_M(s, a) - Q^*_{\bar{M}}(s, a) \right| \leq \min \left\{ d_{sa}(M \| \bar{M}), d_{sa}(\bar{M} \| M) \right\} \,,$$

which concludes the proof. □

4 Similar results to Proposition 1

Similar results to Proposition 1 can be derived. First, an important consequence is the global pseudo-Lipschitz continuity result presented below.

Proposition 8 (Global pseudo-Lipschitz continuity). *For two MDPs M, $\bar{M}$, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,*

$$|Q^*_M(s,a) - Q^*_{\bar{M}}(s,a)| \leq \Delta(M, \bar{M}), \tag{12}$$

with $\Delta(M, \bar{M}) \triangleq \min\left\{d(M\|\bar{M}), d(\bar{M}\|M)\right\}$ *and*

$$d(M\|\bar{M}) \triangleq \frac{1}{1-\gamma} \max_{s,a \in \mathcal{S}\times\mathcal{A}} \left(D_{sa}(M\|\bar{M})\right).$$

From a pure transfer perspective, Equation 12 is interesting since the right hand side does not depend on s, a. Hence, the counterpart of the upper bound of Equation 4, namely,

$$s, a \mapsto Q^*_{\bar{M}}(s,a) + \Delta(M, \bar{M}),$$

is easier to compute. Indeed, $\Delta(M, \bar{M})$ can be computed once and for all, contrarily to $\Delta_{sa}(M, \bar{M})$ that needs to be evaluated for all s, a pair. However, we do not use this result for transfer because it is impractical to compute online. Indeed, estimating the maximum in the definition of $d(M\|\bar{M})$ can be as hard as solving both MDPs, which, when it happens, is too late for transfer to be useful.

Proof of Proposition 8. The proof is by induction. We consider the sequence of value iteration iterates defined for any MDP $M \in \mathcal{M}$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$ by

$$\begin{aligned} Q^0_M(s,a) &= 0\,, \\ Q^{n+1}_M(s,a) &= R^a_s + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s', a'),\ \forall n \in \mathbb{N}\,. \end{aligned}$$

Consider two MDPs $M, \bar{M} \in \mathcal{M}$. It is immediate for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ that

$$\left|Q^0_M(s,a) - Q^0_{\bar{M}}(s,a)\right| \leq d(M\|\bar{M})\,,$$

and, by symmetry, the result holds as well for $d(\bar{M}\|M)$. Suppose that it is true at rank $n \in \mathbb{N}$. Consider rank $n+1$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have that:

$$\begin{aligned} \left|Q^{n+1}_M(s,a) - Q^{n+1}_{\bar{M}}(s,a)\right| &\leq D_{sa}(M\|\bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} \left|Q^n_M(s',a') - Q^n_{\bar{M}}(s',a')\right| \\ &\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} D_{sa}(M\|\bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \frac{1}{1-\gamma} \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} D_{sa}(M\|\bar{M}) \\ &\leq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} D_{sa}(M\|\bar{M}) \left(1 + \frac{\gamma}{1-\gamma}\right) \\ &\leq d(M\|\bar{M})\,. \end{aligned}$$

By symmetry, the results holds as well for $d(\bar{M}\|M)$ which concludes the proof by induction. □

The second result is for the value function and is stated below.

Proposition 9 (Local pseudo-Lipschitz continuity of the optimal value function). *For any two MDPs $M, \bar{M} \in \mathcal{M}$, for all $s \in \mathcal{S}$,*

$$\left|V^*_M(s) - V^*_{\bar{M}}(s)\right| \leq \max_{a \in \mathcal{A}} \Delta_{sa}(M, \bar{M})$$

where the local MDP pseudo-metric $\Delta_{sa}(M, \bar{M})$ has the same definition as in Proposition 1.

Proof of Proposition 9. The proof follows exactly the same steps as the proof of Proposition 1, *i.e.*, by first constructing the value iteration sequence of iterates of the optimal value function, showing the result by induction for rank $n \in \mathbb{N}$ and then concluding with a passage to the limit. □

Another result can be derived for the value of any policy π. For the sake of generality, we state the result for any stochastic policy mapping states to distributions over actions. Note that a deterministic policy is a stochastic policy mapping states to Dirac distributions over actions. First, we state the result for the value function in Proposition 10 and then for the Q function in Proposition 11.

Proposition 10 (Local pseudo-Lipschitz continuity of the value function of any policy)**.** *For any two MDPs $M, \bar{M} \in \mathcal{M}$, for any stochastic stationary policy π, for all $s \in \mathcal{S}$,*

$$\left|V^\pi_M(s) - V^\pi_{\bar{M}}(s)\right| \leq \Delta^\pi_s(M, \bar{M})$$

where $\Delta^\pi_s(M, \bar{M}) \triangleq \min\left\{d^\pi_s(M\|\bar{M}), d^\pi_s(\bar{M}\|M)\right\}$ and $d^\pi_s(M\|\bar{M})$ is defined as the fixed-point of the following fixed-point equation on $d \in \mathcal{F}(\mathcal{S}, \mathbb{R})$:

$$d_s = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left(D^{\gamma V^\pi_{\bar{M}}}_{sa}(M, \bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} d_{s'} \right) .$$

Before proving the Proposition, we show that the fixed point equation admits a unique solution in the following Lemma.

Lemma 2. *Given two MDPs $M, \bar{M} \in \mathcal{M}$, any stochastic stationary policy π, the following equation on $d \in \mathcal{F}(\mathcal{S}, \mathbb{R})$ is a fixed-point equation admitting a unique solution for any $s \in \mathcal{S}$:*

$$d_s = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left(D^{\gamma V^\pi_{\bar{M}}}_{sa}(M, \bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} d_{s'} \right) .$$

We refer to this unique solution as $d^\pi_s(M\|\bar{M})$.

Proof of Lemma 2. Let L be the functional operator that maps any function $d \in \mathcal{F}(\mathcal{S}, \mathbb{R})$ to

$$\begin{array}{rccl} Ld: & \mathcal{S} & \to & \mathbb{R} \\ & s & \mapsto & \sum_{a \in \mathcal{A}} \pi(a \mid s) \left(D^{\gamma V^\pi_{\bar{M}}}_{sa}(M, \bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} d_{s'} \right) \end{array} .$$

Then for f and g, two functions from $\mathcal{S}$ to $\mathbb{R}$, we have that

$$\begin{aligned} Lf_s - Lg_s &= \gamma \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} T^a_{ss'} \left(f_{s'} - g_{s'}\right) \\ &\leq \gamma \left\|f - g\right\|_\infty . \end{aligned}$$

Hence we have that $\left\|Lf - Lg\right\|_\infty \leq \gamma \left\|f - g\right\|_\infty$. Since $\gamma < 1$, L is a contraction mapping in the metric space $\left(\mathcal{F}(\mathcal{S}, \mathbb{R}), \left\|\cdot\right\|_\infty\right)$. This metric space being complete and non-empty, it follows by direct application of the Banach fixed-point theorem that the equation $d = Ld$ admits a unique solution. □

Proof of Proposition 10. Consider a stochastic stationary stationary policy π. The value iteration sequence of iterates $(V^{\pi,n}_M)_{n \in \mathbb{N}}$ of the value function of any MDP $M \in \mathcal{M}$ is defined for all $s \in \mathcal{S}$ by:

$$\begin{aligned} V^{\pi,0}_M(s) &= 0 , \\ V^{\pi,n+1}_M(s) &= \sum_{a \in \mathcal{A}} \pi(a \mid s) \left(R^a_s + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} V^{\pi,n}_M(s') \right) \end{aligned}$$

