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.

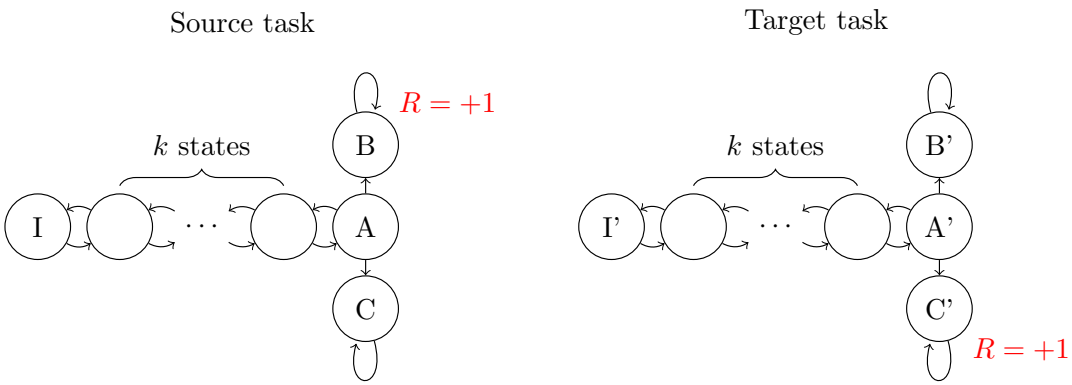


Figure 1: The T-shaped MDP transfer task.

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 $\min\left\{d_{\bar{M}}^{M}, d_{M}^{\bar{M}}\right\}$ 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 $d \in \mathcal{F}(\mathcal{S} \times \mathcal{A}, \mathbb{R})$ to

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

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

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

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 $(\mathcal{F}(\mathcal{S} \times \mathcal{A}, \mathbb{R}), \|\cdot\|_\infty)$. 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 $(Q^n_M)_{n \in \mathbb{N}}$ 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 $(Q^n_M)_{n \in \mathbb{N}}$ 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 $s \in \mathcal{S}$,*

$$\left|V_M^\pi(s) - V_{\bar{M}}^\pi(s)\right| \leq \Delta_s^\pi(M, \bar{M})$$

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

$$d_s = \sum_{a\in\mathcal{A}} \pi(a \mid s) \left(D_{sa}^{\gamma V_M^\pi}(M, \bar{M}) + \gamma \sum_{s'\in\mathcal{S}} T_{ss'}^a d_{s'} \right) .$$

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 $s \in \mathcal{S}$:*

$$d_s = \sum_{a\in\mathcal{A}} \pi(a \mid s) \left(D_{sa}^{\gamma V_M^\pi}(M, \bar{M}) + \gamma \sum_{s'\in\mathcal{S}} T_{ss'}^a d_{s'} \right) .$$

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

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

$$\begin{array}{rccl} Ld: & \mathcal{S} & \to & \mathbb{R} \\ & s & \mapsto & \sum_{a\in\mathcal{A}} \pi(a \mid s) \left(D_{sa}^{\gamma V_M^\pi}(M, \bar{M}) + \gamma \sum_{s'\in\mathcal{S}} T_{ss'}^a d_{s'} \right) \end{array} .$$

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

$$\begin{aligned} Lf_s - Lg_s &= \gamma \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} T_{ss'}^a \left(f_{s'} - g_{s'}\right) \\ &\leq \gamma \left\|f - g\right\|_\infty . \end{aligned}$$

Hence we have that $\|Lf - Lg\|_\infty \leq \gamma \|f - g\|_\infty$. Since $\gamma < 1$, L is a contraction mapping in the metric space $(\mathcal{F}(\mathcal{S}, \mathbb{R}), \|\cdot\|_\infty)$. 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 $(V_M^{\pi,n})_{n\in\mathbb{N}}$ of the value function of any MDP $M \in \mathcal{M}$ is defined for all $s \in \mathcal{S}$ by:

$$\begin{aligned} V_M^{\pi,0}(s) &= 0 , \\ V_M^{\pi,n+1}(s) &= \sum_{a\in\mathcal{A}} \pi(a \mid s) \left(R_s^a + \gamma \sum_{s'\in\mathcal{S}} T_{ss'}^a V_M^{\pi,n}(s') \right) \end{aligned}$$

