State Abstraction Refinement in Model-Free Reinforcement Learning

SIMPAS Retreat — Rouen

Orso Forghieri, Erwan Le Pennec [1] ;
Emmanuel Hyon [2] ;
Hind Castel [3]; .

[1]CMAP Ecole Polytechnique
[2]LIP6, Université Paris Nanterre
[3]Telecom SudParis

12/06/2025

Introduction

Reinforcement Learning

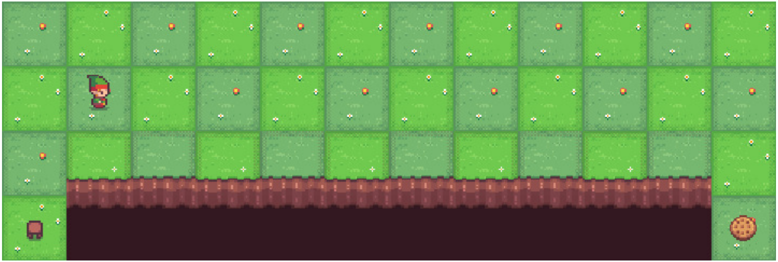

Figure 1: Cliff Walking environment [Towers et al., 2024].

Introduction

Markov Decision Processes

- Observable State s_t, Action a_t, Reward r_t, Next state s_{t+1}
- Optimization problem: $\max_{\pi \in \mathcal{A}^{\mathcal{S}}} \sum_{t \geq 0} \gamma^t r_t$, $\gamma = 0.99$

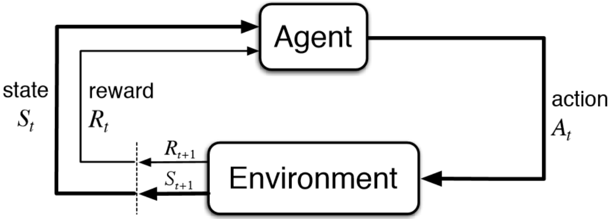

Figure 2: Principle of Reinforcement Learning [Sutton and Barto, 2018].

Introduction

Hierarchical Reinforcement Learning

Why use HRL?

- Solving large RL problems,
- Enhancing explainability and interpretability,
- Ensuring solution quality.

How to implement it?

- **Subgoal discovery** and meta-action learning (temporal abstraction)[1],
- Building reduced representation using **spatial abstraction**, with limited explicit methods[2].

$\rightarrow$ We focus on **state abstraction** discovery in **model-free RL**.

[1] [Pateria et al., 2021, Nachum et al., 2018]

[2] [Abel et al., 2016, Starre et al., 2022]

Outline

Context

Our contribution

In this talk, we present:

- A **hierarchical Q-Learning-based approach** that is:
  1. Based on state abstraction refinement,
  2. Convergence-guaranteed,
  3. Sample efficient.
- A **practical evaluation** on benchmark environments.

Context

Reinforcement Learning Problem

Assuming that we access to samples (s_t, a_t, r_t, s_{t+1}), we aim to compute the optimal action-value function Q^*:

$$Q^*(s,a) = \mathbb{E}_{\pi^*}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s, a_0 = a\right]$$

Q^* is be the solution to the Bellman optimality equation:

$$\begin{aligned} Q^*(s,a) &= \mathbb{E}\left[r_t + \gamma \max_{a'} Q^*(s_{t+1}, a') \mid s_t = s, a_t = a\right] \\ &:= (\mathcal{T}_Q^* Q)(s,a) \end{aligned}$$

Context

Q-Learning Algorithm

Given samples (s_t, a_t, r_t, s_{t+1}), we update Q:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_t(s_t, a_t) \cdot \underbrace{\left[r_t + \gamma \max_{a' \in \mathcal{A}} Q(s_{t+1}, a') - Q(s_t, a_t) \right]}_{\text{TD-Error } \delta_t}$$

where $\sum_{t=1}^{\infty} \alpha_t(s, a) = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2(s, a) < \infty$. Here δ is a sample for $\mathcal{T}_Q^* Q - Q$:

$$\mathbb{E}_{s_{t+1}}[\delta_t | s_t = s, a_t = a] = (\mathcal{T}_Q^* Q - Q)(s, a)$$