Consider two MDPs $M, \bar{M} \in \mathcal{M}$. It is obvious that $\left|V^{\pi,0}_M(s) - V^{\pi,0}_{\bar{M}}(s)\right| \leq d^\pi_s(M\|\bar{M})$ for all $s \in \mathcal{S}$. Suppose the property $\left|V^{\pi,n}_M(s) - V^{\pi,n}_{\bar{M}}(s)\right| \leq d^\pi_s(M\|\bar{M})$ true at rank $n \in \mathbb{N}$ for all $s \in \mathcal{S}$. Consider now the rank $n+1$ and the state $s \in \mathcal{S}$:

$$\begin{aligned} \left|V^{\pi,n+1}_M(s) - V^{\pi,n+1}_{\bar{M}}(s)\right| &\leq \sum_{a \in \mathcal{A}} \pi(a \mid s) \left| R^a_s - \bar{R}^a_s + \gamma \sum_{s' \in \mathcal{S}} \left(T^a_{ss'} V^{\pi,n}_M(s') - \bar{T}^a_{ss'} V^{\pi,n}_{\bar{M}}(s') \right) \right| \\ &\leq \sum_{a \in \mathcal{A}} \pi(a \mid s) \Bigg(\left|R^a_s - \bar{R}^a_s\right| + \gamma \sum_{s' \in \mathcal{S}} V^{\pi,n}_{\bar{M}}(s') \left|T^a_{ss'} - \bar{T}^a_{ss'}\right| \\ &\qquad + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \left|V^{\pi,n}_M(s') - V^{\pi,n}_{\bar{M}}(s')\right| \Bigg) \\ &\leq \sum_{a \in \mathcal{A}} \pi(a \mid s) \left(D^{\gamma V^\pi_{\bar{M}}}_{sa}(M, \bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} d^\pi_{s'}(M\|\bar{M}) \right) \\ &\leq d^\pi_s(M\|\bar{M}) , \end{aligned}$$

where we used Lemma 2 in the last inequality. Since V^{π}_{M} and $V^{\pi}_{\bar{M}}$ are respectively the limits of the sequences $(V_{M}^{\pi,n})_{n\in\mathbb{N}}$ and $\left(V_{\bar{M}}^{\pi,n}\right)_{n\in\mathbb{N}}$, it results from passage to the limit that

$$\left|V^{\pi}_{M}(s) - V^{\pi}_{\bar{M}}(s)\right| \leq d^{\pi}_{s}(M\|\bar{M})\,.$$

By symmetry, we also have $\left|V^{\pi}_{M}(s) - V^{\pi}_{\bar{M}}(s)\right| \leq d^{\pi}_{s}(\bar{M}\|M)$ and we can take the minimum of the two valid upper bounds, yielding:

$$\left|V^{\pi}_{M}(s) - V^{\pi}_{\bar{M}}(s)\right| \leq \min\left\{d^{\pi}_{s}(M\|\bar{M}), d^{\pi}_{s}(\bar{M}\|M)\right\}\,,$$

which concludes the proof. □

Proposition 11 (Local pseudo-Lipschitz continuity of the Q-function of any policy)**.** *For any two MDPs $M, \bar{M} \in \mathcal{M}$, for any stochastic stationary policy π, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\left|Q^{\pi}_{M}(s,a) - Q^{\pi}_{\bar{M}}(s,a)\right| \leq \Delta^{\pi}_{sa}(M,\bar{M})$$

where $\Delta^{\pi}_{sa}(M,\bar{M}) \triangleq \min\left\{d^{\pi}_{sa}(M\|\bar{M}), d^{\pi}_{sa}(\bar{M}\|M)\right\}$ and $d^{\pi}_{sa}(M\|\bar{M})$ is defined as the fixed-point of the following fixed-point equation on $d \in \mathcal{F}\left(\mathcal{S}\times\mathcal{A},\mathbb{R}\right)$:

$$d_{sa} = D_{sa}^{\gamma V^{\pi}_{\bar{M}}}(M,\bar{M}) + \gamma \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} T^{a}_{ss'}\pi(a' \mid s')d_{s'a'}\,.$$

Before proving the Proposition, we show that the fixed point equation admits a unique solution in the following Lemma.

Lemma 3. *Given two MDPs $M, \bar{M} \in \mathcal{M}$, any stochastic stationary policy π, the following equation on $d \in \mathcal{F}\left(\mathcal{S}\times\mathcal{A},\mathbb{R}\right)$ is a fixed-point equation admitting a unique solution for any $(s,a)\ \in \mathcal{S}\times\mathcal{A}$:*

$$d_{sa} = D_{sa}^{\gamma V^{\pi}_{\bar{M}}}(M,\bar{M}) + \gamma \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} T^{a}_{ss'}\pi(a' \mid s')d_{s'a'}\,.$$

We refer to this unique solution as $d^{\pi}_{sa}(M\|\bar{M})$.

Proof of Lemma 3. Let L be the functional operator that maps any function $d \in \mathcal{F}\left(\mathcal{S}\times\mathcal{A},\mathbb{R}\right)$ to

$$\begin{array}{rccl} Ld: & \mathcal{S}\times\mathcal{A} & \to & \mathbb{R} \\ & (s,a) & \mapsto & D_{sa}^{\gamma V^{\pi}_{\bar{M}}}(M,\bar{M}) + \gamma \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} T^{a}_{ss'}\pi(a' \mid s')d_{s',a'}\,. \end{array}$$

Then for f and g, two functions from $\mathcal{S}\times\mathcal{A}$ to $\mathbb{R}$, we have for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ that

$$\begin{aligned} Lf_{sa} - Lg_{sa} &= \gamma \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} T^{a}_{ss'}\pi(a' \mid s')\left(Lf_{s'a'} - Lg_{s'a'}\right) \\ &\leq \gamma \left\|f-g\right\|_{\infty}\,. \end{aligned}$$

Hence we have that $\left\|Lf - Lg\right\|_{\infty} \leq \gamma\left\|f-g\right\|_{\infty}$. Since $\gamma < 1$, L is a contraction mapping in the metric space $\left(\mathcal{F}\left(\mathcal{S}\times\mathcal{A},\mathbb{R}\right), \left\|\cdot\right\|_{\infty}\right)$. This metric space being complete and non-empty, it follows by direct application of the Banach fixed-point theorem that the equation $d = Ld$ admits a unique solution. □

Proof of Proposition 11. Consider a stochastic stationary policy π. The value iteration sequence of iterates $(Q_{M}^{\pi,n})_{n\in\mathbb{N}}$ of the Q function for the policy π and MDP $M \in \mathcal{M}$ is defined for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ by:

$$\begin{aligned} Q_{M}^{\pi,0}(s,a) &= 0\,, \\ Q_{M}^{\pi,n+1}(s,a) &= R^{a}_{s} + \gamma \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} T^{a}_{ss'}\pi(a' \mid s')Q_{M}^{\pi,n}(s',a') \end{aligned}$$

Consider two MDPs $M, \bar{M} \in \mathcal{M}$. It is obvious that $\left|Q_{M}^{\pi,0}(s,a) - Q_{\bar{M}}^{\pi,0}(s,a)\right| \leq d^{\pi}_{sa}(M\|\bar{M})$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Suppose the property $\left|Q_{M}^{\pi,n}(s,a) - Q_{\bar{M}}^{\pi,n}(s,a)\right| \leq d^{\pi}_{sa}(M\|\bar{M})$ true at rank $n\in\mathbb{N}$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Consider now the rank $n+1$

and the state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$
\begin{aligned}
\left| Q_M^{\pi,n+1}(s,a) - Q_{\bar{M}}^{\pi,n+1}(s,a) \right| &\leq \left| R_s^a - \bar{R}_s^a \right| + \gamma \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \pi(a' \mid s') \left| T_{ss'}^a Q_M^{\pi,n}(s',a') - \bar{T}_{ss'}^a Q_{\bar{M}}^{\pi,n}(s',a') \right| \\
&\leq \left| R_s^a - \bar{R}_s^a \right| + \gamma \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \pi(a' \mid s') Q_{\bar{M}}^{\pi,n}(s',a') \left| T_{ss'}^a - \bar{T}_{ss'}^a \right| \\
&\quad + \gamma \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \pi(a' \mid s') T_{ss'}^a \left| Q_M^{\pi,n}(s',a') - Q_{\bar{M}}^{\pi,n}(s',a') \right| \\
&\leq \left| R_s^a - \bar{R}_s^a \right| + \sum_{s' \in \mathcal{S}} \gamma V_{\bar{M}}^{\pi}(s') \left| T_{ss'}^a - \bar{T}_{ss'}^a \right| + \gamma \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \pi(a' \mid s') T_{ss'}^a d_{\pi}^{M,\bar{M}}(s',a') \\
&\leq D_{sa}^{\gamma V_{\bar{M}}^{\pi}}(M, \bar{M}) + \gamma \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} T_{ss'}^a \pi(a' \mid s') d_{s'a'}^{\pi}(M \| \bar{M}) \\
&\leq d_{sa}^{\pi}(M \| \bar{M}) \,,
\end{aligned}
$$

where we used Lemma 3 in the last inequality. Since Q_M^{π} and $Q_{\bar{M}}^{\pi}$ are respectively the limits of the sequences $(Q_M^{\pi,n})_{n \in \mathbb{N}}$ and $\left(Q_{\bar{M}}^{\pi,n}\right)_{n \in \mathbb{N}}$, it results from passage to the limit that

$$
\left| Q_M^{\pi}(s,a) - Q_{\bar{M}}^{\pi}(s,a) \right| \leq d_{sa}^{\pi}(M \| \bar{M}) \,.
$$

