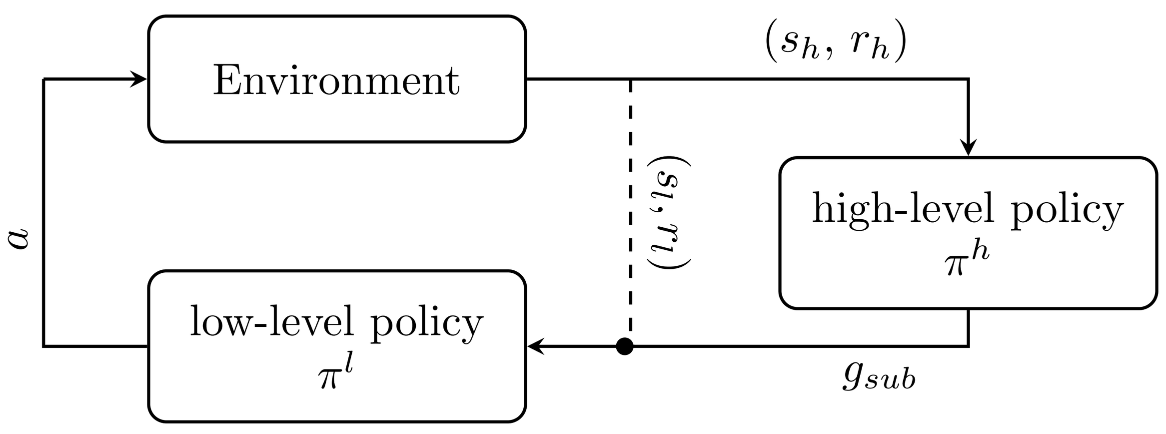
Imperial College London

Brain & Behaviour Lab

SAMPLE COMPLEXITY OF HIERARCHICAL DECOMPOSITIONS IN MARKOV DECISION PROCESSES

Arnaud Robert[1], Ciara Pike-Burke[2], Aldo A. Faisal[1]
[1]Brain & Behaviour Lab: Department of Computing, Imperial College London, UK,
[2]Statistics Section, Department of Mathematics, Imperial College London, UK

GOAL CONDITIONED HIERARCHICAL RL



- **Background:**
  - Hierarchical RL leverages state abstraction [4] and temporal abstraction [7] to improve sample efficiency.
  - Little is known about the reasons for HRL empirical efficiency [5, 1].
- **Contributions:**
  - We formalize the decomposition induced by the hierarchy.
  - We extend the PAC lower bound of [3] to HRL.
  - The bound relates the decomposition characteristics to the sample efficiency.
  - We propose a new HRL algorithm.

HIERARCHICAL EPISODIC FIXED-HORIZON MDP

$$\mathcal{M}_O = \langle \mathcal{S}, \mathcal{A}, P, r, p_0, H \rangle$$

$$\mathcal{M}_l = \langle \mathcal{S}_l \times \mathcal{A}_h, \mathcal{A}, P_l, r_l, p_{0,l}, H_l \rangle \qquad \mathcal{M}_h = \langle \mathcal{S}_h, \mathcal{A}_h, P_h, r_h, p_{0,h}, H_h \rangle$$

LOW-LEVEL MDP

- **State space:** Low-level states consist of $(s_l, a_h) \in \mathcal{S}_l \times \mathcal{A}_h$. Where s_l is the low-level component of the original state $s \in \mathcal{S}$, with $s = (s_l, s_h)$ and a_h is the sub-goal.
- **Action space:** The low-level action corresponds to the original action space $\mathcal{A}$.
- **Transition function:** P_l is a restriction of P on $\mathcal{S}_l \times \mathcal{A}$.
- **Reward function:** The low-level reward function is $r_l(s_l, a_h) = 2r(s_l, a_h)$.
- **Initial state distribution:** $p_{0,l}$ spans the entire low-level state space.
- **Horizon:** The low-level horizon satisfies $H_l = \frac{H}{H_h}$.

HIGH-LEVEL MDP

- **State space:** As any state $s \in \mathcal{S}$ can be represented as a tuple $s = (s_l, s_h)$ the high-level state is s_h.
- **Action space:** $\mathcal{A}_h(s_h)$ corresponds to the set of sub-goals available in state s_h.
- **Transition function:** The probability of observing s'_h is given by $P_h(s'_h | s_h, a_h, \pi_l)$.
- **Reward function:** The high-level reward function is the sum of cumulated low-level reward: $r_h(s_h, a_h) = \sum_{t=1}^{H_l} r_l(s_l, a_h)$.
- **Initial state distribution:** $p_{0,h}$ is a restriction of p_0 on $\mathcal{S}_h$.
- **Horizon:** H_h must satisfy $H_h = \frac{H}{H_l}$.

