Supplementary Material for "Model-Based Reinforcement Learning Exploiting State-Action Equivalence"

Mahsa Asadi MAHSA.ASADI@INRIA.FR
Mohammad Sadegh Talebi SADEGH.TALEBI@INRIA.FR
Hippolyte Bourel HIPPOLYTE.BOUREL@ENS-RENNES.FR
Odalric-Ambrym Maillard ODALRIC.MAILLARD@INRIA.FR
Inria Lille – Nord Europe

Editors: Wee Sun Lee and Taiji Suzuki

Appendix A. Pseudo-Codes of UCRL2 and C-UCRL

In this section, we provide the pseudo-codes of UCRL2, C-UCRL($\mathcal{C}, \boldsymbol{\sigma}$), and C-UCRL.

Algorithm 1 UCRL2 with input parameter $\delta \in (0,1]$ (Jaksch et al., 2010)

Initialize: For all (s,a), set $N_0(s,a) = 0$ and $v_0(s,a) = 0$. Set $t_0 = 0$, $t = 1$, $k = 1$, and observe the initial state s_1;
for episodes $k \geqslant 1$ **do**
  Set $t_k = t$;
  Set $N_{t_k}(s,a) = N_{t_{k-1}}(s,a) + v_k(s,a)$ for all (s,a);
  Compute empirical estimates $\widehat{\mu}_{t_k}(s,a)$ and $\widehat{p}_{t_k}(\cdot|s,a)$ for all (s,a);
  Compute $\pi^+_{t_k} = \texttt{EVI}\left(\widehat{\mu}_{t_k}, \widehat{p}_{t_k}, N_{t_k}, \frac{1}{\sqrt{t_k}}, \frac{\delta}{SA}\right)$ — see Algorithm 2;
  while $v_k(s_t, \pi^+_{t_k}(s_t)) < \max\{1, N_{t_k}(s_t, \pi^+_{t_k}(s_t))\}$ **do**
    Play action $a_t = \pi^+_{t_k}(s_t)$, and observe the next state s_{t+1} and reward $r_t(s_t, a_t)$;
    Set $v_k(s_t, a_t) = v_k(s_t, a_t) + 1$;
    Set $t = t + 1$;
  end while
end for

Appendix B. Proof of Lemma 7

Let us consider the case when a switch occurs between index 1 and 2, that is $\sigma_q(1) = \sigma_p(2)$ and $\sigma_q(2) = \sigma_p(1)$. In this situation, we thus have $p(\sigma_p(1)) > p(\sigma_p(2))$ but $p(\sigma_q(1)) \leqslant p(\sigma_q(2))$. Then, we study $\sum_{i=1,2} |p(\sigma_p(i)) - q(\sigma_q(i))|$. First, we note that if $q(\sigma_q(1)) < p(\sigma_p(1))$ and $q(\sigma_q(2)) < p(\sigma_p(2))$, then

$$\begin{aligned}|p(\sigma_p(1)) - q(\sigma_q(1))| + |p(\sigma_p(2)) - q(\sigma_q(2)))| &= p(\sigma_p(1)) - q(\sigma_q(2)) + p(\sigma_p(2)) - q(\sigma_q(1)) \\ &= |p(\sigma_p(1)) - q(\sigma_p(1))| + |p(\sigma_p(2)) - q(\sigma_p(2))|\,.\end{aligned}$$

Likewise, the same equality occurs if $q(\sigma_q(1)) > p(\sigma_p(1))$ and $q(\sigma_q(2)) > p(\sigma_p(2))$. Now, in the remaining intermediate cases (that is $q(\sigma_p(1)) < p(\sigma_p(2)) < q(\sigma_p(2)) < p(\sigma_p(1))$,

Algorithm 2 $\texttt{EVI}(\mu, p, N, \varepsilon, \delta)$ (Jaksch et al., 2010)