By symmetry, we also have $\left| Q_M^{\pi}(s,a) - Q_{\bar{M}}^{\pi}(s,a) \right| \leq d_{sa}^{\pi}(\bar{M} \| M)$ and we can take the minimum of the two valid upper bounds, yielding for all $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$
\left| Q_M^{\pi}(s,a) - Q_{\bar{M}}^{\pi}(s,a) \right| \leq \min \left\{ d_{sa}^{\pi}(M \| \bar{M}), d_{sa}^{\pi}(\bar{M} \| M) \right\} \,,
$$

which concludes the proof. □

5 Proof of Proposition 2

Proof of Proposition 2. The result is clear for all $(s, a) \notin K$ since the Lipschitz bounds are provably greater than Q_M^*. For $(s, a) \in K$, the result is shown by induction. Let us consider the Dynamic Programming (Bellman 1957) sequences converging to Q_M^* and U whose definitions follow for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and for $n \in \mathbb{N}$:

$$
\begin{cases} Q_M^0(s,a) = 0 \\ Q_M^{n+1}(s,a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \max_{a' \in \mathcal{A}} Q_M^n(s',a') \end{cases} ,
$$

$$
\begin{cases} U^0(s,a) = 0 \\ U^{n+1}(s,a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \max_{a' \in \mathcal{A}} U^n(s',a') \end{cases}
$$

Obviously, we have at rank $n = 0$ that $Q_M^0(s,a) \leq U^0(s,a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Suppose the property true at rank $n \in \mathbb{N}$ and consider rank $n + 1$:

$$
\begin{aligned}
Q_M^{n+1}(s,a) - U^{n+1}(s,a) &= \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \left(\max_{a' \in \mathcal{A}} Q_M^n(s',a') - \max_{a' \in \mathcal{A}} U^n(s',a') \right) \\
&\leq \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \max_{a' \in \mathcal{A}} \left(Q_M^n(s',a') - U^n(s',a') \right) \\
&\leq 0 \,.
\end{aligned}
$$

Which concludes the proof by induction. The result holds by passage to the limit since the considered Dynamic Programming sequences converge to the true functions. □

6 Proof of Proposition 3

Proof of Proposition 3. Consider two tasks $M = (T, R)$ and $\bar{M} = (\bar{T}, \bar{R})$, with K and $\bar{K}$ the respective sets of state-action pairs where their learned models $\hat{M} = (\hat{T}, \hat{R})$ and $\hat{\bar{M}} = (\hat{\bar{T}}, \hat{\bar{R}})$ are known with accuracy ϵ in $\mathcal{L}_1$-norm with probability at least $1 - \delta$, *i.e.*, we have that,

$$
\mathbf{Pr} \left(\begin{array}{lccl} \left| R_s^a - \hat{R}_s^a \right| & \leq & \epsilon, & \forall \, (s,a) \in K \quad \text{and} \\ \left\| T_{ss'}^a - \hat{T}_{ss'}^a \right\|_1 & \leq & \epsilon, & \forall \, (s,a) \in K \quad \text{and} \\ \left| \bar{R}_s^a - \hat{\bar{R}}_s^a \right| & \leq & \epsilon, & \forall \, (s,a) \in \bar{K} \quad \text{and} \\ \left\| \bar{T}_{ss'}^a - \hat{\bar{T}}_{ss'}^a \right\|_1 & \leq & \epsilon, & \forall \, (s,a) \in \bar{K} \end{array} \right) \leq 1 - \delta. \tag{13}
$$

Importantly, notice that the probabilistic event of Inequality 13 is the intersection of at most $4SA$ individual events of estimating either R_s^a, $T_{ss'}^a$, $\bar{R}_s^a$ or $\bar{T}_{ss'}^a$ with precision ϵ. Each one of those individual events is itself true with probability at least $1-\delta'$, where $\delta' \in (0,1]$ is a parameter. For *all* the individual events to be true at the same time, *i.e.* for Inequality 13 to be verified, one must apply Boole's inequality and set $\delta' = \delta/(4SA)$ to ensure a total probability — *i.e.*, probability of the intersection of all the individual events — of at least $1-\delta$.

We demonstrate now the result for each one of the three cases

(i) $(s,a) \in K \cap \bar{K}$,

(ii) $(s,a) \in K \cap \bar{K}^c$ and

(iii) $(s,a) \in K^c \cap \bar{K}^c$,

the case $(s,a) \in K^c \cap \bar{K}$ being the symmetric of case (ii).

(i) If $(s,a) \in K \cap \bar{K}$, then we have ϵ-close estimates of both models with high probability, as described by Inequality 13. By definition:

$$D_{sa}^{\gamma V^*_{\bar{M}}}(M,\bar{M}) = \left|R_s^a - \bar{R}_s^a\right| + \gamma \sum_{s' \in \mathcal{S}} V^*_{\bar{M}}(s') \left|T_{ss'}^a - \bar{T}_{ss'}^a\right| . \tag{14}$$

The first term of the right hand side of Equation 14 respects the following sequence of inequalities with probability at least $1-\delta$:

$$\begin{aligned} \left|R_s^a - \bar{R}_s^a\right| &\leq \left|R_s^a - \hat{R}_s^a\right| + \left|\hat{R}_s^a - \hat{\bar{R}}_s^a\right| + \left|\bar{R}_s^a - \hat{\bar{R}}_s^a\right| \\ &\leq \left|\hat{R}_s^a - \hat{\bar{R}}_s^a\right| + 2\epsilon . \end{aligned} \tag{15}$$

The second term of the right hand side of Equation 14 respects the following sequence of inequalities with probability at least $1-\delta$:

$$\begin{aligned} \gamma \sum_{s' \in \mathcal{S}} V^*_{\bar{M}}(s') \left|T_{ss'}^a - \bar{T}_{ss'}^a\right| &\leq \gamma \sum_{s' \in \mathcal{S}} \bar{V}(s') \left(\left|T_{ss'}^a - \hat{T}_{ss'}^a\right| + \left|\hat{T}_{ss'}^a - \hat{\bar{T}}_{ss'}^a\right| + \left|\bar{T}_{ss'}^a - \hat{\bar{T}}_{ss'}^a\right| \right) \\ &\leq \gamma \max_{s'} \bar{V}(s') \sum_{s' \in \mathcal{S}} \left|T_{ss'}^a - \hat{T}_{ss'}^a\right| + \gamma \sum_{s' \in \mathcal{S}} \bar{V}(s') \left|\hat{T}_{ss'}^a - \hat{\bar{T}}_{ss'}^a\right| + \\ &\quad \gamma \max_{s'} \bar{V}(s') \sum_{s' \in \mathcal{S}} \left|\bar{T}_{ss'}^a - \hat{\bar{T}}_{ss'}^a\right| \\ &\leq \gamma \sum_{s' \in \mathcal{S}} \bar{V}(s') \left|\hat{T}_{ss'}^a - \hat{\bar{T}}_{ss'}^a\right| + 2\epsilon\gamma \max_{s' \in \mathcal{S}} \bar{V}(s'). \end{aligned} \tag{16}$$

Replacing the Inequalities 15 and 16 in Equation 14 yields

$$\begin{aligned} D_{sa}(M \| \bar{M}) &\leq \left|\hat{R}_s^a - \hat{\bar{R}}_s^a\right| + \gamma \sum_{s' \in \mathcal{S}} \bar{V}(s') \left|\hat{T}_{ss'}^a - \hat{\bar{T}}_{ss'}^a\right| + 2\epsilon + 2\epsilon\gamma \max_{s' \in \mathcal{S}} \bar{V}(s') \\ &\leq D_{sa}^{\gamma \bar{V}}(\hat{M}, \hat{\bar{M}}) + 2\epsilon \left(1 + \gamma \max_{s' \in \mathcal{S}} \bar{V}(s')\right) , \end{aligned}$$

which holds with probability at least $1-\delta$ and proves the Theorem for case (i).