→ Convergence of Q-Learning as a Stochastic Approximation method.[3]

[3] [Jaakkola et al., 1993]

Context

State Abstraction

Definition (Abstract MDP [Li et al., 2006])

Given $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, R)$ s.t. $\mathcal{S} = \bigsqcup_k S_k$, the abstract MDP

$$(\mathcal{K}, \mathcal{A}, \tilde{T}, \tilde{R})$$

is defined using

- Abstract state space $\mathcal{K} = \{s_k\}$
- Averaged transition $\tilde{T} = \omega \cdot T \cdot \phi$
- Averaged reward $\tilde{R} = \omega \cdot R$

where

- $\omega \in [0,1]^{K \times \mathcal{S}}$ weights with sum 1 on each region
- $\phi = (\mathbb{1}_{s \in S_k})_{s,k}$

$\rightarrow$ Find aggregation of states without direct access to the abstract environment. Good abstraction $\iff$ Similar states gathered.

Context

Suited abstraction

An adapted abstraction gather similar states. Either with close Q^*-value[4]:

$$s, s' \in s_A \iff \max_{a \in \mathcal{A}} \left| Q^*(s,a) - Q^*(s',a) \right| \leq \varepsilon$$

Or with close Bellman operator update[5]:

$$s, s' \in s_A \iff \max_{a \in \mathcal{A}} \left| \mathcal{T}_Q^* Q(s,a) - \mathcal{T}_Q^* Q(s,a) \right| \leq \varepsilon$$

for Q close to Q_A^*. $\rightarrow$ In model-free, unknowns Q^* and $\mathcal{T}_Q^*$! We sample transitions and TD-errors to approximate $\mathcal{T}_Q^*$.

[4] [Abel et al., 2016]
[5] [Forghieri et al., 2024]

Abstraction Refinement Process

Main algorithm

To solve the original RL problem, we:

1. Start with a trivial representation $\mathcal{K}_0 = \{\mathcal{S}\}$,
2. Iterate the following steps:
   - Compute the optimal Q-value Q_A^* of abstraction $\mathcal{K}_t$ through

$$Q_{A,t+1}(s, a_t) = Q_{A,t}(s, a_t) + \alpha_t(s, a_t) \cdot \delta_t \cdot \frac{\mathbb{1}_{s \in s_{A,t}}}{|s_{A,t}|},$$

   - Store visits $n_{s,a}$ and empirical mean $\hat{\mu}_{s,a}$ of δ_t that occurs on (s, a),
   - Refine abstraction $\mathcal{K}_t$ by separating states such that

$$\max_{a \in \mathcal{A}} |\hat{\mu}_{s,a} - \hat{\mu}_{s',a}| > \varepsilon + f_\theta(n_{s,a}) + f_\theta(n_{s',a'})$$

   where $f_\theta(n) = \sqrt{a \log(\log_c n + 1) + b_\theta} / \sqrt{n}$,
3. Return the last abstraction $\mathcal{K}_t$ and action-value function Q.

Abstraction Refinement Process

Abstraction refinement

Remark

The **refinement condition**

$$\max_{a \in \mathcal{A}} \left| \hat{\mu}_{s,a} - \hat{\mu}_{s',a} \right| > \varepsilon + f_\theta(n_{s,a}) + f_\theta(n_{s',a'})$$

is a proxy w.p. $1 - \theta$ for the condition

$$s, s' \text{ such that } \left| (\mathcal{T}_Q^* Q_A^*)(s, \cdot) - (\mathcal{T}_Q^* Q_A^*)(s', \cdot) \right| \geq \varepsilon$$

where $(\mathcal{T}_Q^* Q)(s, a) = \mathbb{E}\left[r_t + \gamma \max_{a'} Q^*(s_{t+1}, a') \mid s_t = s, a_t = a \right]$.

