Lipschitz Lifelong Reinforcement Learning Appendix

Anonymous Author(s)
Affiliation
Address
email

1 Example of negative transfer in Song et al. [2016]

In their paper, Song et al. [2016] propose two transfer methods based on the metric between MDPs they introduce, stemming from the bi-simulation metric introduced by Ferns et al. [2004]. The intuition is that, for a new target task, the value function of the closest source task in terms of that metric is used as an initialisation. However, if no similar source task is available, using the closest task's value function as an initialization can lead to negative transfer. We here understand negative transfer as the fact that it prevents a learning algorithm to converge to the optimal policy while interacting with a new task. We make the hypothesis that the learning algorithm acts greedily w.r.t. the current Q-value function. This is for example the behaviour of the R-MAX algorithm [Brafman and Tennenholtz, 2002]. We now illustrate a negative transfer case with an example. Let us consider the 2-states MDP of Figure 1. We assume that the transitions are deterministic and the initial state is

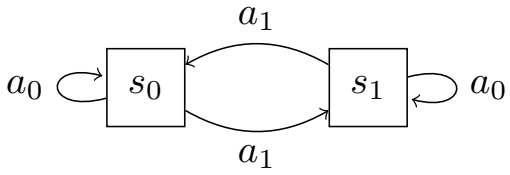

Figure 1: 2-states MDP

always s_0. In the first MDP $M_1 \in \mathcal{M}$, the reward is 0 everywhere except for $R^{a_0}_{s_0} = 1$. In the second MDP $M_2 \in \mathcal{M}$, the reward is 0 everywhere except for $R^{a_1}_{s_1} = 1$. With a discount factor $\gamma = 0.9$, the value functions and Q-functions of both MDPs are summarized in Table 2 Using the weighted

	$V^*_{M_1}(\cdot)$	$Q^*_{M_1}(\cdot, a_0)$	$Q^*_{M_1}(\cdot, a_1)$	$V^*_{M_2}(\cdot)$	$Q^*_{M_2}(\cdot, a_0)$	$Q^*_{M_2}(\cdot, a_1)$
s_0	10	10	8.1	4.74	4.26	4.74
s_1	9	8.1	9	5.26	4.74	5.26

Figure 2: Value functions and Q-functions of MDPs M_1 and M_2

transfer technique from M_1 to M_2 proposed by Song et al. [2016] (Definition 4.1), the Q-function described below is used as an initialization for the exploration of M_2.

$$Q^{\text{transfer}}_{M_2}(s_0, a_0) = 2.03$$
$$Q^{\text{transfer}}_{M_2}(s_0, a_1) = 2.25$$
$$Q^{\text{transfer}}_{M_2}(s_1, a_0) = 2.5$$
$$Q^{\text{transfer}}_{M_2}(s_1, a_1) = 2.03$$

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

First, $Q_{M_2}^{\text{transfer}}$ does not respect the principle of "optimism under the face of uncertainty" that often results in sound and efficient exploration [Strehl et al., 2009, Brafman and Tennenholtz, 2002, Sutton et al., 1998]. Further, a greedy policy w.r.t. $Q_{M_2}^{\text{transfer}}$ would never discover the state-action pair s_1, a_1 in M_2 which is the maximum-reward pair. Instead, the agent would go from s_0 to s_1 and perform self-loops thereafter.

As a conclusion, this negative transfer example motivates the need for distance between MDPs not only to account for the best-source task to use for transfer but also to discourage the transfer when the distance is too high. The approach we develop in this paper used the distance to build optimistic upper-bounds on the Q-function. Those upper-bounds are simply of no use when the distance is too high which is equivalent as avoiding transfer.

