# Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes


Ronan Fruit[†]

Matteo Pirotta[*]

Alessandro Lazaric[*]

[†] SequeL – INRIA Lille

[*] FAIR – Facebook Paris



NeurIPS 2018, Montreal, December 5th

# Exploration–exploitation in RL with Misspecified State Space

Ronan Fruit[†]

Matteo Pirotta[*]

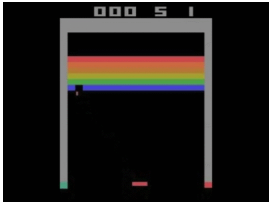
Alessandro Lazaric[*]

[†]SequeL – INRIA Lille
[*]FAIR – Facebook Paris

NeurIPS 2018, Montreal, December 5th

# Misspecified states: Examples

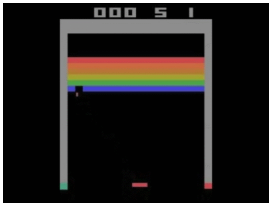
## 1 Breakout [Mnih et al., 2015]



# Misspecified states: Examples

**1** Breakout [Mnih et al., 2015]

Intuitive state space: set of plausible configurations of wall, ball and paddle

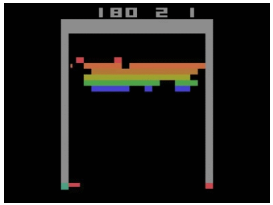


# Misspecified states: Examples

## 1 Breakout [Mnih et al., 2015]

Intuitive state space: set of plausible configurations of wall, ball and paddle


initial state $s_1$

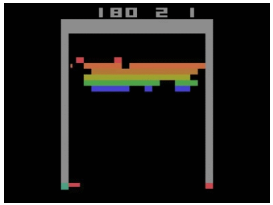



# Misspecified states: Examples

**1** Breakout [Mnih et al., 2015]

Intuitive state space: set of plausible configurations of wall, ball and paddle


initial state $s_1$



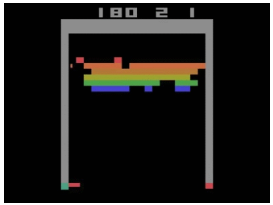

**Plausible state after some time...**

# Misspecified states: Examples

**1** Breakout [Mnih et al., 2015]

Intuitive state space: set of plausible configurations of wall, ball and paddle


initial state $s_1$




**Plausible state after some time...**

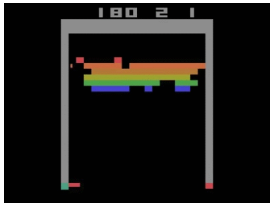

**Non reachable from $s_1$**

# Misspecified states: Examples

## 1 Breakout [Mnih et al., 2015]

Intuitive state space: set of plausible configurations of wall, ball and paddle


initial state $s_1$




**Plausible state after some time...**

**Cannot be observed!**

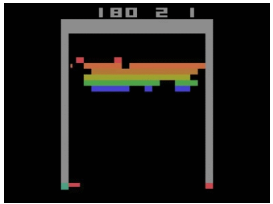

**Non reachable from $s_1$**

# Misspecified states: Examples

## 1 Breakout [Mnih et al., 2015]

Intuitive state space: set of plausible configurations of wall, ball and paddle


initial state $s_1$




**Plausible state after some time...**

**Cannot be observed!**


**Non reachable from $s_1$**

Misspecified state space = $\exists$ states non-observable from initial state
$+$ difficult to exclude **explicitly** from the state space

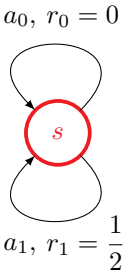
# Why is exploration more challenging with a misspecified state space?