Initialize: $u^{(0)} \equiv 0, u^{(-1)} \equiv -\infty$, $n = 0$;
while $\max_s(u^{(n)}(s) - u^{(n-1)}(s)) - \min_s(u^{(n)}(s) - u^{(n-1)}(s)) > \varepsilon$ **do**
For all (s,a), set $\mu'(s,a) = \mu(s,a) + \beta'_{N(s,a)}(\delta)$;
For all (s,a), set $p'(\cdot|s,a) \in \operatorname{argmax}_{q \in \mathcal{P}(s,a)} \sum_{x \in \mathcal{S}} q(x) u^{(n)}(x)$ where

$$\mathcal{P}(s,a) := \left\{ q \in \Delta^S : \|q - p(\cdot|s,a)\|_1 \leqslant \beta_{N(s,a)}(\delta) \right\};$$

For all s, update $u^{(n+1)}(s) = \max_{a \in \mathcal{A}} \left(\mu'(s,a) + \sum_{x \in \mathcal{S}} p'(x|s,a) u^{(n)}(x) \right)$;
For all s, update $\pi_{n+1}(s) \in \operatorname{Argmax}_{a \in \mathcal{A}} \left(\mu'(s,a) + \sum_{x \in \mathcal{S}} p'(x|s,a) u^{(n)}(x) \right)$;
Set $n = n + 1$;
end while
Output: π_{n+1}

Algorithm 3 $\texttt{C-UCRL}(\mathcal{C}, \boldsymbol{\sigma})$ with input parameter $\delta \in (0,1]$

Initialize: For all $c \in \mathcal{C}$, set $n_0(c) = 0$ and $V_0(c) = 0$. Set $t_0 = 0$, $t = 1$, $k = 1$, and observe the initial state s_1;
for episodes $k \geqslant 1$ **do**
Set $t_k = t$;
Set $n_{t_k}(c) = n_{t_k - 1}(c) + V_{k-1}(c)$ for all c;
Compute empirical estimates $\widehat{\mu}^{\boldsymbol{\sigma}}_{t_k}(c)$ and $\widehat{p}^{\boldsymbol{\sigma}}_{t_k}(\cdot|c)$ for all c;
Compute $\pi^+_{t_k} = \texttt{EVI}\left(\widehat{\mu}^{\boldsymbol{\sigma}}_{t_k}, \widehat{p}^{\boldsymbol{\sigma}}_{t_k}, n_{t_k}, \frac{1}{\sqrt{t_k}}, \frac{\delta}{C}\right)$ — see Algorithm 2;
while $V_k(c_t) < \max\{1, n_{t_k}(c_t)\}$ **do**
Play action $a_t = \pi^+_{t_k}(s_t)$, and observe the next state s_{t+1} and reward $r_t(s_t, a_t)$;
Set $c_t \in \mathcal{C}$ to be the class containing (s_t, a_t);
Set $V_k(c_t) = V_k(c_t) + 1$;
Set $t = t + 1$;
end while
end for