2 Discussion on the use of the term "local MDPs distance"

In Theorem 1, we used the term *local MDP distance* to refer to $d_M^{\bar{M}}(s,a)$, defined with the following fixed point equation:

$$d_M^{\bar{M}}(s,a) = D_{\gamma V_{\bar{M}}^*}\left(\langle R,T\rangle, \langle \bar{R},\bar{T}\rangle\right)(s,a) + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \max_{a' \in \mathcal{A}} d_M^{\bar{M}}(s',a')$$

This is not a distance between MDPs in the mathematical sense of the term. The first reason for that is that it is a local function, hence it does not take into account the whole state-action space. Similarly to an $\mathcal{L}_p$-norm, summation over $\mathcal{S} \times \mathcal{A}$ would result in a pseudo-metric. Besides, the pseudo-metric used in Corollary 1 is a mathematical pseudo-metric and is an $\mathcal{L}_\infty$-norm on the whole $\left[D_{\gamma V_{\bar{M}}^*}\left(\langle R,T\rangle, \langle \bar{R},\bar{T}\rangle\right)(s,a)\right]_{s,a}$ vector. We chose to use the terminology *local MDP distance* for $d_M^{\bar{M}}(s,a)$ because it is intuitive and we found other terms like *similarity measure* more confusing.

3 Proof of Theorem 1

Statement: ***Local Lipschitz continuity.*** *For any two MDPs M and $\bar{M}$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$|Q_M^*(s,a) - Q_{\bar{M}}^*(s,a)| \leq d_M^{\bar{M}}(s,a)$$

where $d_M^{\bar{M}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined with the following fixed-point equation:

$$d_M^{\bar{M}}(s,a) = D_{\gamma V_{\bar{M}}^*}\left(\langle R,T\rangle, \langle \bar{R},\bar{T}\rangle\right)(s,a) + \gamma \sum_{s'} T_{ss'}^a \max_{a'} d_M^{\bar{M}}(s',a')$$

Proof. From Dynamic Programming [Bellman, 1957], the optimal state-action value function can be expressed as the limit of a sequence of iterates:

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

It can easily be proven that this sequence is increasing monotonically, i.e. for all $N \geq n$, $Q_M^N(s,a) \geq Q_M^n(s,a), \forall s,a \in \mathcal{S} \times \mathcal{A}$. Notably, by passage to the limit, $Q_M^*(s,a) \geq Q_M^n(s,a), \forall n,s,a \in \mathbb{N} \times \mathcal{S} \times \mathcal{A}$, and by definition of the state value function $V_M^*(s) \geq Q_M^n(s,a), \forall n,s,a \in \mathbb{N} \times \mathcal{S} \times \mathcal{A}$. This result will be used in the remaining of the reasoning. The proof is by induction. For rank $n = 0$,

the result is obvious. Let us consider rank $n+1$, for $s, a \in \mathcal{S} \times \mathcal{A}$:

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

Which, by remarking that the fixed point equation $d_{\bar{M}}^{\bar{M}}(s,a) = D_{\gamma V_{\bar{M}}^*}\left(\langle R, T\rangle, \langle \bar{R}, \bar{T}\rangle\right)(s,a) + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \max_{a'} d_{\bar{M}}^{\bar{M}}(s',a')$ as a unique solution using the Banach fixed-point theorem, concludes the proof by induction. □

4 Similar results to Theorem 1

Similar results to Theorem 1 can be derived with a similar proof as in Section 3. The first result is for the value function and is stated below.

Theorem 3. ***Local Lipschitz continuity of the value function.*** *For any two MDPs M and $\bar{M}$, for all $s \in \mathcal{S}$,*

$$|V_M^*(s) - V_{\bar{M}}^*(s)| \leq \max_{a \in \mathcal{A}} d_{\bar{M}}^{\bar{M}}(s,a)$$

where ***the local MDP distance*** *$d_{\bar{M}}^{\bar{M}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined with the following fixed-point equation:*

$$d_{\bar{M}}^{\bar{M}}(s,a) = D_{\gamma V_{\bar{M}}^*}\left(\langle R, T\rangle, \langle \bar{R}, \bar{T}\rangle\right)(s,a) + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \max_{a' \in \mathcal{A}} d_{\bar{M}}^{\bar{M}}(s',a')$$

Another result can be derived for any policy π common th both MDPs. For the sake of generality, we here state the result for any stochastic policy mapping states to distributions over actions. A deterministic policy is a stochastic policy choosing the selected action with probability 1 and the others with probability 0.