Theoretical Guarantees

Convergence guarantee

Theorem

With probability 1,

(i) *the previous process converges to a given state abstraction* $\mathcal{K}$ *and its Q-value function* Q_A^*. *This function approximates the optimal action-value function* Q^* *under the condition:*

$$\|Q_A^* - Q^*\|_\infty \leq \frac{\varepsilon}{1-\gamma},$$

(ii) *the abstraction* $\mathcal{K}$ *satisfies the following property:*

$$\forall s_A \in \mathcal{K}, \max_{s,s' \in s_A} \max_{a \in \mathcal{A}} \left| Q^*(s,a) - Q^*(s',a) \right| \leq \frac{2\varepsilon}{1-\gamma}.$$

→ Convergence + adapted resulting representation

Theoretical Guarantees

Visiting adapted abstractions

Theorem

If a visited abstraction gathers similar states

$$\max_{s,s' \in s_A} \max_{a \in \mathcal{A}} \left| (\mathcal{T} Q_A^*)(s,a) - (\mathcal{T} Q_A^*)(s',a) \right| \leq \frac{\varepsilon}{2} \,,$$

and if

$$\| Q_A^* - Q_{A,t_0} \|_\infty \leq \frac{\varepsilon}{2}$$

then, with probability $1-\theta$*, the current abstraction will be the final one:*

$$\forall t \geq t_0, \; \max_{a \in \mathcal{A}} \left| \hat{\mu}_{s,a} - \hat{\mu}_{s',a} \right| \leq \varepsilon + f(n_{s,a}) + f(n_{s',a}) \,.$$

$\rightarrow$ If we visit an adapted abstraction, we keep it!

Application

Cliff Walking Environment

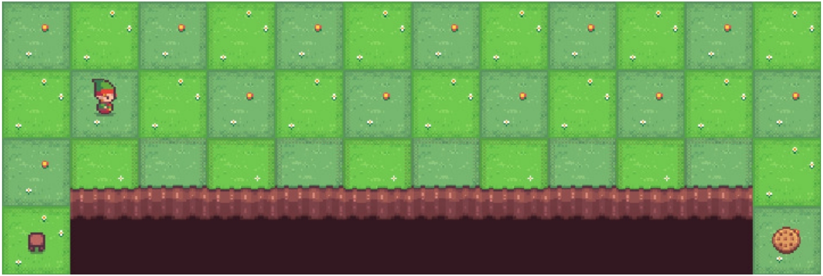

Figure 3: Cliff Walking environment [Towers et al., 2024].

Application

Cliff walking optimal value function

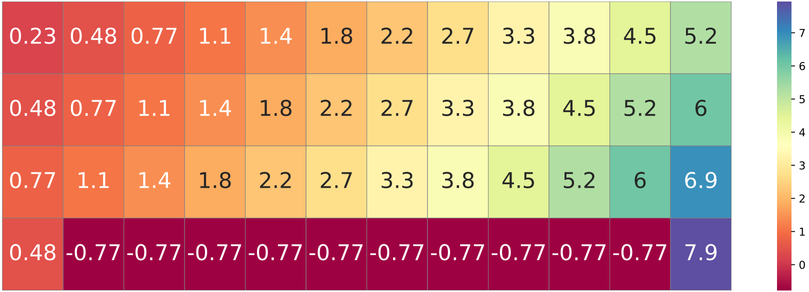

Figure 4: Optimal value function V^* for the Cliff Walking environment. $\gamma = 0.9$, $|\mathcal{S}| = 48$.

Application

Learning curve on Cliff Walking

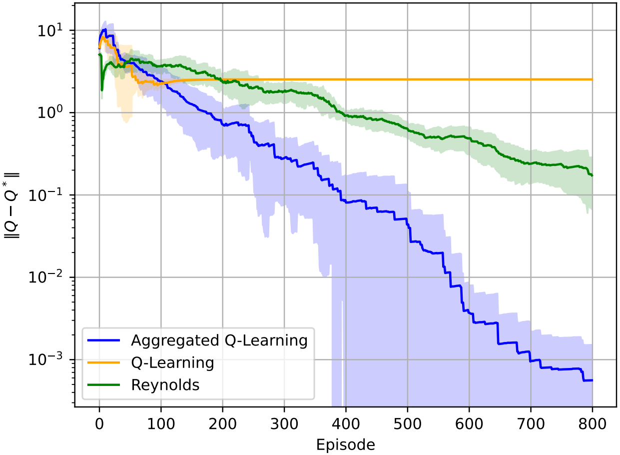

Figure 5: Error to optimal along episodes of learning. $\gamma = 0.9$, 128 steps per episode.

