Lipschitz Lifelong Reinforcement Learning

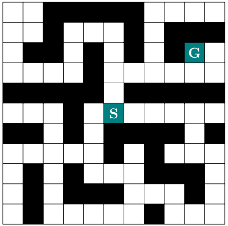
Erwan Lecarpentier [1] David Abel [2] Kavosh Asadi [2]

Yuu Jinnai [2] Emmanuel Rachelson [1] Michael L. Littman [2]

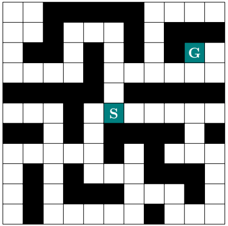
[1] ISAE-SUPAERO, Université de Toulouse, [2] Brown University

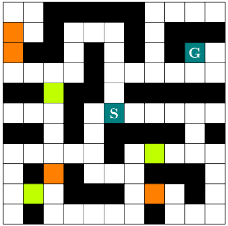
June 21, 2024

The pros and cons of transfer in RL

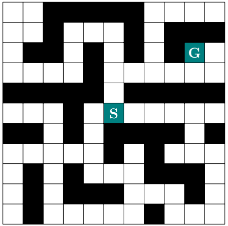


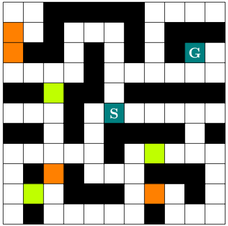
The pros and cons of transfer in RL

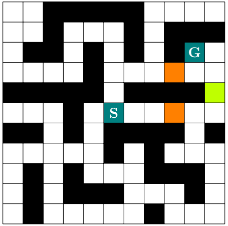




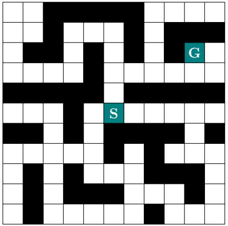
The pros and cons of transfer in RL

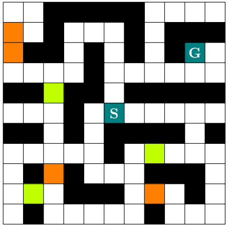


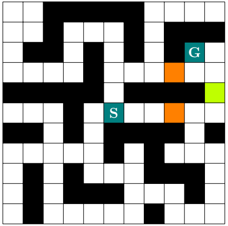


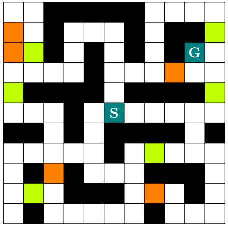


The pros and cons of transfer in RL









Motivations for transfer

- Resilience (*e.g.* sim2real, environment perturbations)
- Lifelong learning

But transfer can be detrimental

How can one guarantee transfer will be beneficial?

This work = attempt at formalizing safe value function transfer + perspectives.

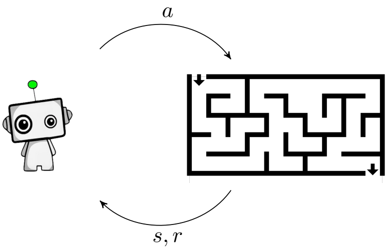
Outline

Spoiler

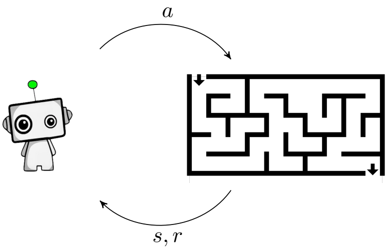
Contributions

- Theoretical study of the Lipschitz **Continuity of V^* and Q^* in the MDP space**;
- Proposal of a **practical, non-negative, transfer method** based on a local distance between MDPs;
- Proposal and study of a **PAC-MDP algorithm** applying this transfer method in the Lifelong RL setting.

