A Appendix

A.1 Proof of Theorem

Theorem 3.1 states that 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 collect

$$\mathbb{E}[N_e] = \Omega\Bigg(\max\Bigg(\frac{|\mathcal{S}_l||\mathcal{A}_h||\mathcal{A}|H_l^2}{\epsilon^2} \ln\big(\frac{1}{\delta + c_l}\big), \frac{|\mathcal{S}_h||\mathcal{A}_h|H_h^2}{\epsilon^2} \ln\big(\frac{1}{\delta + c_h}\big)\Bigg)\Bigg) \tag{7}$$

episodes until its policy is (ϵ, δ)-accurate.

Proof. An ϵ-accurate pair of policies (π_l, π_h) satisfies $|V_o^{\pi_l^*,\pi_h^*} - V_o^{\pi_l,\pi_h}| \leq \epsilon$. Note that by the triangle inequality, if $|V_o^{\pi_l^*,\pi_h^*} - V_o^{\pi_l^*,\pi_h}| + |V_o^{\pi_l^*,\pi_h} - V_o^{\pi_l,\pi_h}| \leq \epsilon$, then we will have $|V_o^{\pi_l^*,\pi_h^*} - V_o^{\pi_l,\pi_h}| \leq \epsilon$. We, therefore, focus on showing:

(i) the number of samples required to guarantee $|V_o^{\pi_l^*,\pi_h^*} - V_o^{\pi_l^*,\pi_h}| \leq \epsilon/2$ is bounded by $\Omega\Bigg(\frac{|\mathcal{S}_h||\mathcal{A}_h|H_h^2}{\epsilon^2} \ln\big(\frac{1}{\delta+c_h}\big)\Bigg)$

(ii) the number of samples required to guarantee $|V_o^{\pi_l^*,\pi_h} - V_o^{\pi_l,\pi_h}| \leq \epsilon/2$ is bounded by $\Omega\Bigg(\frac{|\mathcal{S}_l||\mathcal{A}_H||\mathcal{A}|H_l^2}{\epsilon^2} \ln\big(\frac{1}{\delta+c_l}\big)\Bigg)$

Then once we have both (i) and (ii), we know that after

$$\Omega\Bigg(\max\Bigg(\frac{|\mathcal{S}_L||\mathcal{A}_H||\mathcal{A}|H_L^2}{\epsilon^2} \ln\big(\frac{1}{\delta + c_l}\big), \frac{|\mathcal{S}_H||\mathcal{A}_H|H_H^2}{\epsilon^2} \ln\big(\frac{1}{\delta + c_h}\big)\Bigg)\Bigg)$$

episodes, we will have $|V_o^{\pi_l^*,\pi_h^*} - V_o^{\pi_l^*,\pi_h}| + |V_o^{\pi_l^*,\pi_h} - V_o^{\pi_l,\pi_h}| \leq \epsilon$ and so $|V_o^{\pi_l^*,\pi_h^*} - V_o^{\pi_l,\pi_h}| \leq \epsilon$.

Part (i) Note that only learning the high-level policy when the low-level policy is optimal, is equivalent to learning an ϵ-accurate high-level policy interacting with $\mathcal{M}_h$ with a stationary transition function (since the low-level behaviour is not evolving anymore). Hence we can bound the number of episodes N_h required to have: $|V_h^* - V_h^{\pi_l^*,\pi_h}| \leq \epsilon$, by directly applying Eq. (4) to the high-level MDP to get

$$\mathbb{E}[N_h] = \Omega\Bigg(\frac{|\mathcal{S}_h||\mathcal{A}_h|H_h^2}{\epsilon^2} \ln\big(\frac{1}{\delta + c_h}\big)\Bigg)$$

To be able to use this result to construct the bound of interest, we need to make sure these results are valid under the original MDP: $|V_o^{\pi_l^*,\pi_h^*} - V_o^{\pi_l^*,\pi_h}| \leq \epsilon$. In particular, the reward functions are not the same for $\mathcal{M}_o$ and $\mathcal{M}_h$. By decomposition, r_h includes the bonus (or the absence of penalty) the high-level gives to the low-level for completing the task. To compensate for that the low-level reward is re-scaled with a penalty twice larger per step. This ensure that $|V_o^{\pi_l^*,\pi_h^*} - V_o^{\pi_l^*,\pi_h}| \leq 2|V_h^* - V_h^{\pi_l^*,\pi_h}|$. Hence after $\mathbb{E}[N_h]$ episodes, we have $|V_o^* - V_o^{\pi_l^*,\pi_h}| \leq 2\epsilon$