Application

Associated Abstraction

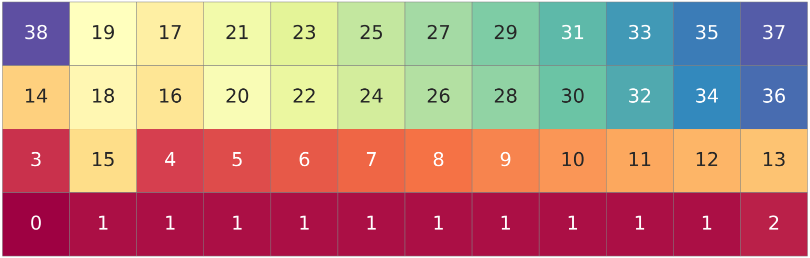

Figure 6: Abstraction found using AggQL on the Cliff Walking environment.

Application
Mountain Car Environment

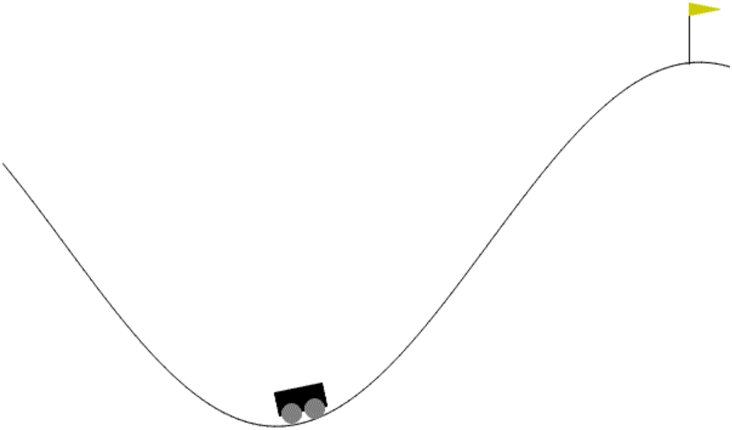

Figure 7: Mountain Car environment [Towers et al., 2024].

Application

Mountain Car Optimal Value

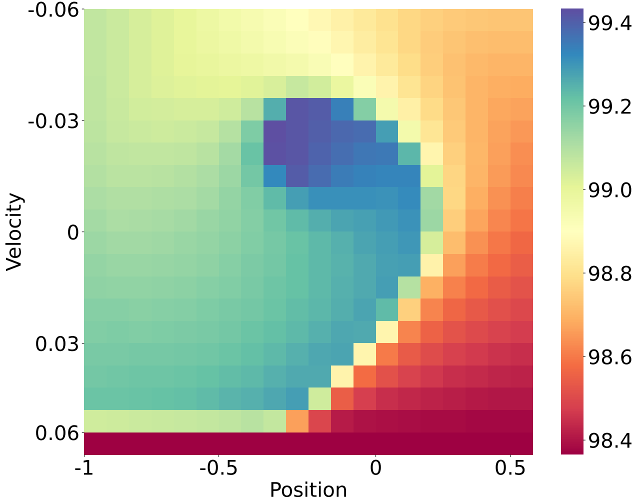

Figure 8: Optimal value function. $\gamma = 0.99$, $|\mathcal{S}| = 400$.

Application

Learning curve on Mountain Car

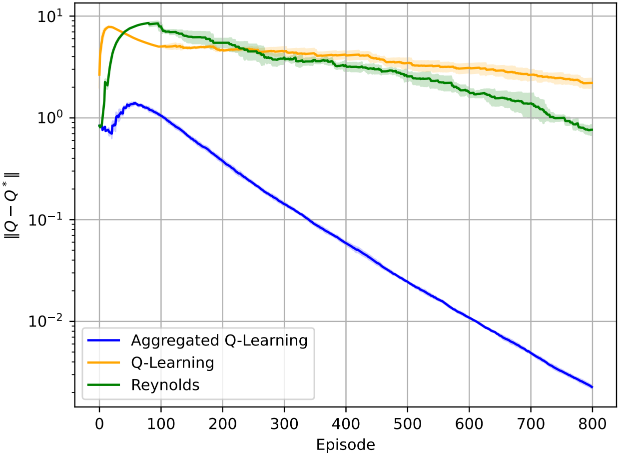

Figure 9: Error to optimal along episodes of learning. $\gamma = 0.99$, 1024 steps per episode.