Reinforcement Learning

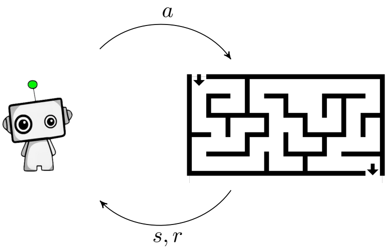


Reinforcement Learning



States	Actions	Transitions	Rewards
$s_t \in \mathcal{S}$	$a_t \in \mathcal{A}$	$T^{a_t}_{s_t s_{t+1}}$	$R^{a_t}_{s_t}$

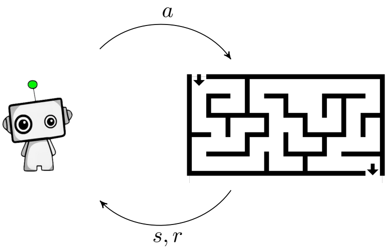
Reinforcement Learning



States	Actions	Transitions	Rewards
$s_t \in \mathcal{S}$	$a_t \in \mathcal{A}$	$T^{a_t}_{s_t s_{t+1}}$	$R^{a_t}_{s_t}$

Policy: what to do in s?

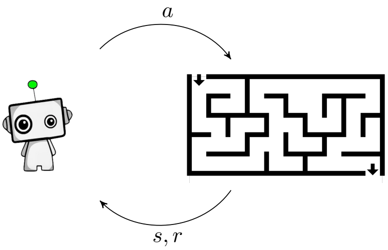
Reinforcement Learning



States	Actions	Transitions	Rewards
$s_t \in \mathcal{S}$	$a_t \in \mathcal{A}$	$T_{s_t s_{t+1}}^{a_t}$	$R_{s_t}^{a_t}$

Policy: $\pi : s \mapsto a$

Reinforcement Learning



States $s_t \in \mathcal{S}$

Actions $a_t \in \mathcal{A}$

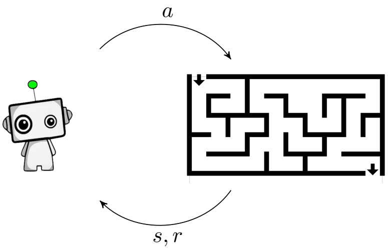
Transitions $T^{a_t}_{s_t s_{t+1}}$

Rewards $R^{a_t}_{s_t}$

Policy: $\pi : s \mapsto a$

Expected value: $V^\pi(s) = \mathbb{E}_{\text{trajectories}}\left[\text{trajectory's return}\right]$

Reinforcement Learning



States $s_t \in \mathcal{S}$

Actions $a_t \in \mathcal{A}$

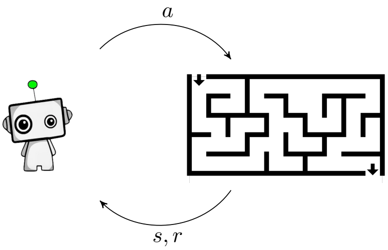
Transitions $T_{s_t s_{t+1}}^{a_t}$

Rewards $R_{s_t}^{a_t}$

Policy: $\pi : s \mapsto a$

Expected value: $V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{s_t}^{a_t} \mid s_0 = s, s_{t+1} \sim T_{s_t s_{t+1}}^{a_t}, a_t = \pi(s_t)\right]$

Reinforcement Learning

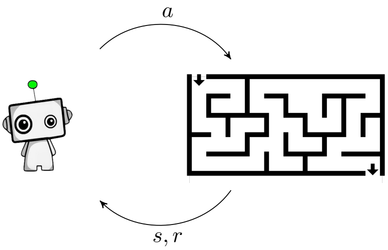


States	Actions	Transitions	Rewards
$s_t \in \mathcal{S}$	$a_t \in \mathcal{A}$	$T^{a_t}_{s_t s_{t+1}}$	$R^{a_t}_{s_t}$

Policy: $\pi : s \mapsto a$

Expected value: $Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^{a_t}_{s_t} \mid s_0 = s, a_0 = a, s_{t+1} \sim T^{a_t}_{s_t s_{t+1}}, a_t = \pi(s_t)\right]$

Reinforcement Learning



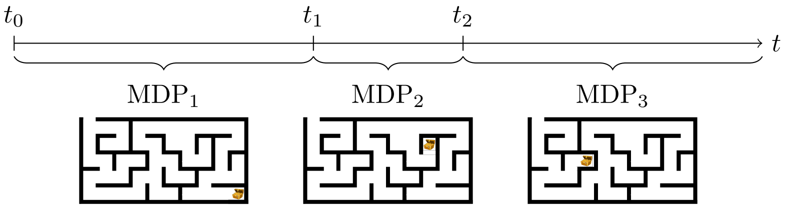
States	Actions	Transitions	Rewards
$s_t \in \mathcal{S}$	$a_t \in \mathcal{A}$	$T^{a_t}_{s_t s_{t+1}}$	$R^{a_t}_{s_t}$