SAMPLE-COMPLEXITY OF REINFORCEMENT LEARNING

Definition[2]: An algorithm satisfies a PAC bound N if, for a given input $\epsilon > 0$ and $\delta < 1$, it satisfies the following condition for any episodic fixed-horizon MDP. With probability at least $1 - \delta$, the algorithm plays policies that are at least ϵ-optimal after at most N episodes. That is, with probability at least $1 - \delta$

$$\max\{k \in \mathbb{N} : \Delta_k > \epsilon\} \leq N,$$

where N is a polynomial that can depend on the properties of the problem instance.

SAMPLE-COMPLEXITY OF HIERARCHICAL RL

Theorem: There exist positive constants c_l, c_h and δ_0 such that for every $\delta \in (0, \delta_0)$ and for every algorithm A that satisfies a PAC guarantee for (ϵ, δ) and outputs a deterministic policy, there is a fixed horizon MDP such that A must interact for

$$\mathbb{E}[N] = \Omega\left(\max\left(\frac{|\mathcal{S}_l||\mathcal{A}_h||\mathcal{A}|H_l^2}{\epsilon^2}\ln\left(\frac{1}{\delta + c_l}\right), \frac{|\mathcal{S}_h||\mathcal{A}_h|H_h^2}{\epsilon^2}\ln\left(\frac{1}{\delta + c_h}\right)\right)\right) \tag{1}$$

episodes until the policy is (ϵ, δ)-accurate. Full proof in [6].

HIERARCHICAL Q-LEARNING

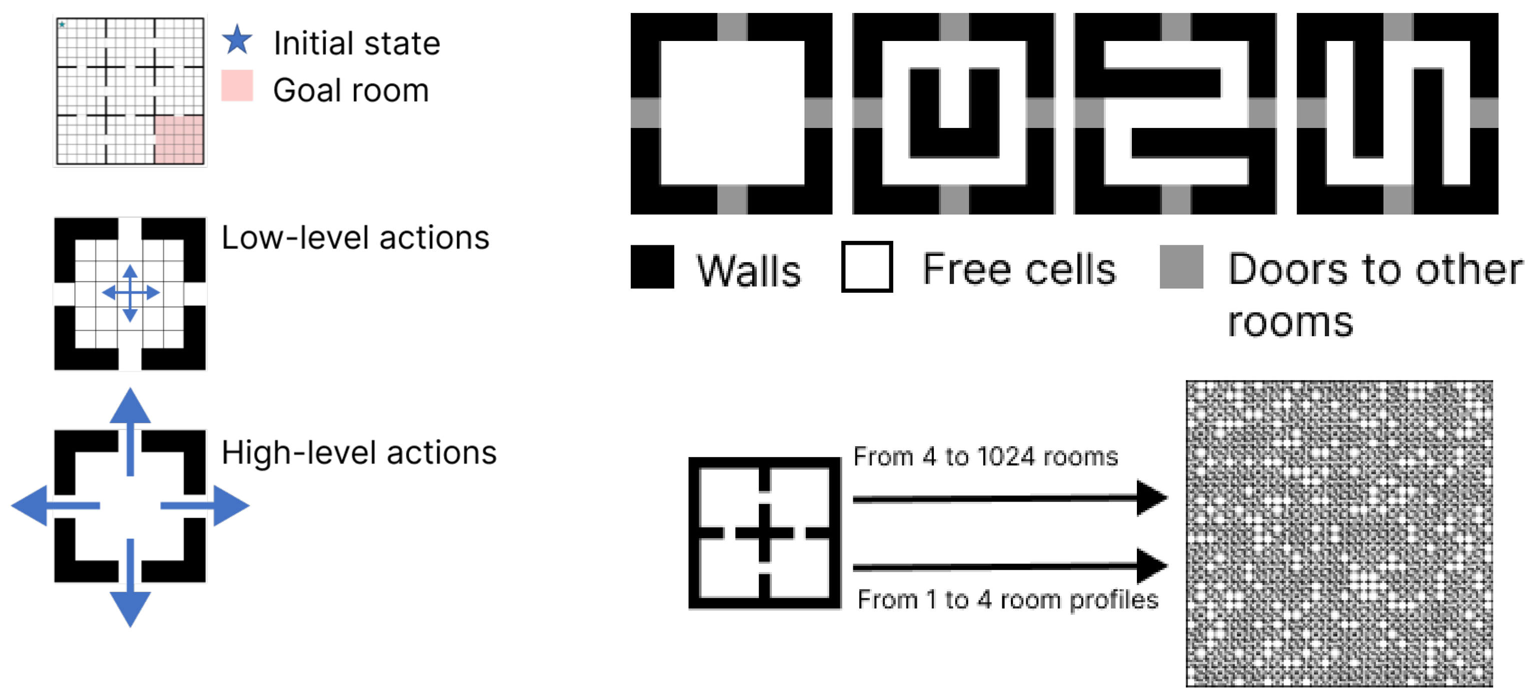
Algorithm 1: Stationary Hierarchical Q-learning (SHQL)