(ii) If $(s,a) \in K \cap \bar{K}^c$, then we do not have an ϵ-close estimate of $\bar{T}_{s\cdot}^a$ and $\bar{R}_s^a$. Similarly to the proof of case (i), we upper bound sequentially the two terms of the right hand side of Equation 14. With probability at least $1-\delta$, we have the following:

$$\begin{aligned} \left|R_s^a - \bar{R}_s^a\right| &\leq \left|R_s^a - \hat{R}_s^a\right| + \left|\hat{R}_s^a - \bar{R}_s^a\right| \\ &\leq \epsilon + \max_{\bar{R} \in [0,1]} \left|\hat{R}_s^a - \bar{R}\right| . \end{aligned} \tag{17}$$

Similarly, with probability at least $1-\delta$, we have:

$$\begin{aligned} \gamma \sum_{s' \in \mathcal{S}} V^*_{\bar{M}}(s') \left|T_{ss'}^a - \bar{T}_{ss'}^a\right| &\leq \gamma \sum_{s' \in \mathcal{S}} \bar{V}(s') \left(\left|T_{ss'}^a - \hat{T}_{ss'}^a\right| + \left|\hat{T}_{ss'}^a - \bar{T}_{ss'}^a\right| \right) \\ &\leq \gamma \max_{s' \in \mathcal{S}} \bar{V}(s')\epsilon + \gamma \max_{\bar{T} \in \mathcal{V}_S} \sum_{s' \in \mathcal{S}} \bar{V}(s') \left|\hat{T}_{ss'}^a - \bar{T}_{s'}\right| , \end{aligned} \tag{18}$$

where $\mathcal{V}_S$ is the set of probability vectors of size S. Combining inequalities 17 and 18, we get the following with probability at least $1-\delta$, by noticing $D_{sa}^{\gamma V^*_{\bar{M}}}(M, \bar{M})$ on the left hand side:

$$D_{sa}(M\|\bar{M}) \leq \max_{\bar{m}\in\mathcal{M}} D_{sa}^{\gamma\bar{V}}(\hat{M}, \bar{m}) + \epsilon\left(1 + \gamma \max_{s'} \bar{V}(s')\right),$$

which is the expected result for case (ii).

(iii) If $(s,a) \in K^c \cap \bar{K}^c$, then we do not have ϵ-close estimates of both tasks. In such a case, the result

$$D_{sa}(M\|\bar{M}) \leq \max_{m,\bar{m}\in\mathcal{M}^2} D_{sa}^{\gamma\bar{V}}(m, \bar{m})$$

is straightforward by remarking that, as a consequence of Inequality 13, we have that $V^*_{\bar{M}}(s) \leq \bar{V}(s)$ with probability at least $1-\delta$. □

7 Analytical calculation of $\hat{D}_{sa}(M\|\bar{M})$ in Proposition 3

Consider two tasks $M = (T, R)$ and $\bar{M} = (\bar{T}, \bar{R})$, with K and $\bar{K}$ the respective sets of state-action pairs where their learned models $\hat{M} = (\hat{T}, \hat{R})$ and $\hat{\bar{M}} = (\hat{\bar{T}}, \hat{\bar{R}})$ are known with accuracy ϵ in $\mathcal{L}_1$-norm with probability at least $1-\delta$. We note V_{max}, a known upper bound on the maximum achievable value. In the worst case where one does not have any information on the value of V_{max}, setting $V_{\text{max}} = \frac{1}{1-\gamma}$ is a valid upper bound. We detail the computation of $\hat{D}_{sa}(M\|\bar{M})$ for each cases: 1) $(s,a) \in K \cap \bar{K}$, 2) $(s,a) \in K \cap \bar{K}^c$, and 3) $(s,a) \in K^c \cap \bar{K}^c$. The case $(s,a) \in K^c \cap \bar{K}$ being the symmetric of case 2), the same calculations apply.

1) If $(s,a) \in K \cap \bar{K}$, we have

$$\begin{aligned}
\hat{D}_{sa}(M\|\bar{M}) &= D_{sa}^{\gamma\bar{V}}(\hat{M}, \hat{\bar{M}}) + 2B \\
&= \left|\hat{R}_s^a - \hat{\bar{R}}_s^a\right| + \gamma \sum_{s'\in\mathcal{S}} \bar{V}(s') \left|\hat{T}_{ss'}^a - \hat{\bar{T}}_{ss'}^a\right| + 2\epsilon\left(1 + \gamma \max_{s'\in\mathcal{S}} \bar{V}(s')\right).
\end{aligned}$$

Since (s,a) is a known state-action pair, everything is known and computable in this last equation. Note that $\max_{s'\in\mathcal{S}} \bar{V}(s')$ can be tracked along the updates of $\bar{V}$ and thus its computation does not induce any additional computational complexity.

2) If $(s,a) \in K \cap \bar{K}^c$, we have

$$\begin{aligned}
\hat{D}_{sa}(M\|\bar{M}) &= \max_{\bar{\mu}\in\mathcal{M}} D_{sa}^{\gamma\bar{V}}(\hat{M}, \bar{\mu}) + B \\
&= \max_{\bar{R}_s^a, \bar{T}_{ss'}^a} \left(\left|\hat{R}_s^a - \bar{R}_s^a\right| + \gamma \sum_{s'\in\mathcal{S}} \bar{V}(s') \left|\hat{T}_{ss'}^a - \bar{T}_{ss'}^a\right|\right) + \epsilon\left(1 + \gamma \max_{s'\in\mathcal{S}} \bar{V}(s')\right), \\
&= \max_{r\in[0,1]} \left|\hat{R}_s^a - r\right| + \gamma \max_{\substack{t\in[0,1]^S \\ \text{s.t. } \sum_{s'\in\mathcal{S}} t_{s'}=1}} \left(\sum_{s'\in\mathcal{S}} \bar{V}(s') \left|\hat{T}_{ss'}^a - t_{s'}\right|\right) + \epsilon\left(1 + \gamma \max_{s'\in\mathcal{S}} \bar{V}(s')\right).
\end{aligned}$$

First, we have

$$\max_{r\in[0,1]} \left|\hat{R}_s^a - r\right| = \max\left\{\hat{R}_s^a, 1 - \hat{R}_s^a\right\}.$$

Maximizing over the variable $t \in [0,1]^S$ such that $\sum_{s'\in\mathcal{S}} t_{s'} = 1$ is equivalent to maximizing a convex combination of the positive vector $\bar{V}$ whose terms are not independent as they must sum to one. This is easily solvable as a linear programming problem. A straightforward (simplex-like) resolution procedure consists in progressively adding mass on the terms that will maximize the convex combination as follows:

- $t_{s'} = 0, \forall s' \in \mathcal{S}$
- l = Sort states by decreasing values of $\bar{V}$
- While $\sum_{s\in\mathcal{S}} t_s < 1$
  - s' = pop first state in l
  - Assign $t_{s'} \leftarrow \arg\max_{t\in[0,1]} \left|\hat{T}_{ss'}^a - t\right|$ to s' (note that $t_{s'} \in \{0,1\}$)
  - If $\sum_{s\in\mathcal{S}} t_s > 1$, then $t_{s'} \leftarrow 1 - \sum_{s\in\mathcal{S}\setminus s'} t(s)$

This allows calculating the maximum over transition models.

Notice that there is a simpler computation that almost always yields the same result (when it does not, it provides an upper bound) and does not require the burden of the previous procedure. Consider the subset of states for which $\bar{V}(s') = \max_{s\in\mathcal{S}} \bar{V}(s)$ (often these are states in $\bar{K}^c$). Among those states, let us suppose there exists s^+, unreachable from (s,a), according to $\hat{T}$, *i.e.*, $\hat{T}^a_{ss^+} = 0$. If $\bar{M}$ has not been fully explored, as is often the case in RMax, there may be many such states. Then the distribution t with all its mass on s^+ maximizes the $\max_{t\in[0,1]^S}$ term. Conversely, if such a state does not exist (that is, if for all such states $\hat{T}^a_{ss^+} > 0$), then $\max_{s\in\mathcal{S}} \bar{V}(s)$ is an upper bound on the $\max_{t\in[0,1]^S}$ term. Therefore:

$$\max_{t\in[0,1]^S} \left(\sum_{s'\in\mathcal{S}} \bar{V}(s') \left| \hat{T}^a_{ss'} - t_{s'} \right| \right) \leq \max_{s\in\mathcal{S}} \bar{V}(s)\,,$$

with equality in many cases.