Policy: $\pi : s \mapsto a$

Expected value: $Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^{a_t}_{s_t} \mid s_0 = s, a_0 = a, s_{t+1} \sim T^{a_t}_{s_t s_{t+1}}, a_t = \pi(s_t)\right]$

Optimal value function: $Q^*(s,a) = \max_\pi Q^\pi(s,a)$

Lifelong Reinforcement Learning



Key result

The value function is Lipschitz continuous wrt MDP space.

$$\left|Q^*_M(s,a) - Q^*_{\bar{M}}(s,a)\right| \leq \mathrm{distance}_{\mathcal{M}}(M, \bar{M})$$

Idea

The closer two MDPs, the closer their optimal value functions.

$\downarrow$

Can we do value transfer with that?

Idea

The closer two MDPs, the closer their optimal value functions.

$\downarrow$

Can we do value transfer with that?

$$|Q^*_M(s,a) - Q^*_{\bar{M}}(s,a)| \leq \mathrm{distance}_{\mathcal{M}}(M, \bar{M})$$

$$\downarrow$$

$$Q^*_M(s,a) \leq U(s,a)$$
$$U(s,a) := Q^*_{\bar{M}}(s,a) + \mathrm{distance}_{\mathcal{M}}(M, \bar{M})$$

Why is this important?

Good upper bound on Q^*_M
$\Rightarrow$ more efficient exploration
$\Rightarrow$ possibly faster inference of Q^*_M

What can we say about $Q^*_M - Q^*_{\bar{M}}$? (1/5)

Heavy notations inside.
Proceed with caution.

What can we say about $Q^*_M - Q^*_{\bar{M}}$? (2/5)

Q^*_M is the fixed point of the sequence:

$$Q^{n+1}_M(s,a) = R^a_s + \mathbb{E}_{s' \sim T^a_{ss'}} \left[\max_{a' \in \mathcal{A}} Q^n_M(s', a') \right]$$
$$= R^a_s + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s', a')$$

Let's suppose that

$$\left| Q^n_M(s,a) - Q^n_{\bar{M}}(s,a) \right| \leq d_{sa}(M \| \bar{M})$$

What can we say about $Q^*_M - Q^*_{\bar{M}}$? (3/5)

$$\left| Q^{n+1}_M(s,a) - Q^{n+1}_{\bar{M}}(s,a) \right| = \left| R^a_s - \bar{R}^a_s + \gamma \sum_{s' \in \mathcal{S}} \left[T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s',a') - \bar{T}^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_{\bar{M}}(s',a') \right] \right|,$$

What can we say about $Q^*_M - Q^*_{\bar{M}}$? (3/5)

$$\left| Q^{n+1}_M(s,a) - Q^{n+1}_{\bar{M}}(s,a) \right| = \left| R^a_s - \bar{R}^a_s + \gamma \sum_{s' \in \mathcal{S}} \left[T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s',a') - \bar{T}^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_{\bar{M}}(s',a') \right] \right|,$$

$$\ldots$$

$$\leq \left| R^a_s - \bar{R}^a_s \right| + \sum_{s' \in \mathcal{S}} \gamma V^*_{\bar{M}}(s') \left| T^a_{ss'} - \bar{T}^a_{ss'} \right| + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} \left| Q^n_M(s',a') - Q^n_{\bar{M}}(s',a') \right|,$$

What can we say about $Q^*_M - Q^*_{\bar{M}}$? (3/5)

$$\left| Q^{n+1}_M(s,a) - Q^{n+1}_{\bar{M}}(s,a) \right| = \left| R^a_s - \bar{R}^a_s + \gamma \sum_{s' \in \mathcal{S}} \left[T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s',a') - \bar{T}^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_{\bar{M}}(s',a') \right] \right|,$$

$$\dots$$

$$\leq \left| R^a_s - \bar{R}^a_s \right| + \sum_{s' \in \mathcal{S}} \gamma V^*_{\bar{M}}(s') \left| T^a_{ss'} - \bar{T}^a_{ss'} \right| +$$

$$\gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} \left| Q^n_M(s',a') - Q^n_{\bar{M}}(s',a') \right|,$$