Input: $Q^L_{:,:,:} = 0$, $Q^H_{:,:} = 0$, $\text{done}_H = False$, $t = k = 0$

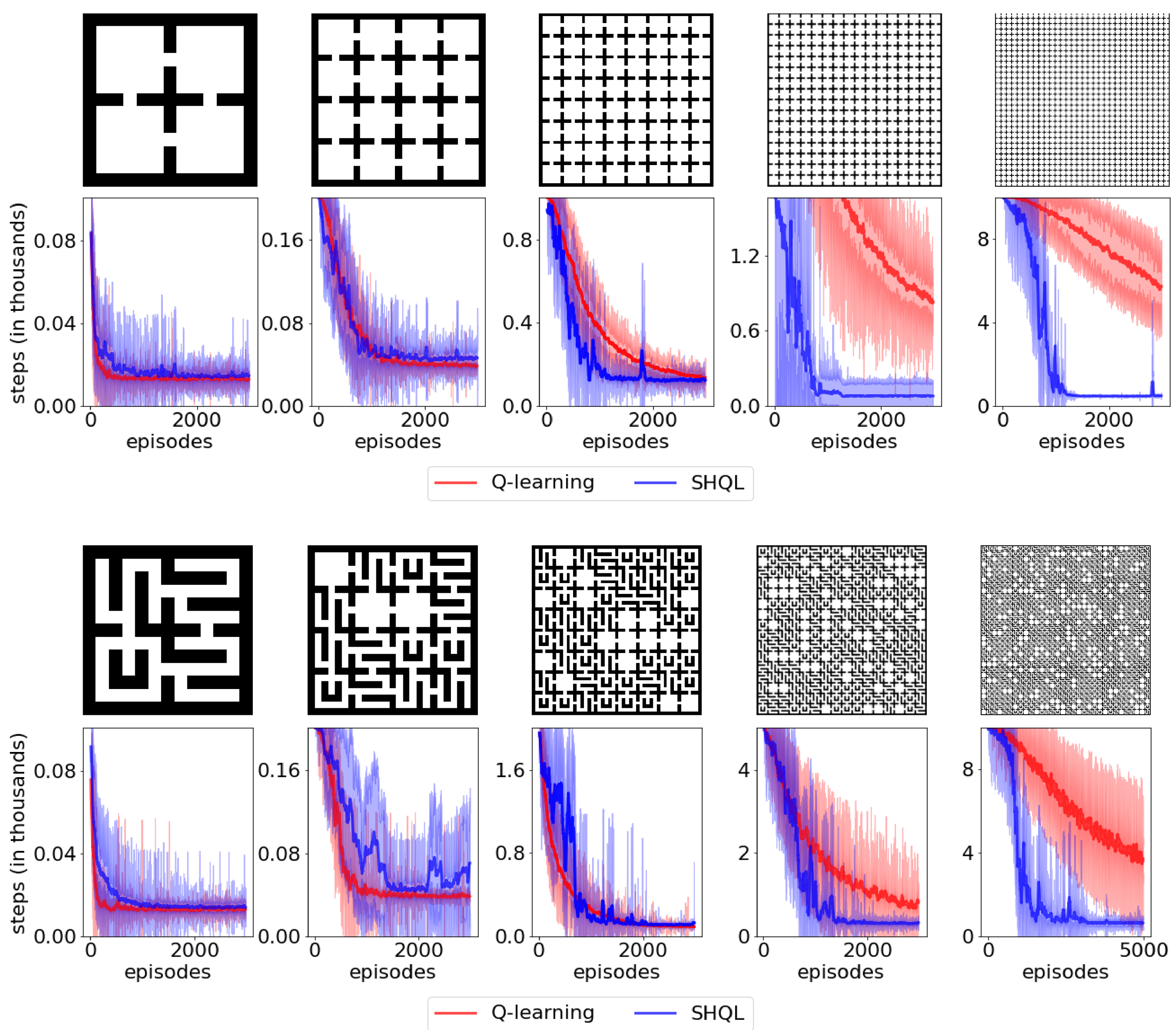
while *not* $done_H$ *and* $k < K$ **do**
- Observe s^H_k, s^L_t
- **while** *not* $done_L$ *and* $t < T$ **do**
  - $a^L_t = \pi^L(s^L_t)$
  - Observe s^L_{t+1}, r^L_t
  - *LowLevelUpdate*$((s^L_t, a^L_t, r^L_t, s^L_{t+1}, g_{sub}))$
  - $s_t = s_{t+1}$
  - $t = t + 1$
- Observe s^H_{k+1} r^H_k
- **if** $done_L$ **then**
  - $Q^H_{nxt} = \max_a Q^H_{s^H_{k+1},a}$
  - $Q^L_{s_k,a_k} = Q^H_{s_k,a_k} + \alpha * (r^H_k + \gamma Q^H_{nxt})$