Theorem 4. ***Local Lipschitz continuity of the value and Q-value functions for any policy.*** *For any two MDPs M and $\bar{M}$, for a stochastic policy π, for all $s, a \in \mathcal{S} \times \mathcal{A}$,*

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

where $d_{M,\bar{M}}^\pi(s)$ is defined with the following fixed-point equation:

$$d_{M,\bar{M}}^\pi(s) = \mathbb{E}_{a \sim \pi}\left[D_{\gamma V_{\bar{M}}^*}\left(\langle R, T\rangle, \langle \bar{R}, \bar{T}\rangle\right)(s,a) + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a d_{M,\bar{M}}^\pi(s') \right]$$

5 Proof of Corollary 1

Statement: ***Global continuity.*** *For any two MDPs M and $\bar{M}$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$|Q_M^*(s,a) - Q_{\bar{M}}^*(s,a)| \leq d_M^{\bar{M}} := \frac{1}{1-\gamma} \max_{s,a \in \mathcal{S} \times \mathcal{A}} \left[D_{\gamma V_{\bar{M}}^*}\left(\langle R, T\rangle, \langle \bar{R}, \bar{T}\rangle\right)(s,a) \right]$$

Proof. The proof is by induction, reasoning on the value iteration iterates introduced in Section 3. The result is obvious at rank $n = 0$, let us consider rans $n + 1$:

$$\begin{aligned}
\left|Q_M^{n+1}(s,a) - Q_{\bar{M}}^{n+1}(s,a)\right| &\leq D_{\gamma V_{\bar{M}}^*}\left(\langle R,T\rangle, \langle \bar{R},\bar{T}\rangle\right)(s,a) \\
&\quad + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \max_{a' \in \mathcal{A}} \left|Q_M^n(s',a') - Q_{\bar{M}}^n(s',a')\right| \\
&\leq \max_{s,a \in \mathcal{S}\times\mathcal{A}} \left[D_{\gamma V_{\bar{M}}^*}\left(\langle R,T\rangle, \langle \bar{R},\bar{T}\rangle\right)(s,a)\right] \\
&\quad + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \frac{1}{1-\gamma} \max_{s,a \in \mathcal{S}\times\mathcal{A}} \left[D_{\gamma V_{\bar{M}}^*}\left(\langle R,T\rangle, \langle \bar{R},\bar{T}\rangle\right)(s,a)\right] \\
&\leq \max_{s,a \in \mathcal{S}\times\mathcal{A}} \left[D_{\gamma V_{\bar{M}}^*}\left(\langle R,T\rangle, \langle \bar{R},\bar{T}\rangle\right)(s,a)\right] \left(1 + \frac{\gamma}{1-\gamma}\right) \\
&\leq d_M^{\bar{M}}
\end{aligned}$$

□

6 Proof of Property 1

Statement: *Given a partially known MDP $M = \langle \mathcal{S}, \mathcal{A}, R, T\rangle$, K the set of known state-action pairs and a set of Lipschitz bounds $\left\{U_{\bar{M}_1}, U_{\bar{M}_2}, \cdots\right\}$, the **total upper-bound** U defined below is an upper-bound on Q_M^* for all $s, a \in \mathcal{S} \times \mathcal{A}$.*

$$U(s,a) \triangleq \begin{cases} R_s^a + \gamma \sum_{s' \in \mathcal{S}} T_{ss'}^a \max_{a'} U(s',a') & \text{if } (s,a) \in K \\ \min\left[\frac{1}{1-\gamma}, U_{\bar{M}_1}(s,a), U_{\bar{M}_2}(s,a), \cdots\right] & \text{else} \end{cases} \tag{1}$$

Proof. 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 by induction. Let us consider the Dynamic Programming [Bellman, 1957] sequences converging to Q_M^* and U at rank n whose definitions follow:

$$\begin{cases} Q_{M,0}^*(s,a) = 0 \\ Q_{M,n}^*(s,a) = R_s^a + \gamma \sum_{s'} T_{ss'}^a \max_{a'} Q_{M,n-1}^*(s',a') \end{cases}$$

$$\begin{cases} U_0(s,a) = 0 \\ U_n(s,a) = R_s^a + \gamma \sum_{s'} T_{ss'}^a \max_{a'} U_{n-1}(s',a') \end{cases}$$

Obviously, $Q_{M,0}^*(s,a) \leq U_0(s,a)$. Suppose the property true at rank n and consider rank $n+1$:

$$\begin{aligned}
Q_{M,n+1}^*(s,a) - U_{n+1}(s,a) &= \gamma \sum_{s'} T_{ss'}^a \left(\max_{a'} Q_{M,n}^*(s',a') - \max_{a'} U_n(s',a')\right) \\
&\leq \gamma \sum_{s'} T_{ss'}^a \max_{a'} \left(Q_{M,n}^*(s',a') - U_n(s',a')\right) \\
&\leq 0
\end{aligned}$$