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

What can we say about $Q^*_M - Q^*_{\bar{M}}$? (3/5)

$$\left| Q^{n+1}_M(s,a) - Q^{n+1}_{\bar{M}}(s,a) \right| = \left| R^a_s - \bar{R}^a_s + \gamma \sum_{s' \in \mathcal{S}} \left[T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s',a') - \bar{T}^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_{\bar{M}}(s',a') \right] \right|,$$

$$\dots$$

$$\leq \left| R^a_s - \bar{R}^a_s \right| + \sum_{s' \in \mathcal{S}} \gamma V^*_{\bar{M}}(s') \left| T^a_{ss'} - \bar{T}^a_{ss'} \right| +$$

$$\gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} \left| Q^n_M(s',a') - Q^n_{\bar{M}}(s',a') \right|,$$

$$\leq D_{sa}(M \| \bar{M}) + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a'} d_{s'a'}(M \| \bar{M}) \triangleq d_{sa}(M \| \bar{M}).$$

What can we say about $Q^*_M - Q^*_{\bar{M}}$? (3/5)

$$\left|Q^{n+1}_M(s,a) - Q^{n+1}_{\bar{M}}(s,a)\right| = \left|R^a_s - \bar{R}^a_s + \gamma \sum_{s' \in \mathcal{S}} \left[T^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_M(s',a') - \bar{T}^a_{ss'} \max_{a' \in \mathcal{A}} Q^n_{\bar{M}}(s',a') \right] \right|,$$

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

By induction,

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

What can we say about $Q^*_M - Q^*_{\bar{M}}$? (4/5)

Sooooooooo...

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

With

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

And

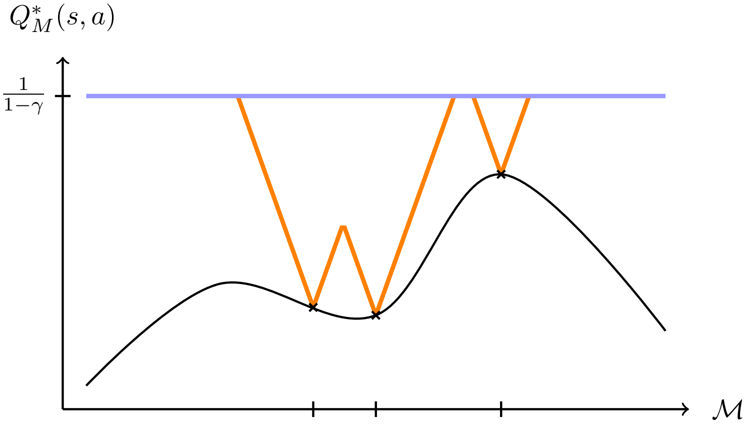
$$D_{sa}(M \| \bar{M}) = \left| R^a_s - \bar{R}^a_s \right| + \sum_{s' \in \mathcal{S}} \gamma V^*_{\bar{M}}(s') \left| T^a_{ss'} - \bar{T}^a_{ss'} \right|$$

$D_{sa}(M \| \bar{M})$: pseudo-metric between M and $\bar{M}$.
$d_{sa}(M \| \bar{M})$: local MDP dissimilarity.

What can we say about $Q^*_M - Q^*_{\bar{M}}$? (5/5)

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

Graphically



Rmax in a nutshell

Rmax (Brafman and Tennenholtz, 2002)

Optimistic model initialization: $\hat{R}_s^a = R_{\max}$, $\hat{T}_{s,s}^a = 1$, then:

- Solve model $\rightarrow Q$.
- Explore greedily wrt Q, store samples.
- When enough samples in (s, a), update $\hat{R}_s^a$ and $\hat{T}_{s,s'}^a$.
- Repeat.

Main intuition: try to disprove optimism where it indicates the most potential.
If Q is an upper-bound on Q^*, then exploring greedily wrt Q will shrink this upper bound.
Notation: $K :=$ set of known state-action pairs.