3) If $(s,a) \in K^c \cap \bar{K}^c$, the resolution is trivial and we have

$$\begin{aligned}
\hat{D}_{sa}(M\|\bar{M}) &= \max_{\mu,\bar{\mu}\in\mathcal{M}^2} D^{\gamma\bar{V}}_{sa}(\mu,\bar{\mu}) \\
&= \max_{R^a_s, T^a_{ss'}, \bar{R}^a_s, \bar{T}^a_{ss'}} \left(\left| R^a_s - \bar{R}^a_s \right| + \gamma \sum_{s'\in\mathcal{S}} \bar{V}(s') \left| T^a_{ss'} - \bar{T}^a_{ss'} \right| \right) \\
&= \max_{r,\bar{r}\in[0,1]} |r - \bar{r}| + \gamma \max_{\substack{t,\bar{t}\in[0,1]^S \\ \text{s.t. } \sum_{s\in\mathcal{S}} t_s = 1 \\ \text{and } \sum_{s\in\mathcal{S}} \bar{t}_s = 1}} \sum_{s'\in\mathcal{S}} \bar{V}(s') \left| t_{s'} - \bar{t}_{s'} \right| \\
&= 1 + 2\gamma \max_{s\in\mathcal{S}} \bar{V}(s)\,.
\end{aligned}$$

Overall, computing the value of the provided upper bound in the three cases allows to compute $\hat{D}_{sa}(M\|\bar{M})$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

8 Proof of Proposition 4

Lemma 4. *Given two tasks $M, \bar{M} \in \mathcal{M}$, K the set of state-action pairs for which (R,T) is known with accuracy ϵ in $\mathcal{L}_1$-norm with probability at least $1-\delta$. If $\gamma(1+\epsilon) < 1$, this equation on $\hat{d} \in \mathcal{F}(\mathcal{S}\times\mathcal{A}, \mathbb{R})$ is a fixed-point equation admitting a unique solution.*

$$\hat{d}_{s,a} = \begin{cases} \hat{D}_{sa}(M\|\bar{M}) + \gamma\Big(\sum_{s'\in\mathcal{S}} \hat{T}^a_{ss'} \max_{a'\in\mathcal{A}} \hat{d}_{s',a'} + \epsilon \max_{(s',a')\in\mathcal{S}\times\mathcal{A}} \hat{d}_{s',a'} \Big) & \textit{if } (s,a)\in K, \\ \hat{D}_{sa}(M\|\bar{M}) + \gamma \max_{(s',a')\in\mathcal{S}\times\mathcal{A}} \hat{d}_{s',a'} & \textit{else.} \end{cases}$$

We refer to this unique solution as $\hat{d}_{sa}(M\|\bar{M})$.

Proof of Lemma 4. Let L be the functional operator that maps any function $d \in \mathcal{F}(\mathcal{S}\times\mathcal{A}, \mathbb{R})$ to

$$\begin{aligned}
Ld:\ \ \mathcal{S}\times\mathcal{A} &\ \to\ \mathbb{R} \\
(s,a) &\ \mapsto\ \begin{cases} \hat{D}_{sa}(M\|\bar{M}) + \gamma \left(\sum_{s'\in\mathcal{S}} \hat{T}^a_{ss'} \max_{a'\in\mathcal{A}} d_{s',a'} + \epsilon \max_{(s',a')\in\mathcal{S}\times\mathcal{A}} d_{s',a'} \right) & \text{if } (s,a)\in K, \\ \hat{D}_{sa}(M\|\bar{M}) + \gamma \max_{(s',a')\in\mathcal{S}\times\mathcal{A}} d_{s',a'} & \text{otherwise.} \end{cases}
\end{aligned}$$

Let f and g be two functions from $\mathcal{S}\times\mathcal{A}$ to $\mathbb{R}$. If $(s,a)\in K$, we have

$$\begin{aligned}
Lf_{sa} - Lg_{sa} &= \gamma \sum_{s'\in\mathcal{S}} T^a_{ss'} \left(\max_{a'\in\mathcal{A}} f_{s'a'} - \max_{a'\in\mathcal{A}} g_{s'a'} \right) + \gamma\epsilon \left(\max_{(s',a')\in\mathcal{S}\times\mathcal{A}} f_{s'a'} - \max_{(s',a')\in\mathcal{S}\times\mathcal{A}} g_{s'a'} \right) \\
&\leq (\gamma + \gamma\epsilon) \left(\max_{(s',a')\in\mathcal{S}\times\mathcal{A}} f_{s'a'} - \max_{(s',a')\in\mathcal{S}\times\mathcal{A}} g_{s'a'} \right) \\
&\leq \gamma(1+\epsilon) \max_{(s',a')\in\mathcal{S}\times\mathcal{A}} \left(f_{s'a'} - g_{s'a'} \right) \\
&\leq \gamma(1+\epsilon) \left\| f - g \right\|_\infty\,.
\end{aligned}$$

If $(s, a) \notin K$, we have

$$\begin{aligned} Lf_{sa} - Lg_{sa} &= \gamma \left(\max_{(s',a') \in \mathcal{S} \times \mathcal{A}} f_{s'a'} - \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} g_{s'a'} \right) \\ &\leq \gamma \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \left(f_{s'a'} - g_{s'a'} \right) \\ &= \gamma(1 + \epsilon) \left\| f - g \right\|_\infty . \end{aligned}$$

In both cases, $\|Lf - Lg\|_\infty \leq \gamma(1 + \epsilon) \|f - g\|_\infty$. If $\gamma(1 + \epsilon) < 1$, L is a contraction mapping in the metric space $(\mathcal{F}(\mathcal{S} \times \mathcal{A}, \mathbb{R}), \| \cdot \|_\infty)$. This metric space being complete and non-empty, it follows from Banach fixed-point theorem that $d = Ld$ admits a single solution. □

Proof of Proposition 4. Consider two MDPs $M, \bar{M} \in \mathcal{M}$. Before proving the result, notice that we shall put ourselves in the case of Proposition 3, for the upper bound on the pseudometric between models $\hat{D}_{sa}(M \| \bar{M})$ to be true upper bounds with probability at least $1 - \delta$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. As seen in the proof of Proposition 3, this implies learning any reward or transition function with precision ϵ in $\mathcal{L}_1$-norm with probability at least $1 - \delta/(4SA)$.

The proof is done by induction, by calculating the values of $d_{sa}(M \| \bar{M})$ and $\hat{d}_{sa}(M \| \bar{M})$ following the value iteration algorithm. Those values can respectively be computed via the sequences of iterates $(d^n)_{n \in \mathbb{N}}$ and $(\hat{d}^n)_{n \in \mathbb{N}}$ defined as follows for all $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\begin{aligned} d^0_{sa}(M \| \bar{M}) &= 0 \\ d^{n+1}_{sa}(M \| \bar{M}) &= D_{sa}(M \| \bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} d^n_{s'a'}(M \| \bar{M}) , \end{aligned}$$

and,

$$\begin{aligned} &\hat{d}^0_{sa}(M \| \bar{M}) = 0, \\ &\hat{d}^{n+1}_{sa}(M \| \bar{M}) = \begin{cases} \hat{D}_{sa}(M \| \bar{M}) + \gamma \left(\sum_{s' \in \mathcal{S}} \hat{T}^{a}_{ss'} \max_{a' \in \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) + \epsilon \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) \right) & \text{if } (s, a) \in K, \\ \hat{D}_{sa}(M \| \bar{M}) + \gamma \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) \text{ otherwise.} \end{cases} \end{aligned}$$

The proof at rank $n = 0$ is trivial. Let us assume the proposition $d^n_{sa}(M \| \bar{M}) \leq \hat{d}^n_{sa}(M \| \bar{M})$, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$ true at rank $n \in \mathbb{N}$ and consider rank $n + 1$. There are two cases, depending on the fact that (s, a) is in K or not.

If $(s, a) \in K$, we have

$$\begin{aligned} d^{n+1}_{sa}(M \| \bar{M}) - \hat{d}^{n+1}_{sa}(M \| \bar{M}) = &D_{sa}(M \| \bar{M}) - \hat{D}_{sa}(M \| \bar{M}) \\ &+ \gamma \sum_{s' \in \mathcal{S}} \left(T^a_{ss'} \max_{a' \in \mathcal{A}} d^n_{s'a'}(M \| \bar{M}) - \hat{T}^{a}_{ss'} \max_{a' \in \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) \right) \\ &- \gamma \epsilon \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) . \end{aligned}$$

Using Proposition 3, we have that $\hat{D}_{sa}(M \| \bar{M})$ is an upper bound on $D_{sa}(M \| \bar{M})$ with probability at least $1 - \delta$. Hence

$$\mathbf{Pr} \left(D_{sa}(M \| \bar{M}) - \hat{D}_{sa}(M \| \bar{M}) \leq 0 \right) \geq 1 - \delta.$$

This plus the fact that $d^n_{sa}(M \| \bar{M}) \leq \hat{d}^n_{sa}(M \| \bar{M})$, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$ by induction hypothesis, we have with probability at least $1 - \delta$,

$$\begin{aligned} d^{n+1}_{sa}(M \| \bar{M}) - \hat{d}^{n+1}_{sa}(M \| \bar{M}) &\leq \gamma \sum_{s' \in \mathcal{S}} \max_{a' \in \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) \left(T^a_{ss'} - \hat{T}^{a}_{ss'} \right) - \gamma \epsilon \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) \\ &\leq \gamma \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) \sum_{s' \in \mathcal{S}} \left(T^a_{ss'} - \hat{T}^{a}_{ss'} \right) - \gamma \epsilon \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) \\ &\leq \gamma \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) \left(\left\| T - \hat{T} \right\|_1 - \epsilon \right) . \end{aligned}$$

Since $\mathbf{Pr} \left(\left\| T - \hat{T} \right\|_1 \leq \epsilon \right) \geq 1 - \delta$, we have with probability at least $1 - \delta$,

$$d^{n+1}_{sa}(M \| \bar{M}) - \hat{d}^{n+1}_{sa}(M \| \bar{M}) \leq \gamma \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}^n_{s'a'}(M \| \bar{M}) \left(\epsilon - \epsilon \right) = 0,$$

which concludes the proof in the first case case.