Application

Associated Abstraction

25	25	24	23	23	22	21	21	20	19	18	18	17	16	15	14	14	14	13	13
25	25	24	23	23	22	22	21	20	20	19	18	17	16	15	14	13	13	13	13
25	25	24	23	23	22	22	22	21	21	20	19	18	17	15	14	13	13	12	12
25	25	24	24	23	23	23	23	24	25	24	22	20	18	16	15	13	12	12	12
25	25	24	24	24	24	24	26	32	37	37	35	29	21	17	15	14	12	11	11
26	25	25	25	25	25	26	29	38	37	37	36	36	33	21	16	14	12	11	11
26	25	25	25	25	26	27	30	38	37	37	36	35	35	31	18	14	12	11	10
26	26	26	26	26	27	28	30	34	37	36	35	35	34	34	23	15	12	11	10
27	26	26	26	27	27	28	29	31	33	34	34	34	34	34	27	15	12	10	9
27	27	27	27	27	28	28	29	30	31	32	32	33	33	34	28	14	11	10	8
27	27	27	27	28	28	28	29	30	30	31	32	32	33	33	24	13	10	9	8
28	28	28	28	28	28	29	29	30	30	31	32	32	33	33	18	11	9	8	7
28	28	28	28	28	29	29	29	30	30	31	32	32	33	26	12	9	8	7	6
29	29	29	29	29	29	29	30	30	31	31	32	32	31	14	9	7	6	6	5
29	29	29	29	29	29	30	30	31	31	32	32	32	18	9	7	6	5	5	4
29	29	30	30	30	30	30	31	31	32	32	32	18	9	7	5	5	4	4	3
30	30	30	30	30	30	31	31	32	32	32	18	8	6	5	4	3	3	3	3
30	30	30	30	31	31	31	32	32	33	24	7	4	3	3	3	2	2	2	2
24	25	25	25	25	25	25	26	25	11	5	2	2	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 10: Abstraction found using AggQL on the Mountain Car environment.

Conclusion

In this work:

- We proposed a model-free Q-Learning-based algorithm that:
  - Refines state abstraction dynamically,
  - Increase sample efficiency of QL.
- We brougth theoretical guarantees:
  - On the convergence itself,
  - On the provided abstraction.
- We led a practical evaluation.

We consider to extend this work by:

- Evaluating the robustness of AggQL in highly stochastic environments,
- Incorporating deep learning techniques to handle high-dimensional state spaces.

Abel, D., Hershkowitz, D., and Littman, M. (2016).
Near optimal behavior via approximate state abstraction.
In *International Conference on Machine Learning*, pages 2915–2923. PMLR.

Forghieri, O., Le Pennec, E., Castel, H., and Hyon, E. (2024).
Progressive state space disaggregation for infinite horizon Dynamic Programming.
In *34th International Conference on Automated Planning and Scheduling*.

Jaakkola, T., Jordan, M., and Singh, S. (1993).
Convergence of stochastic iterative dynamic programming algorithms.
Advances in neural information processing systems, 6.

Li, L., Walsh, T. J., and Littman, M. L. (2006).
Towards a unified theory of state abstraction for MDPs.
In *AI&M*.

Nachum, O., Gu, S. S., Lee, H., and Levine, S. (2018).

Data-efficient Hierarchical Reinforcement Learning.
Advances in neural information processing systems, 31.

Pateria, S., Subagdja, B., Tan, A.-h., and Quek, C. (2021).
Hierarchical Reinforcement Learning: A comprehensive survey.
ACM Computing Surveys (CSUR), 54(5):1–35.

Starre, R. A., Loog, M., and Oliehoek, F. A. (2022).
Model-based Reinforcement Learning with state abstraction: A survey.
In *Benelux Conference on Artificial Intelligence*, pages 133–148. Springer.

Sutton, R. S. and Barto, A. G. (2018).
Reinforcement Learning: An introduction.
MIT press.

Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., Goulao, M., Kallinteris, A., Krimmel, M., KG, A., et al. (2024).

Gymnasium: A standard interface for Reinforcement Learning environments.
arXiv preprint arXiv:2407.17032.