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

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

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

$$\left|V_M^\pi(s) - V_{\bar{M}}^\pi(s)\right| \leq d_s^\pi(M\|\bar{M})\,.$$

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

$$\left|V_M^\pi(s) - V_{\bar{M}}^\pi(s)\right| \leq \min\left\{d_s^\pi(M\|\bar{M}), d_s^\pi(\bar{M}\|M)\right\}\,,$$

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 $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

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

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

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

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 $d \in \mathcal{F}(\mathcal{S}\times\mathcal{A},\mathbb{R})$ is a fixed-point equation admitting a unique solution for any $(s,a) \in \mathcal{S}\times\mathcal{A}$:*

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

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

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

$$\begin{array}{rccl} Ld: & \mathcal{S}\times\mathcal{A} & \to & \mathbb{R} \\ & (s,a) & \mapsto & D_{sa}^{\gamma V_{\bar{M}}^\pi}(M,\bar{M}) + \gamma \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} T_{ss'}^a \pi(a' \mid s') d_{s',a'}\,. \end{array}$$

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

$$\begin{aligned} Lf_{sa} - Lg_{sa} &= \gamma \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} T_{ss'}^a \pi(a' \mid s') \left(Lf_{s'a'} - Lg_{s'a'}\right) \\ &\leq \gamma \left\|f - g\right\|_\infty\,. \end{aligned}$$

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 $\left(\mathcal{F}(\mathcal{S}\times\mathcal{A},\mathbb{R}), \left\|\cdot\right\|_\infty\right)$. 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 11. Consider a stochastic stationary policy π. The value iteration sequence of iterates $\left(Q_M^{\pi,n}\right)_{n\in\mathbb{N}}$ 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:

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

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

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

Importantly, notice that the probabilistic event of Inequality 13 is the intersection of at most $4SA$ individual events of estimating either R^a_s, $T^a_{ss'}$, $\bar{R}^a_s$ or $\bar{T}^a_{ss'}$ 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:

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

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

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

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

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

Replacing the Inequalities 15 and 16 in Equation 14 yields

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

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 $\bar{T}^a_{s\cdot}$ and $\bar{R}^a_s$. 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:

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

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

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

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 $V_{\max}=\frac{1}{1-\gamma}$ 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.*

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

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

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

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

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

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

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

In both cases, $\left\| Lf - Lg \right\|_\infty \leq \gamma(1+\epsilon) \left\| f - g \right\|_\infty$. 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,

$$\hat{d}^0_{sa}(M \| \bar{M}) = 0,$$

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

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

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

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 $1 - \delta$,

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

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^2A. Overall, the cost of all the DP computations in a complete run of LRMax is

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

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

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

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:

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

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 $[0.8, 1.0]$.

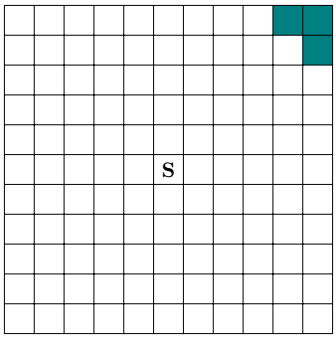


Figure 2: The tight grid-world environment.

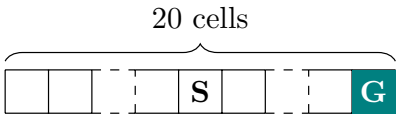


Figure 3: The corridor grid-world environment.