Conversely, if $(s, a) \notin K$, we have

$$d_{sa}^{n+1}(M \| \bar{M}) - \hat{d}_{sa}^{n+1}(M \| \bar{M}) = D_{sa}(M \| \bar{M}) - \hat{D}_{sa}(M \| \bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^{a} \max_{a' \in \mathcal{A}} d_{s'a'}^{n}(M \| \bar{M}) - \gamma \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}_{s'a'}^{n}(M \| \bar{M}) \,.$$

Using the same reasoning than in case $(s, a) \in K$, we have with probability higher than $1 - \delta$,

$$\begin{aligned} d_{sa}^{n+1}(M \| \bar{M}) - \hat{d}_{sa}^{n+1}(M \| \bar{M}) &\leq \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^{a} \max_{a' \in \mathcal{A}} \hat{d}_{s'a'}^{n}(M \| \bar{M}) - \gamma \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}_{s'a'}^{n}(M \| \bar{M}) \\ &\leq \gamma \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}_{s'a'}^{n}(M \| \bar{M}) - \gamma \max_{(s',a') \in \mathcal{S} \times \mathcal{A}} \hat{d}_{s'a'}^{n}(M \| \bar{M}) \\ &\leq 0, \end{aligned}$$

which concludes the proof in the second case. □

9 Proof of Proposition 6

Proof of Proposition 6. The cost of LRMax is constant on most time steps since the action is greedily chosen w.r.t. the upper bound on the optimal Q-function, which is a lookup table. Let $N \in \mathbb{N}$ be the number of source tasks that have been learned by LRMax during a lifelong RL experiment. When updating a new state-action pair, *i.e.*, labeling it as a known pair, the algorithm performs $2N$ Dynamic Programming (DP) computations to update the induced Lipschitz bounds (Equation 8) plus one DP computation to update the total-bound (Equation 6). In total, we apply $(2N + 1)$ DP computations for each state-action pair update. As at most SA state-action pairs are updated during the exploration of the current MDP, the total number of DP computations is at most $SA(2N + 1)$, for which we use the value iteration algorithm.

We use the value iteration as a Dynamic Programming method. Strehl, Li, and Littman (2009) report the minimum number of iterations needed by the value iteration algorithm to estimate a value function (or Q-function in our case) that is ϵ_Q-close to the optimum in maximum norm. This minimum number is given by

$$\left\lceil \frac{1}{1-\gamma} \ln\left(\frac{1}{\epsilon_Q(1-\gamma)}\right) \right\rceil .$$

Each iteration has a cost $S^2 A$. Overall, the cost of all the DP computations in a complete run of LRMax is

$$\tilde{\mathcal{O}}\left(\frac{S^3 A^2 N}{1-\gamma} \ln\left(\frac{1}{\epsilon_Q(1-\gamma)}\right)\right) .$$

This, plus the constant cost $\mathcal{O}(1)$ applied on each one of the τ decision epochs concludes the proof. □

10 Proof of Proposition 7

Proof of Proposition 7. Consider an algorithm producing ϵ-accurate model estimates $\hat{D}_{sa}(M \| \bar{M})$ for a subset K of $\mathcal{S} \times \mathcal{A}$ after interacting with any two MDPs $M, \bar{M} \in \mathcal{M}$. Assume $\hat{D}_{sa}(M \| \bar{M})$ to be an upper bound of $D_{sa}(M \| \bar{M})$ for any $(s, a) \notin K$. These assumptions are guaranteed with high probability by Proposition 3 while running Algorithm 1 in the lifelong RL setting. Then, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and any two MDPs $M, \bar{M} \in \mathcal{M}$, we have that

$$\begin{array}{rcll} \hat{D}_{sa}(M \| \bar{M}) & = & D_{sa}(M \| \bar{M}) \pm \epsilon & \text{if } (s, a) \in K \\ \hat{D}_{sa}(M \| \bar{M}) & \geq & D_{sa}(M \| \bar{M}) & \text{else.} \end{array}$$

Particularly, $\hat{D}_{sa}(M \| \bar{M}) + \epsilon \geq D_{sa}(M \| \bar{M})$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and any $M, \bar{M} \in \mathcal{M}$. By definition of $D_{\max}(s, a)$, this implies that, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\max_{M, \bar{M} \in \tilde{\mathcal{M}}} \hat{D}_{sa}(M \| \bar{M}) + \epsilon \geq D_{\max}(s, a) \,, \tag{19}$$

where $\tilde{M}$ is the set of possible tasks in the considered lifelong RL experiment. Consider $\hat{\mathcal{M}}$, the set of sampled MDPs which allows to define $\hat{D}_{\max}(s, a) = \max_{M, \bar{M} \in \hat{\mathcal{M}}} \hat{D}_{sa}(M \| \bar{M})$ as the maximum model distance for all the experienced MDPs at $(s, a) \in \mathcal{S} \times \mathcal{A}$. We have that

$$\hat{D}_{\max}(s, a) = \max_{M, \bar{M} \in \tilde{\mathcal{M}}} \hat{D}_{sa}(M \| \bar{M}) \,,$$

only if two MDPs maximizing the right hand side of this equation belong to $\hat{M}$. If it is the case, then Equation 19 imply that

$$\hat{D}_{\max}(s, a) + \epsilon \geq D_{\max}(s, a) \,. \tag{20}$$

Overall, we require the two MDPs maximizing $\max_{M,\bar{M}\in\tilde{\mathcal{M}}} \hat{D}_{sa}(M\|\bar{M})$ to be sampled for Equation 20 to hold. Let us now derive the probability that those two MDPs have been sampled. We note them M_1 and M_2. There may exist more candidates for the maximization but, for the sake of generality, we put ourselves in the case where only two MDPs achieve the maximization. Let us consider drawing $m \in \mathbb{N}$ tasks. We note p_1 (respectively p_2) the probability of sampling M_1 (respectively M_2) each time a task is sampled. We note X_1 (respectively X_2) the random variable of the first occurrence of the task M_1 (respectively M_2) among the m trials. Hence, the probability of sampling M_1 for the first time at trial $k \in \{1, \dots, m\}$ is given by the geometric law and is equal to

$$\mathbf{Pr}\left(X_1 = k\right) = p_1 \left(1 - p_1\right)^{k-1} .$$

Additionally, the probability of sampling M_1 at least once in the first m trials is given by the cumulative distribution function:

$$\mathbf{Pr}\left(X_1 \leq m\right) = 1 - \left(1 - p_1\right)^m . \tag{21}$$

We are interested in the probability of the event that M_1 *and* M_2 have been sampled in the m first trials, *i.e.* $\mathbf{Pr}\left(X_1 \leq m \cap X_2 \leq m\right)$. Following the rule of addition for probabilities, we have that,

$$\mathbf{Pr}\left(X_1 \leq m \cap X_2 \leq m\right) = \mathbf{Pr}\left(X_1 \leq m\right) + \mathbf{Pr}\left(X_2 \leq m\right) - \mathbf{Pr}\left(X_1 \leq m \cup X_2 \leq m\right) .$$

Given that the event of sampling either M_1 or M_2 during a single trial happens with probability $p_1 + p_2$, we have by analogy with Equation 21 that $\mathbf{Pr}\left(X_1 \leq m \cup X_2 \leq m\right) = 1 - (1 - (p_1 + p_2))^m$. As a result, the following holds:

$$\begin{aligned}
\mathbf{Pr}\left(X_1 \leq m \cap X_2 \leq m\right) &= 1 - (1 - p_1)^m + 1 - (1 - p_2)^m - (1 - (1 - (p_1 + p_2))^m) \\
&= 1 - (1 - p_1)^m - (1 - p_2)^m + (1 - (p_1 + p_2))^m \\
&\geq 1 - 2(1 - p_{\min})^m + (1 - 2p_{\min})^m .
\end{aligned}$$