# Why is exploration more challenging with a misspecified state space?

- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**

# Why is exploration more challenging with a misspecified state space?

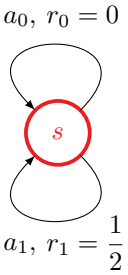
- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**




Example 1 of Ortner [2008]

# Why is exploration more challenging with a misspecified state space?

- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**

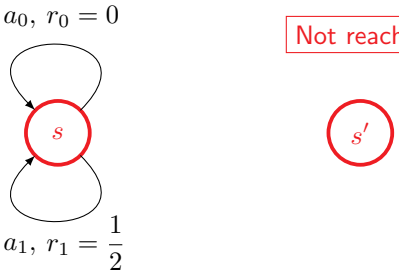



Optimism (UCB, etc.) = Optimal Strategy

Example 1 of Ortner [2008]

# Why is exploration more challenging with a misspecified state space?

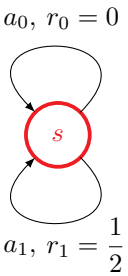
- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**



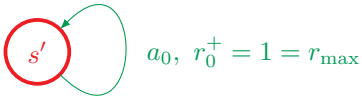

Example 1 of Ortner [2008]

# Why is exploration more challenging with a misspecified state space?

- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**




Not reachable from $s$

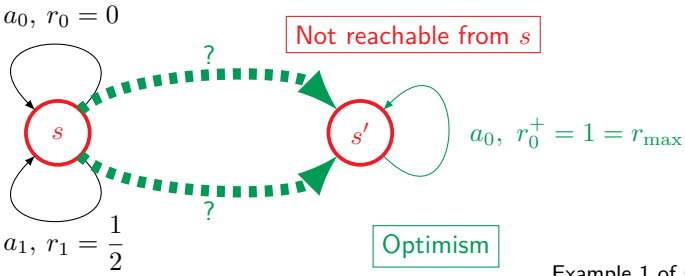



Optimism

Example 1 of Ortner [2008]

# Why is exploration more challenging with a misspecified state space?

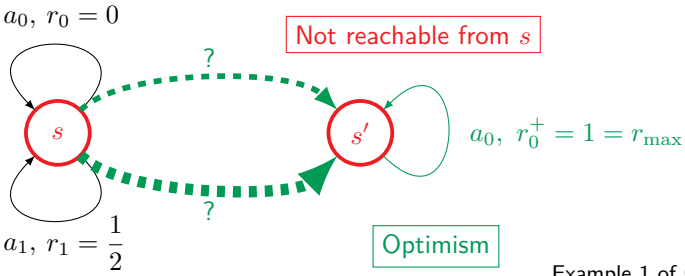
- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**




Example 1 of Ortner [2008]

# Why is exploration more challenging with a misspecified state space?

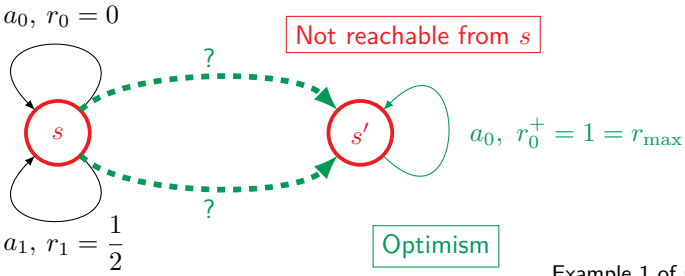
- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**




Example 1 of Ortner [2008]

# Why is exploration more challenging with a misspecified state space?

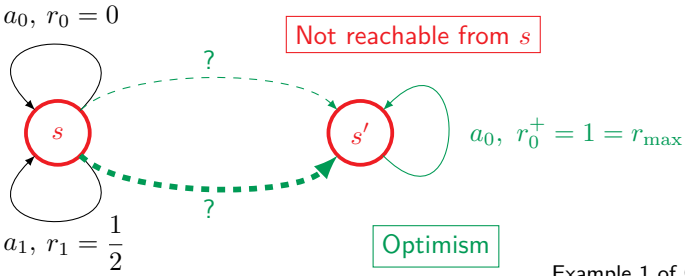
- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**