Rmax in a nutshell

Properties

- Learns a model online.
- Finds an ϵ-optimal policy with high probability in polynomial time (PAC-MDP).
- One of the only algorithms with a guaranteed convergence rate.

But limited to (small) discrete state/action spaces in its original formulation.

Lipschitz Rmax — the idea

In Rmax, Q acts as an admissible heuristic for exploration.

Any tighter upper-bound than $\frac{1}{1-\gamma}$ will improve Rmax's convergence.

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

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

Upper bound on Q^*_M:

$$Q_M(s,a) \triangleq \begin{cases} R^a_s + \gamma \sum_{s' \in \mathcal{S}} T^a_{ss'} \max_{a' \in \mathcal{A}} Q_M(s',a') \text{ if } (s,a) \in K, \\ U(s,a) \text{ otherwise,} \end{cases}$$

$\rightarrow$ Solve by Dynamic Programming.

A computable upper-bound on Q^*_M

So we need to compute $U_{\bar{M}}(s, a)$

$$U_{\bar{M}}(s, a) \triangleq Q^*_{\bar{M}}(s, a) + \min \left\{ d_{sa}(M \| \bar{M}), d_{sa}(\bar{M} \| M) \right\}$$

With

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

And

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

Problem: $d_{sa}(M \| \bar{M})$ can be computed by Dynamic Programming...

... but it requires knowing exactly $V^*_{\bar{M}}$, $T^a_{ss'}$, $\bar{T}^a_{ss'}$, R^a_s, and $\bar{R}^a_s$.

A computable upper-bound $\hat{U}_{\bar{M}}(s, a)$ on $U_{\bar{M}}(s, a)$?

A computable upper-bound on Q^*_M

So we need to compute $U_{\bar{M}}(s, a)$

$$U_{\bar{M}}(s, a) \triangleq Q^*_{\bar{M}}(s, a) + \min \left\{ d_{sa}(M \| \bar{M}), d_{sa}(\bar{M} \| M) \right\}$$

With

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

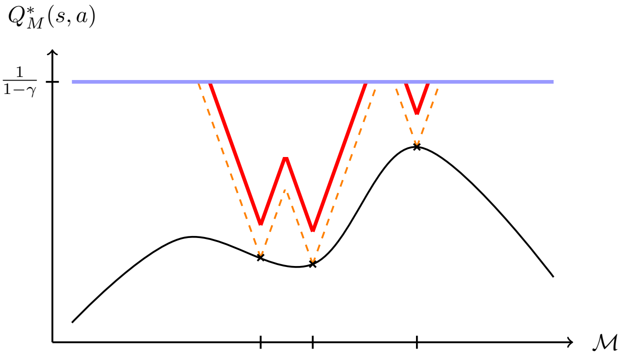
And

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

- Known upper-bound $\rightarrow Q_{\bar{M}}$
- Maximization over the unknown model(s) $\rightarrow \hat{D}_{sa}(M \| \bar{M})$
- Maximize over s' if unknown $\rightarrow \hat{d}_{sa}(M \| \bar{M})$

A computable upper-bound on Q^*_M

$$U_{\bar{M}}(s,a) = Q^*_{\bar{M}}(s,a) + \Delta_{sa}(M,\bar{M})$$
$$\downarrow$$
$$\hat{U}_{\bar{M}}(s,a) = Q_{\bar{M}}(s,a) + \hat{\Delta}_{sa}(M,\bar{M})$$



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**
    $a = \arg\max_{a'} Q(s, a')$
    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 $(\hat{T}^a_{ss'}, \hat{R}^a_s)$
      $Q \leftarrow$ 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 $\hat{D}_{sa}(M \| \bar{M})$, $\hat{D}_{sa}(\bar{M} \| M)$
  Compute $\hat{d}_{sa}(M \| \bar{M})$, $\hat{d}_{sa}(\bar{M} \| M)$
  Compute $\hat{U}_{\bar{M}}$
Compute $\hat{U}$
Compute and return Q