As said earlier, Equation 20 holds if M_1 and M_2 have been sampled during the first m trials, which imply that the probability for Equation 20 to hold is at least equal to the probability of sampling both tasks. Formally,

$$\begin{aligned}
\mathbf{Pr}\left(\hat{D}_{\max}(s,a) + \epsilon \geq D_{\max}(s,a)\right) &\geq \mathbf{Pr}\left(X_1 \leq m \cap X_2 \leq m\right) \\
&\geq 1 - 2(1 - p_{\min})^m + (1 - 2p_{\min})^m .
\end{aligned}$$

In turn, if m verifies $2(1 - p_{\min})^m - (1 - 2p_{\min})^m \leq \delta$, then $1 - 2(1 - p_{\min})^m + (1 - 2p_{\min})^m \geq 1 - \delta$ and $\mathbf{Pr}\left(\hat{D}_{\max}(s,a) + \epsilon \geq D_{\max}(s,a)\right) \geq 1 - \delta$, which concludes the proof. □

11 Discussion on an upper bound on distances between MDP models

Section 4.4 introduced the idea of exploiting *prior* knowledge on the maximum distance between two MDP models. This idea begs for a more detailed discussion. Consider two MDPs M and $\bar{M}$. By definition of the local model pseudo metric in Equation 1, the maximum possible distance is given by

$$\max_{M,\bar{M}\in\mathcal{M}^2} D_{sa}(M\|\bar{M}) = \frac{1+\gamma}{1-\gamma}.$$

But this assumes that *any* transition or reward model can define M and $\bar{M}$. In other words, the maximization is made on the whole set of possible MDPs. To illustrate why this is too naive, consider a game within the Arcade Learning Environment (Bellemare et al. 2013). We, as humans, have a strong bias concerning similarity between environments. If the game changes, we still assume groups of pixels will move together on the screen as the result of game actions. For instance, we generally discard possible new games $\bar{M}$ that "teleport" objects across the screen without physical considerations. We also discard new games that allow transitions from a given screen to another screen full of static. These examples illustrate why the knowledge of $D_{\max}$ is very natural (and also why its precise value may be irrelevant). The same observation can be made for the "tight" experiment of Section 5; the set of possible MDPs is restricted by some implicit assumptions that constrain the maximum distance between tasks. For instance, in these experiments, all transitions move to a neighboring state and never "teleport" the agent to the other side of the gridworld. Without the knowledge of $D_{\max}$, LRMax assumes such environments are possible and therefore transfer values very cautiously (with the ultimate goal not to under estimate the optimal Q-function, in order to avoid negative transfer). Overall, the experiments of Section 5 confirm this important insight: safe transfer occurs slowly if no a priori is given on the maximum distance between MDPs. On the contrary, the knowledge of $D_{\max}$ allows a faster and more efficient transfer between environments.

12 The "tight" environment used in experiments of Section 5

The tight environment is a 11×11 grid-world illustrated in Figure 2. The initial state of the agent is the central cell displayed with an "S". The actions are moving 1 cell in one of the four cardinal directions. The reward is 0 everywhere, except for executing an action in one of the three teal cells in the upper-right corner. Each time a task is sampled, a slipping probability of executing another action as the one selected is drawn in $[0, 1]$ and the reward received in each one of the teal cells is picked in $[0.8, 1.0]$.

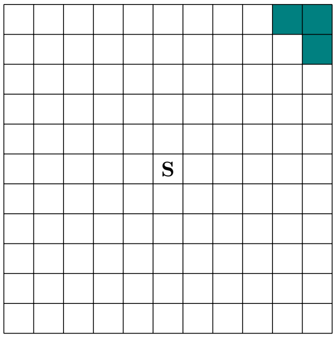


Figure 2: The tight grid-world environment.

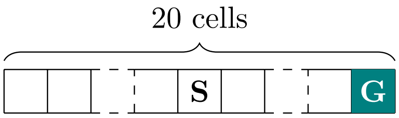


Figure 3: The corridor grid-world environment.

13 Additional lifelong RL experiments

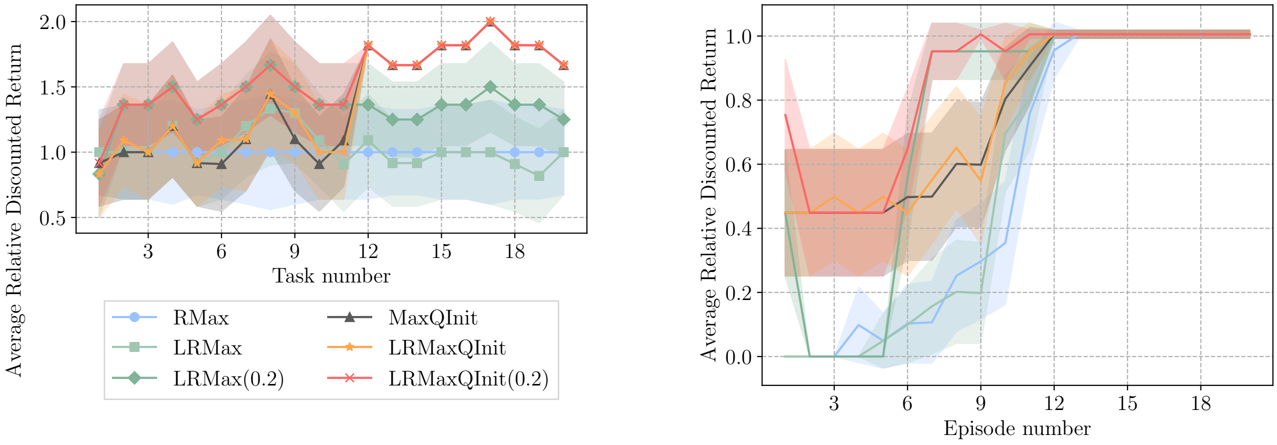
We ran additional experiments on the corridor grid-world environment represented in Figure 3. The initial state of the agent is the central cell labeled with the letter “S”. The actions are $\{$left, right$\}$ and the goal is to reach the cell labeled with the letter “G” on the extreme right. A reward $R > 0$ is received when reaching the goal and 0 otherwise. At each new task, a new value of R is sampled in $[0.8, 1]$. The transition function is fixed and deterministic.

The key insight in this experiment is not to lose time exploring the left part of the corridor. We ran 20 episodes of 11 time steps for each one of the 20 sampled tasks. Results are displayed in Figure 4a and 4b, respectively for the average relative discounted return over episodes and over tasks. Similarly as in Section 5, we observe in Figure 4a that LRMax benefits from the transfer method as early as the second task. The MaxQInit algorithm benefits from the transfer from task number 12. Prior knowledge $D_{\max}$ decreases the sample complexity of LRMax as reported earlier and the combination of LRMax with MaxQInit outperforms all other methods by providing a tighter upper bound on the optimal Q-value function. This decrease of sample complexity is also observed in the episode-wise display of Figure 4b where the convergence happens more quickly on average for LRMax and even more for MaxQInit. This figure allows to see the three learning stages of LRMax reported in Section 5.

We also ran lifelong RL experiments in the maze grid-world of Figure 5. The tasks consists in reaching the goal cell labeled with a “G” while the initial state of the agent is the central cell, labeled with an “S”. Two walls configurations are possible, yielding two different tasks with probability $\frac{1}{2}$ of being sampled in the lifelong RL setting. The first task corresponds to the case where orange walls are actually walls and green cells are normal white cells where the agent can go. The second task is the converse, where green walls are walls and orange cells are normal white cells. We run 100 episodes of length 15 time steps and sample a total of 30 different tasks. Results can be found in Figure 6. In this experiment, we observe the increase of performance of LRMax as the value of $D_{\max}$ decreases. The three stages behavior of LRMax reported in Section 5 does not appear in this case. We tested the performance of using the online estimation of the local model distances of Proposition 7 in the algorithm referred by LRMax in Figure 6. Once enough tasks have been sampled, the estimate on the model local distance is used with high confidence on its value and refines the upper bound computed analytically in Equation 7. Importantly, this instance of LRMax achieved the best result in this particular environment, demonstrating the usefulness of this result. This method being similar to the MaxQInit estimation of maximum Q-values, we unsurprisingly observe that both algorithms feature a similar performance in the maze environment.

14 Prior $D_{\max}$ use experiment

Consider two MDPs $M, \bar{M} \in \mathcal{M}$. Each time a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ is updated, we compute the local distance upper bound $\hat{D}_{sa}(M \| \bar{M})$ (Equation 7) for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. In this computation, one can leverage the knowledge of $D_{\max}$ to select $\min \left\{ \hat{D}_{sa}(M \| \bar{M}), D_{\max} \right\}$. We show that LRMax relies less and less on $D_{\max}$ as knowledge on the current task increases. For this experiment, we used the two grid-worlds environments displayed in Figures 7a and 7b.