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. □

7 Proof of Property 3

Statement: ***Computational complexity.*** *The total computation complexity of Lipschitz R-MAX is*

$$\mathcal{O}\left(B + \frac{S^2A^2(S + \ln(A))(N+1)}{(1-\gamma)} \ln \frac{1}{\epsilon(1-\gamma)}\right)$$

with B the number of time steps, ϵ the precision of the value iteration algorithm and N the memory size.

Proof. We follow the proof of the computational complexity of R-MAX proposed by Strehl et al. [2009]. The cost of Lipschitz R-MAX is constant on most time steps since the action is greedily chosen w.r.t. the upper-bound on the optimal Q-value function which is a lookup table. When updating a new state-action pair (labelling it as a known pair), the algorithm performs N DP computations to update the Lipschitz bounds plus one DP computation to update the total-bound. The cost of one DP computation is given by [Strehl et al., 2009]:

$$\mathcal{O}\left(SA(S+\ln(A))\frac{1}{1-\gamma}\ln\frac{1}{\epsilon(1-\gamma)}\right)$$

The result comes out by remarking that at most SA state-action pairs are updated, each resulting in $(N+1)$ DP computations. □

8 Proof of Theorem 2

Statement: *Consider an algorithm producing ϵ-accurate model estimates for a subset of $\mathcal{S}\times\mathcal{A}$ after interacting with an MDP. For all $s,a\in\mathcal{S}\times\mathcal{A}$, after sampling m MDPs, the probability of successfully estimating $D_{\max}(s,a)$ is:*

$$Pr(\hat{D}_{\max}(s,a)\geq D_{\max}(s,a))\geq 1-2(1-p_{\min})^m+(1-2p_{\min})^m$$

where $p_{\min}=\min_{M\in\mathcal{M}}Pr(M)$ is a lower bound on the sampling probability of an MDP.

Proof. Consider a fixed state-action pair $s,a\in\mathcal{S}\times\mathcal{A}$. In the worst case, there is only one pair of MDPs $M_1^*,M_2^*=\arg\max_{M_1,M_2\in\mathcal{M}^2}\left[D_{\gamma V^*_{M_1}}\left(\langle R_1,T_1\rangle,\langle R_2,T_2\rangle\right)(s,a)\right]$ yielding the maximum distance. Hence, the problem amounts to know the probability of sampling those two MDPs after m samples.

Let us write X the random variable of the number of samples required for sampling either M_1^* or M_2^* for the first time. Given $p_{\min}$, using the geometric distribution we have that:

$$P(X=i)\geq 2(1-2p_{\min})^{i-1}p_{\min}$$

The factor 2 comes from the fact that both MDPs can be samples for the first time for it to be a success. Let us write Y the random variable of the number of samples required for sampling the remaining MDP for the first time. We have the following result using the geometric distribution for the conditional $P(Y=k|X=i)$:

$$\begin{aligned}P(Y=k)&=\sum_{i=1}^{k-1}P(Y=k,X=i)\\&=\sum_{i=1}^{k-1}P(Y=k|X=i)P(X=i)\\&\geq 2\sum_{i=1}^{k-1}(1-p_{\min})^{k-i-1}(1-2p_{\min})^{i-1}p_{\min}^2\end{aligned}\tag{2}$$

$P(Y=k)$ is the probability of first success at step k. For $\hat{D}_{\max}(s,a)$ to estimate $D_{\max}(s,a)$ in m steps, we require that this success occurs any time during the first m steps, so we have:

$$Pr(\hat{D}_{\max}(s,a)\geq D_{\max}(s,a))=\sum_{k=2}^{m}P(Y=k)$$

Using Equation 2, we can deduce our result when remarking that necessarily $p_{\min} \leq 1/2$:

$$
\begin{aligned}
Pr(\hat{D}_{\max}(s,a) \geq D_{\max}(s,a)) &\geq 2p_{\min}^2 \sum_{k=2}^{m} \sum_{i=1}^{k-1} (1-p_{\min})^{k-i-1}(1-2p_{\min})^{i-1} \\
&\geq 2p_{\min}^2 \sum_{k=0}^{m-2} \sum_{i=0}^{k} (1-p_{\min})^{k-i}(1-2p_{\min})^{i} \\
&\geq 2p_{\min}^2 \sum_{k=0}^{m-2} (1-p_{\min})^{k} \sum_{i=0}^{k} \left(\frac{1-2p_{\min}}{1-p_{\min}}\right)^{i} \\
&\geq 2p_{\min}^2 \sum_{k=0}^{m-2} (1-p_{\min})^{k} \frac{1}{p} \left(1 - p_{\min} - \frac{(1-2p_{\min})^{k+1}}{(1-p_{\min})^{k}}\right) \\
&\geq 2p_{\min} \sum_{k=0}^{m-2} \left((1-p_{\min})^{k+1} - (1-2p_{\min})^{k+1}\right) \\
&\geq 2p_{\min}(1-p_{\min}) \frac{1-(1-p_{\min})^{m-1}}{1-(1-p_{\min})} \\
&- 2p_{\min}(1-2p_{\min}) \frac{1-(1-2p_{\min})^{m-1}}{1-(1-2p_{\min})} \\
&\geq 1 - 2(1-p_{\min})^{m} + (1-2p_{\min})^{m}
\end{aligned}
$$

□

9 Environments used in experiments

We here include the figures of the environments used in the experiments of the paper (Section 5). The rewards in the teal cells of Figures 4 and 5 are defined as:

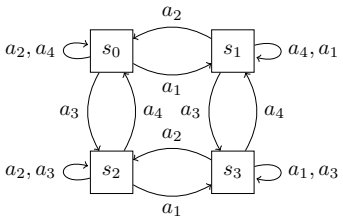

Figure 3: 4-state MDP

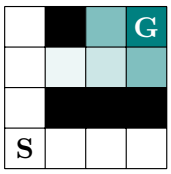

Figure 4: 4 times 4 heat-map grid-world. Slip probability is 10%.

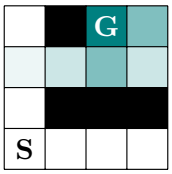

Figure 5: 4 times 4 heat-map grid-world. Slip probability is 5%.

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

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).

10 Informations 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.

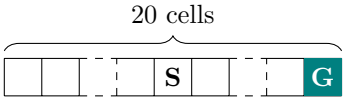

Figure 6: Corridor task. The goal is on the extreme right cell and yields reward 1. Other states yield 0 reward.

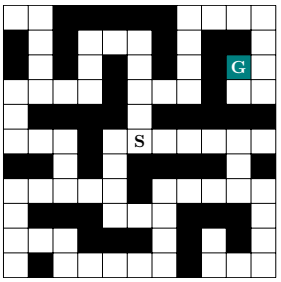

Figure 7: Maze A) task, the slip probability is sampled in $[0, 0.1]$

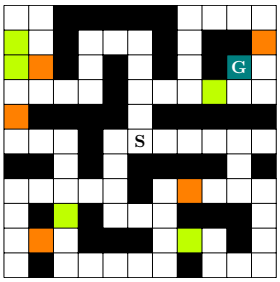

Figure 8: Maze B) task, activated walls are either the green or the orange ones.

Task	Number of experiment repetitions	Number of sampled tasks	Number of episodes	Maximum length of episodes	Upper bound on the number of collected transition samples (s, a, r, s')
4-state Figure 3	10	2	1000	1000	20,000,000
Heat-map Figures 4, 5	100	2	100	30	600,000
Corridor Figure 6	1	20	20	10	4,000
Maze A) Figure 7	1	20	100	1,000	2,000,000
Maze B) Figure 8	1	30	100	1,000	3,000,000

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.

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 behaviours.

References

Richard Bellman. *Dynamic programming*. Princeton, USA: Princeton University Press, 1957.

Ronen I. Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 162–169. AUAI Press, 2004.

Jerzy Neyman. 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, 1937.

Jinhua Song, Yang Gao, Hao Wang, and Bo An. Measuring the distance between finite Markov decision processes. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pages 468–476. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10(Nov):2413–2444, 2009.

Richard S. Sutton, Andrew G. Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.