Unfolding the computation

$\hat{D}_{sa}(M \| \bar{M})$	Model distance upper-bound (analytical resolution)
$\to \hat{d}_{sa}(M \| \bar{M})$	Model dissimilarity upper-bound (dynamic programming)
$\to \hat{U}_{\bar{M}}$	Upper-bound on Lipschitz bound $Q_{\bar{M}}(s, a) + \hat{\Delta}_{sa}(M, \bar{M})$
$\to \hat{U}$	Minimum over all upper-bounds
$\to Q$	Upper bound on Q^*_M (dynamic programming)

Remarks

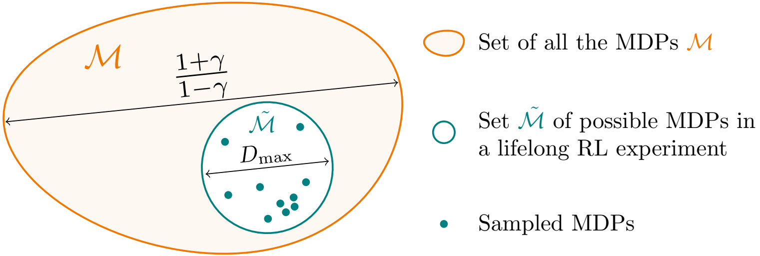
- Shrinking $\hat{D}_{sa}(M \| \bar{M})$ has an influence on $\hat{d}_{sa}(M \| \bar{M})$ in all state-action pairs.
- Smaller $\hat{d}_{sa}(M \| \bar{M})$ induce tighter $\hat{U}$ bounds
- Shrinking $\hat{U}(s, a)$ has an influence on Q in all state-action pairs

Consequence: any information that can help reduce $\hat{D}_{sa}(M \| \bar{M})$ will greatly facilitate value transfer and improve Lipschitz Rmax.

Prior knowledge on model distance

Recall: $\hat{D}_{sa}(M \| \bar{M})$ is an upper-bound on $D_{sa}(M \| \bar{M})$.

How is it computed? Worst case distance between models.
Why? Because models are only partially known.



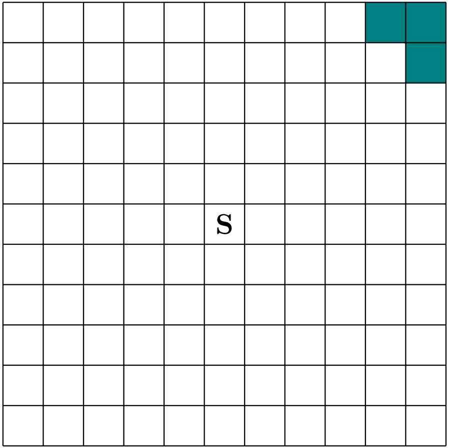
Knowledge of $D_{\max}$ (even a very conservative hypothesis) will strongly tighten $\hat{U}$.

Empirical evaluation

Claims:

- LRmax allows for early performance increase (resilience)
- LRmax is more sample efficient than Rmax
- LRmax avoids negative transfer

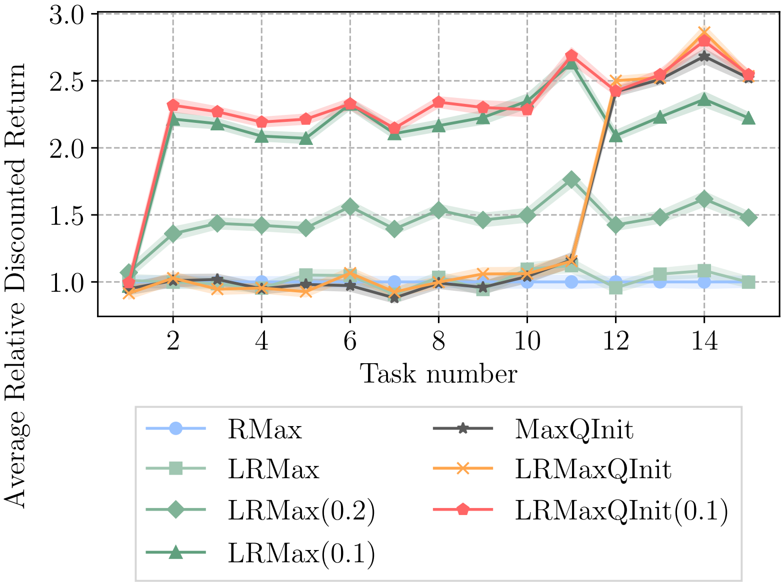
The “tight” environment from (Abel et al., 2018)