(a) Average discounted return vs. tasks

(b) Average discounted return vs. episodes

Figure 4: Results of the corridor lifelong RL experiment with 95% confidence interval.

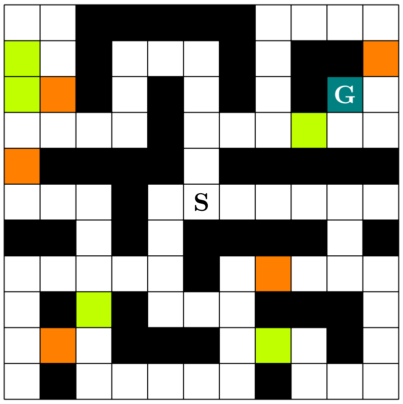


Figure 5: The maze grid-world environment. The walls correspond to the black cells and either the green ones or the orange ones.

The rewards collected with any actions performed in the teal cells of both tasks are defined as:

$$
\begin{aligned}
R_a^s &= \exp\left(- \frac{(s_x - g_x)^2 + (s_y - g_y)^2}{2\sigma^2} \right) , \\
\forall s &= (s_x, s_y) \in \mathcal{S}, a \in \mathcal{A},
\end{aligned}
$$

where (s_x, s_y) are the coordinates of the current state, (g_x, g_y) the coordinate of the goal cell labelled with a G and σ is a span parameter equal to 1 in the first environment and 1.5 in the second environment. The agent starts at the cell labelled with the S letter. Black cells represent unreachable cells (walls). We run LRMax twice on the two different maze grid-worlds and record for each model update the proportion of times $D_{\max}$ is smaller than $\hat{D}_{sa}(M \| \bar{M})$ in Figure 8 via the % use of $D_{\max}$.

With maximum value $D_{\max} = 19$, $\hat{D}_{sa}(M \| \bar{M})$ is systematically lesser than $D_{\max}$, resulting in 0% use. Conversely, with minimum value $D_{\max} = 0$, the use expectedly increases to 100%. The in-between value of $D_{\max} = 10$ displays a linear decay of the use. This suggests that, at each update, $\hat{D}_{sa}(M \| \bar{M}) \leq D_{\max}$ is only true for one more unique s, a pair, resulting in a constant decay of the use. With fewer prior ($D_{\max} = 15$ or 17), updating one single s, a pair allows $\hat{D}_{sa}(M \| \bar{M})$ to drop under $D_{\max}$ for more than one pair, resulting in less use of the prior knowledge. The conclusion of this experiment if that $D_{\max}$ is only useful at the beginning of the exploration, while LRMax relies more on its own bound $\hat{D}_{sa}(M \| \bar{M})$ when partial knowledge of the task has been acquired.

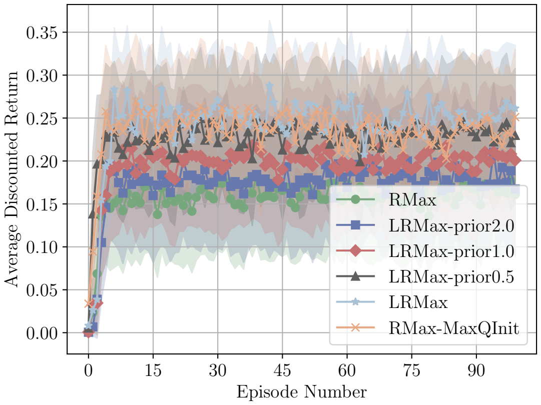
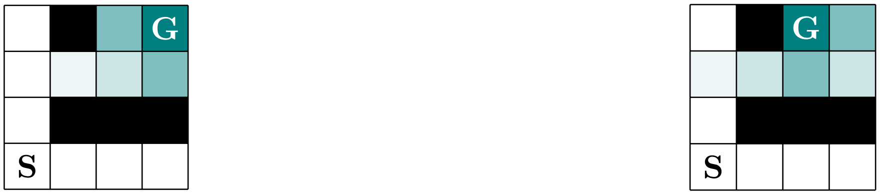


Figure 6: Averaged discounted return over tasks for the maze grid-world lifelong RL experiment.



(a) 4 times 4 heat-map grid-world. Slipping probability is 10%.

(b) 4 times 4 heat-map grid-world. Slipping probability is 5%.

Figure 7: The two grid-worlds of the prior use experiment.

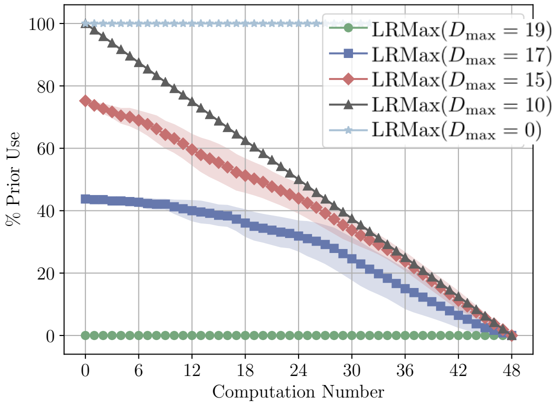


Figure 8: Proportion of times where $D_{\max} \leq \hat{D}_{sa}(M \| \bar{M})$, *i.e.*, use of the prior, vs computation of the Lipschitz bound. Each curve is displayed with 95% confidence intervals.

Villani, C. 2008. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media.

Wilson, A.; Fern, A.; Ray, S.; and Tadepalli, P. 2007. Multi-Task Reinforcement Learning: A Hierarchical Bayesian Approach. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, 1015–1022.

Task	Number of experiment repetitions	Number of sampled tasks	Number of episodes	Maximum length of episodes	Total number of collected transition samples (s, a, r, s')
“Tight” task of Figures 3a 3b and 3c	10	15	2000	10	3,000,000
“Tight” task of Figure 3d	100	2	2000	10	4,000,000
Corridor task Section 13	1	20	20	11	4400
Maze task Section 13	1	30	100	15	45000
Heat-map Section 14	100	2	100	30	600,000

Table 1: Summary of the number of experiment repetition, number of sampled tasks, number of episodes, maximum length of episodes and upper bounds on the number of collected samples.

15 Discussion on RMax precision parameters ϵ, δ, n_{known}

We used $n_{known} = 10$, $\delta = 0.05$ and $\epsilon = 0.01$. Theoretically, n_{known} should be a lot larger ($\approx 10^5$) in order to reach an accuracy $\epsilon = 0.01$ according to Strehl, Li, and Littman (2009). However, it is common practice to assume such small values of n_{known} are sufficient to reach an acceptable model accuracy ϵ. Interestingly, empirical validation did not confirm this assumption for any RMax-based algorithm. We keep these values nonetheless for the sake of comparability between algorithms and consistency with the literature. Despite such absence of accuracy guarantees, RMax-based algorithms still perform surprisingly well and are robust to model estimation uncertainties.

16 Information about the Machine Learning reproducibility checklist

For the experiments run in Section 5, the computing infrastructure used was a laptop using a single 64-bit CPU (model: Intel(R) Core(TM) i7-4810MQ CPU @ 2.80GHz). The collected samples sizes and number of evaluation runs for each experiment is summarized in Table 1.

The displayed confidence intervals for any curve presented in the paper is the 95% confidence interval (Neyman 1937) on the displayed mean. No data were excluded neither pre-computed. Hyper-parameters were determined to our appreciation, they may be sub-optimal but we found the results convincing enough to display interesting behaviors.

References

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research* 47: 253–279.

Bellman, R. 1957. *Dynamic Programming*. Princeton, USA: Princeton University Press.

Ferns, N.; Panangaden, P.; and Precup, D. 2004. Metrics for Finite Markov Decision Processes. In *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence (UAI 2004)*, 162–169. AUAI Press.

Neyman, J. 1937. X—outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236(767): 333–380.

Puterman, M. L. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Song, J.; Gao, Y.; Wang, H.; and An, B. 2016. Measuring the Distance Between Finite Markov Decision Processes. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*, 468–476.

Strehl, A. L.; Li, L.; and Littman, M. L. 2009. Reinforcement Learning in Finite MDPs: PAC Analysis. *Journal of Machine Learning Research* 10(Nov): 2413–2444.

Taylor, M. E.; and Stone, P. 2009. Transfer Learning for Reinforcement Learning Domains: A Survey. *Journal of Machine Learning Research* 10(Jul): 1633–1685.

Watkins, C. J. C. H.; and Dayan, P. 1992. Q-learning. *Machine Learning* 8(3-4): 279–292.