Part (ii) By a similar argument to Part (i), we can bound the number of episodes in the low-level MDP required to obtain an ϵ-optimal low-level policy for a fixed high-level policy π_h. In particular, a lower bound on the number of episodes N_l required to have $|V_l^{\pi_h,\pi_l^*} - V_l^{\pi_l,\pi_h}| \leq \epsilon$ can directly be obtained from Eq. (4):

$$\mathbb{E}[N_l] = \Omega\Bigg(\frac{|\mathcal{S}_l||\mathcal{A}_H||\mathcal{A}|H_l^2}{\epsilon^2} \ln\big(\frac{1}{\delta + c_l}\big)\Bigg).$$

We are interested in comparing the policies when they interact with the original MDP. The issue is that there is a difference of scale between $V_o^{\pi_l,\pi_h}$ and $V_l^{\pi_l,\pi_h}$. Episodes are shorter by a factor of H_h in the low-level MDP. So we need to ensure that $|V_l^{\pi_h,\pi_l^*} - V_l^{\pi_l,\pi_h}| \leq \frac{\epsilon}{H_h}$. But by construction, this re-scaling is not necessary as a single episode in the original MDP corresponds to at most H_h episodes in the low-level MDP as a single episode in $\mathcal{M}_o$ with x sub-goals correspond to x episodes in $\mathcal{M}_l$.

This leads us to a lower bound on the number of episodes needed to obtain an ϵ-accurate pair of policies as the one stated in the theorem. □

A.2 Additional experiments

In the experimental section (Sec. 5) we used several room layouts. In the main paper, we only provide learning curves for mazes that are composed of rooms without any obstacles or mazes that are composed of all the possible room layouts depicted in the rightmost plot of figure 2. To complete our experiment we show below in (Fig. 4 and Fig. 5) the learning curve obtained when mazes are built from two or three different room layouts. Note also that those results were used to plot the evolution of the bound ratio in the rightmost plot of figure 3.

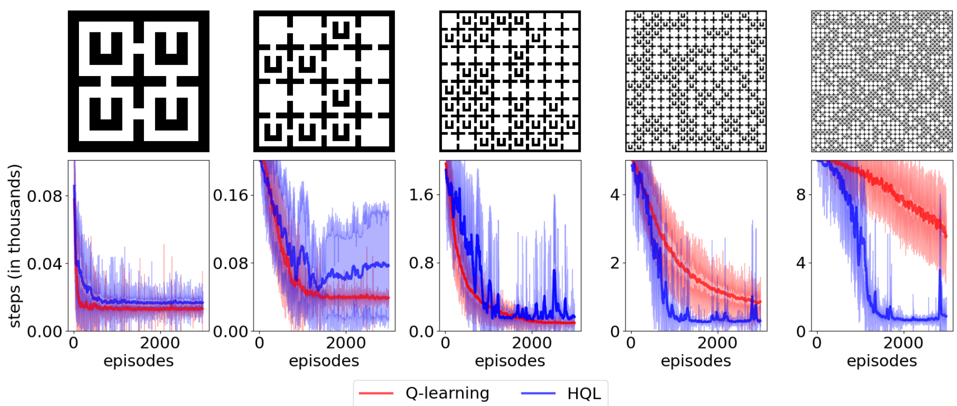

Figure 4: Shows learning curves on various maze sizes with two different room instances, either the room is empty or it has a U-shape obstacle in it. The performance of the agent is measured in the number of steps it requires to solve the task.

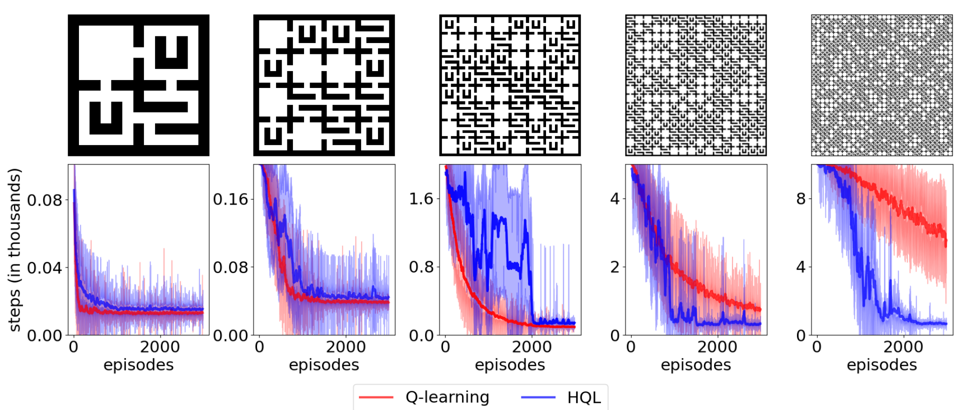

Figure 5: Shows learning curves on various maze sizes with three different room instances, either the room is empty or it has either a U-shape obstacle or the room is stripped with horizontal walls. The performance of the agent is measured in the number of steps it requires to solve the task.