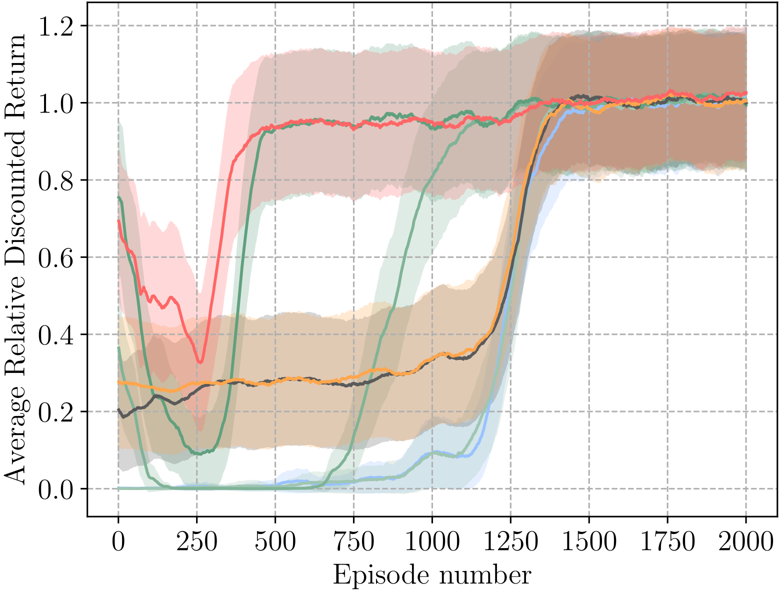
Variations:

- rewards are picked in $[0.8; 1]$
- probability of slipping is picked in $[0; 0.1]$

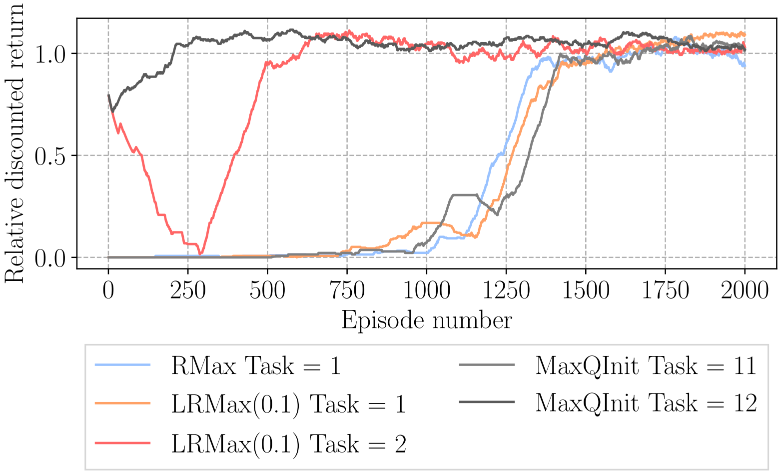
Average discounted return vs. tasks: early transfer among tasks



Average discounted return vs. episodes: faster convergence



Discounted return for specific tasks



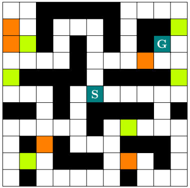
Conclusion

Contributions

- Theoretical study of the Lipschitz **Continuity of V^* and Q^* in the MDP space**;
- Proposal of a **practical, non-negative, transfer method** based on a local distance between MDPs;
- Proposal and study of a **PAC-MDP algorithm** applying this transfer method in the Lifelong RL setting.

Perspectives

- Other algorithms than Rmax?
- Robustness instead of resilience
- Extension to an algorithm that uses value function approximation?



Lecarpentier E, Abel D, Asadi K, Jinnai Y, Rachelson E, Littman M L (AAAI 2021)
Lipschitz Lifelong Reinforcement Learning
`https://arxiv.org/abs/2001.05411`

Thanks for your attention!

Join us at EWRL17!

EWRL

Oct 28-30 2024, Toulouse

THE 17TH EUROPEAN WORKSHOP ON REINFORCEMENT LEARNING

Influence of D_{max}