13 Additional lifelong RL experiments

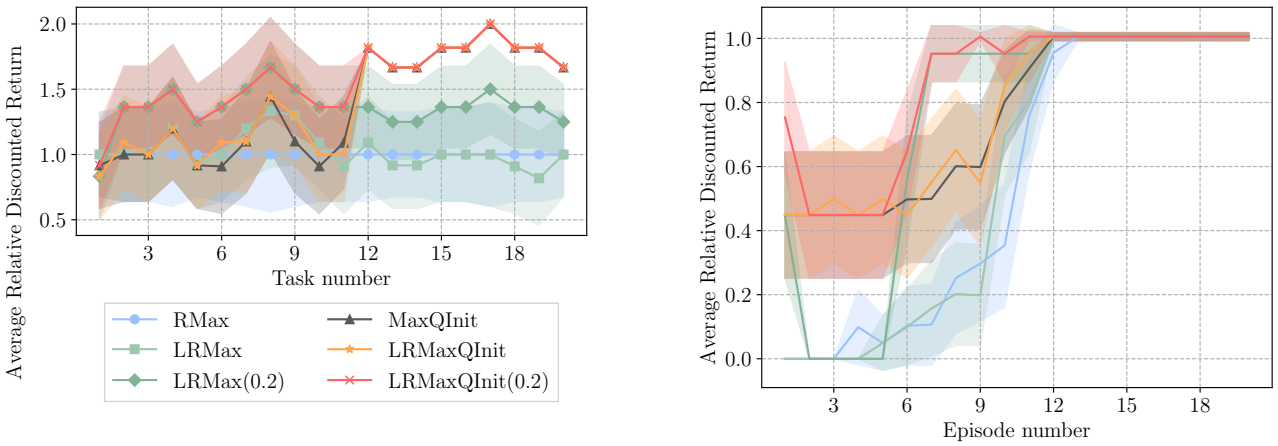
We ran additional experiments on the corridor grid-world environment represented in Figure 3. The initial state of the agent is the central cell labeled with the letter "S". The actions are $\{$left, right$\}$ and the goal is to reach the cell labeled with the letter "G" on the extreme right. A reward $R > 0$ 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.

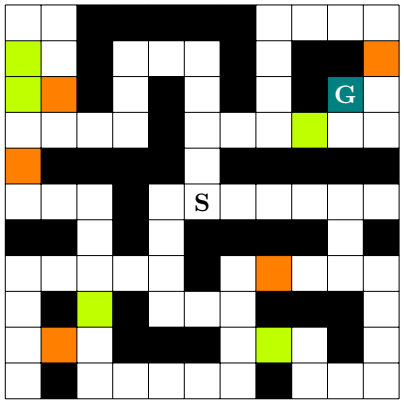


Figure 5: The maze grid-world environment. The walls correspond to the black cells and either the green ones or the orange ones.

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

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

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.

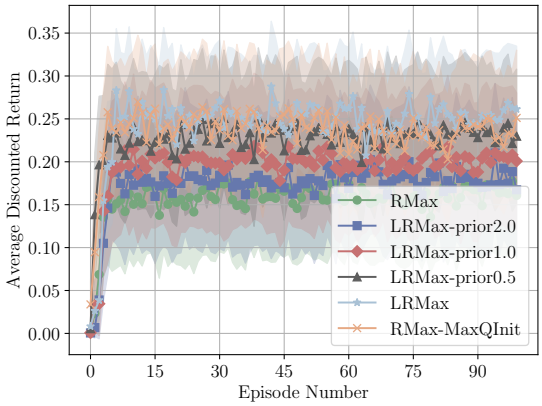
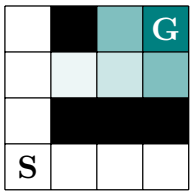
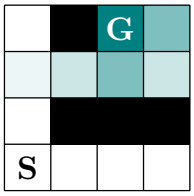


Figure 6: Averaged discounted return over tasks for the maze grid-world lifelong RL experiment.



(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.

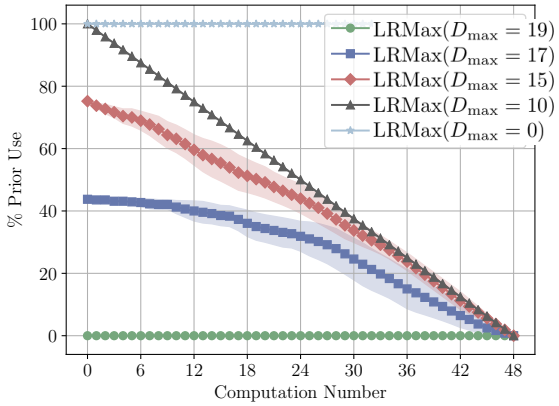


Figure 8: Proportion of times where $D_{\max} \leq \hat{D}_{sa}(M \| \bar{M})$, *i.e.*, use of the prior, vs computation of the Lipschitz bound. Each curve is displayed with 95% confidence intervals.

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.