Algorithm 4 $\texttt{C-UCRL}$ with input parameter $\delta \in (0,1]$

Initialize: For all (s,a), set $N_0(s,a) = 0$ and $v_0(s,a) = 0$. For all $c \in \mathcal{C}$, set $n_0(c) = 0$ and $V_0(c) = 0$. Set $t_0 = 0$, $t = 1$, $k = 1$, and observe the initial state s_1;
for episodes $k \geqslant 1$ **do**
Set $t_k = t$;
Set $N_{t_k}(s,a) = N_{t_k - 1}(s,a) + v_k(s,a)$ for all (s,a);
Set $n_{t_k}(c) = n_{t_k - 1}(c) + V_{k-1}(c)$ for all c;
Compute empirical estimates $\boldsymbol{\sigma}_{t_k}$;
Find $\mathcal{C}_{t_k}$ using $\texttt{ApproxEquivalence}$;
Compute empirical estimates $\widehat{\mu}^{\boldsymbol{\sigma}_{t_k}}_{t_k}(c)$ and $\widehat{p}^{\boldsymbol{\sigma}_{t_k}}_{t_k}(\cdot|c)$ for all $c \in \mathcal{C}_{t_k}$;
Compute $\pi^+_{t_k} = \texttt{EVI}\left(\widehat{\mu}^{\boldsymbol{\sigma}_{t_k}}_{t_k}, \widehat{p}^{\boldsymbol{\sigma}_{t_k}}_{t_k}, n_{t_k}, \frac{1}{\sqrt{t_k}}, \frac{\delta}{SA}\right)$ — see Algorithm 2;
while $v_k(s_t, \pi^+_{t_k}(s_t)) < \max\{1, N_{t_k}(s_t, \pi^+_{t_k}(s_t))\}$ and $V_k(c_t) < \max\{1, n_{t_k}(c_t)\}$ **do**
Play action $a_t = \pi^+_{t_k}(s_t)$, and observe the next state s_{t+1} and reward $r_t(s_t, a_t)$;
Set $c_t \in \mathcal{C}_{t_k}$ to be the class containing (s_t, a_t);
Set $V_k(c_t) = V_k(c_t) + 1$;
Set $v_k(s_t, a_t) = v_k(s_t, a_t) + 1$;
Set $t = t + 1$;
end while
end for

$p(\sigma_p(2)) < q(\sigma_p(1)) < q(\sigma_p(2)) < p(\sigma_p(1))$, and $p(\sigma_p(2)) < q(\sigma_p(1)) < p(\sigma_p(1)) < q(\sigma_p(2)))$, it is immediate to check that

$$|p(\sigma_p(1)) - q(\sigma_q(1))| + |p(\sigma_p(2)) - q(\sigma_q(2)))| \leqslant |p(\sigma_p(1)) - q(\sigma_p(1))| + |p(\sigma_p(2)) - q(\sigma_p(2))|\,.$$

Thus, proceeding iteratively for all switch that occurs, and decomposing the permutations σ_p and σ_q into elementary switches, we deduce that almost surely

$$\|q(\sigma_q(\cdot)) - p(\sigma_p(\cdot))\|_1 \leqslant \|q(\sigma_p(\cdot)) - p(\sigma_p(\cdot))\|_1 = \|p - q\|_1\,,$$

thus concluding the lemma. ■

Appendix C. Proof of Proposition 13

First recall that $\Delta := \min\big\{d(\{\ell\},\{\ell'\}) : \ell,\ell' \in \mathcal{S}\times\mathcal{A} \text{ and } \ell,\ell' \text{ are not in the same class}\big\}$. Define

$$\mathcal{E} = \bigcap_{t\in\mathbb{N}}\bigcap_{s,a}\Big\{\|p(\cdot|s,a) - \widehat{p}_t(\cdot|s,a)\|_1 \leqslant \beta_{N_t(s,a)}\big(\tfrac{\delta}{SA}\big)\Big\}\,.$$

Note that $\mathbb{P}(\mathcal{E}) \geqslant 1-\delta$. We will need the following lemma:

Lemma 1 *Assume that the event $\mathcal{E}$ holds. Then, for all t, at every round k of* ApproxEquivalence*, for all $v \in \mathcal{C}^k$, there exists $u \in \mathcal{N}(v)$ such that u and v belong to the same class.*

Proof (of Lemma 1) Fix $t \geqslant 1$ and round k, and consider $v \in \mathcal{C}^k$. Recall that u is a PAC Neighbor of v if it satisfies:

$$\begin{aligned}
&(i) \quad \big\|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot|u) - \widehat{p}_t^{\boldsymbol{\sigma}_{v,t}}(\cdot|v)\big\|_1 - \varepsilon_{u,t} - \varepsilon_{v,t} \leqslant 0\,;\\
&(ii) \quad \big\|\widehat{p}_t(\sigma_{i,t}(\cdot)|i) - \widehat{p}_t(\sigma_{j,t}(\cdot)|j)\big\|_1 - \beta_{N_t(i)}\big(\tfrac{\delta}{SA}\big) - \beta_{N_t(j)}\big(\tfrac{\delta}{SA}\big) \leqslant 0\,, \quad \forall i\in u, \forall j \in v;\\
&(iii) \quad \big\|\widehat{p}_t(\sigma_{\ell,t}(\cdot)|\ell) - \widehat{p}_t^{\boldsymbol{\sigma}_{u\cup v,t}}(\cdot|u\cup v)\big\|_1 - \beta_{N_t(\ell)}\big(\tfrac{\delta}{SA}\big) - \varepsilon_{u\cup v,t} \leqslant 0\,, \quad \forall \ell \in u\cup v.
\end{aligned}$$