Function `LowLevelUpdate`$(s_t, a_t, r_t, s_{t+1}\ g_{sub})$**:**
- $Q^L_{nxt} = \max_a Q^L_{g_{sub},s_{t+1},a}$
- $Q^L_{g_{sub},s_t,a_t} = Q^L_{g_{sub},s_t,a_t} + \alpha * (r^L_t + \gamma Q^L_{nxt})$
- **return** Q^L

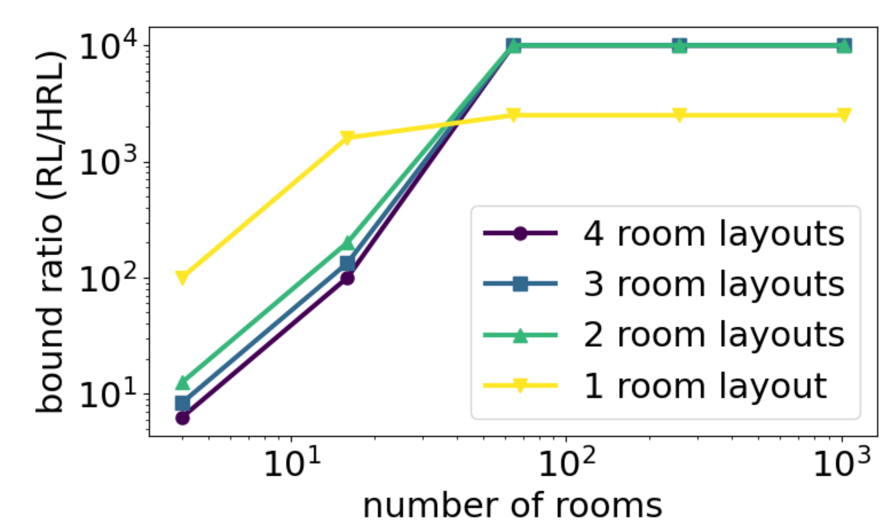
DIVERSITY OF MDPS



PERFORMANCE STATIONARY HQL VS Q-LEARNING



DISCUSSION



Conclusions

- Empirical and theoretical results are aligned.
- Both state and temporal abstractions play a significant role in HRL efficiency.
- We provided theoretical and empirical evidence of these phenomena.

Limitations

- In this work, the decomposition is given. In nature, it should be learned.
- The discrete setting does not allow us to account for generalization over-subgoals.

ACKNOWLEDGEMENTS

AR was supported by an EPSRC CASE studentship supported by Shell and AAF was supported by a UKRI Turing AI Fellowship (EP/V025449/1).

UK Research and Innovation

EPSRC Engineering and Physical Sciences Research Council

REFERENCES

[1] Benjamin Beyret, Ali Shafti, and A Aldo Faisal. "Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 5014–5019.

[2] Christoph Dann. "Strategic exploration in reinforcement learning-new algorithms and learning guarantees". PhD thesis. Google, 2019.

[3] Christoph Dann and Emma Brunskill. "Sample complexity of episodic fixed-horizon reinforcement learning". In: *Advances in Neural Information Processing Systems* 28 (2015).

[4] Peter Dayan and Geoffrey E Hinton. "Feudal reinforcement learning". In: *Advances in neural information processing systems* 5 (1992).

[5] Ofir Nachum et al. "Data-efficient hierarchical reinforcement learning". In: *Advances in neural information processing systems* 31 (2018).

[6] Arnaud Robert, Ciara Pike-Burke, and Aldo A Faisal. "Sample Compexity of Hierarchical Decompositions in Markov Decision Processes". In: *New Frontiers in Learning, Control, and Dynamical Systems, ICML workshop* (2023).

[7] Richard S Sutton, Doina Precup, and Satinder Singh. "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning". In: *Artificial intelligence* 112.1-2 (1999), pp. 181–211.