Example 1 of Ortner [2008]

# Why is exploration more challenging with a misspecified state space?

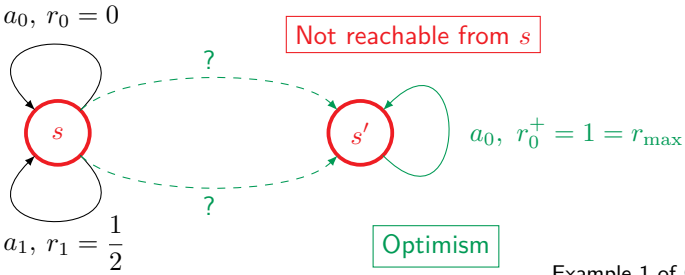
- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**




Example 1 of Ortner [2008]

# Why is exploration more challenging with a misspecified state space?

- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**



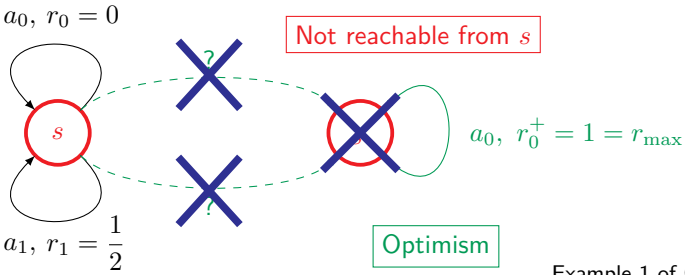

Example 1 of Ortner [2008]

**Problem:** The action played keeps changing: it is $a_0$ half of the time and $a_1$ the other half $\implies$ **linear regret**!

# Why is exploration more challenging with a misspecified state space?

- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**

Why not **ignore** $s'$?

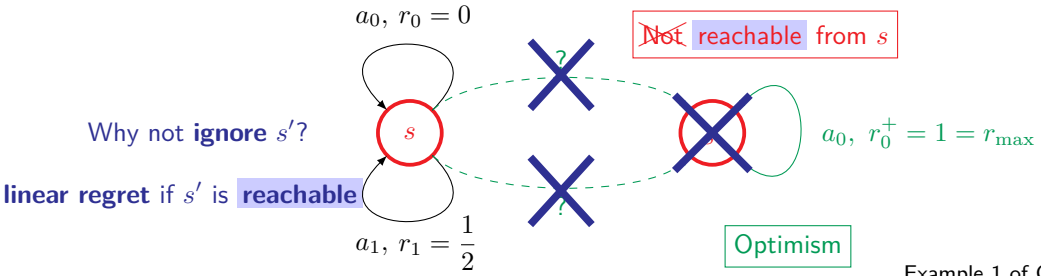



Example 1 of Ortner [2008]

**Problem:** The action played keeps changing: it is $a_0$ half of the time and $a_1$ the other half $\Longrightarrow$ **linear regret**!

# Why is exploration more challenging with a misspecified state space?

- All existing methods known to efficiently balance exploration and exploitation in RL with theoretical guarantees rely on the optimism in the face of uncertainty principle
- All such methods **fail to learn** when the state space is **misspecified**




Example 1 of Ortner [2008]

**Problem:** The action played keeps changing: it is $a_0$ half of the time and $a_1$ the other half $\implies$ **linear regret**!

# Our work

**Regret** of existing methods: $\widetilde{O}\left(D\, S\, \sqrt{AT}\right)$

Diameter — Total number of states

**Misspecified** state space $\iff$ $D = +\infty$ (infinite diameter)

TUCRL: first algorithm able to **adapt** to the reachable part of the MDP

**Regret:** $\widetilde{O}\left(D^{\mathrm{C}}\, S^{\mathrm{C}} \sqrt{AT}\right)$

Reachable diameter — Number of reachable states

Come to see our poster # 161 !