In order to prove the lemma, it suffices to show that under $\mathcal{E}$, there exists $u \subset \mathcal{S}\times\mathcal{A}$ satisfying (i)–(iii) and $d(u,v) = 0$. To this end, we will show that the event $\mathcal{E}$ implies the following: For all $u \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
&(i') \quad \big\|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot|u) - \widehat{p}_t^{\boldsymbol{\sigma}_{v,t}}(\cdot|v)\big\|_1 \leqslant d(u,v) + \varepsilon_{u,t} + \varepsilon_{v,t}\,;\\
&(ii') \quad \big\|\widehat{p}_t(\sigma_{i,t}(\cdot)|i) - \widehat{p}_t(\sigma_{j,t}(\cdot)|j)\big\|_1 \leqslant d(\{i\},\{j\}) + \beta_{N_t(i)}\big(\tfrac{\delta}{SA}\big) + \beta_{N_t(j)}\big(\tfrac{\delta}{SA}\big)\,, \quad \forall i\in u, \forall j\in v;\\
&(iii') \quad \big\|\widehat{p}_t(\sigma_{\ell,t}(\cdot)|\ell) - \widehat{p}_t^{\boldsymbol{\sigma}_{u\cup v,t}}(\cdot|u\cup v)\big\|_1 \leqslant d(\{\ell\},u\cup v) + \beta_{N_t(\ell)}\big(\tfrac{\delta}{SA}\big) + \varepsilon_{u\cup v,t}\,, \quad \forall \ell \in u\cup v.
\end{aligned}$$

Now, (i')–(iii') imply that there exists $u \in \mathcal{N}(v)$ such that u and v belong to the same class, and the lemma follows. It remains to prove (i')–(iii').

Proof of (i')**.** Consider $u \in \mathcal{S} \times \mathcal{A}$. Then, the non-expansive property of the norm function implies

$$\begin{aligned}\|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot|u) - \widehat{p}_t^{\boldsymbol{\sigma}_{v,t}}(\cdot|v)\|_1 &\leqslant \|p^{\boldsymbol{\sigma}_u}(\cdot|u) - p^{\boldsymbol{\sigma}_v}(\cdot|v)\|_1 + \|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot|u) - p^{\boldsymbol{\sigma}_u}(\cdot|u)\|_1 + \|\widehat{p}_t^{\boldsymbol{\sigma}_{v,t}}(\cdot|v) - p^{\boldsymbol{\sigma}_v}(\cdot|v)\|_1 \\ &= d(u,v) + \underbrace{\|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot|u) - p^{\boldsymbol{\sigma}_u}(\cdot|u)\|_1}_{A_1} + \underbrace{\|\widehat{p}_t^{\boldsymbol{\sigma}_{v,t}}(\cdot|v) - p^{\boldsymbol{\sigma}_v}(\cdot|v)\|_1}_{A_2}\,.\end{aligned}$$

The term A_1 is upper bounded as follows:

$$\begin{aligned}A_1 &= \sum_{x\in\mathcal{S}} \|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(x|u) - p^{\boldsymbol{\sigma}_u}(x|u)\|_1 \\ &= \sum_{x\in\mathcal{S}} \Big|\frac{1}{n_t(u)} \sum_{(s,a)\in u} N_t(s,a)\Big(\widehat{p}_t(\sigma_{s,a,t}(x)|s,a) - p(\sigma_{s,a}(x)|s,a)\Big)\Big| \\ &\leqslant \frac{1}{n_t(u)} \sum_{(s,a)\in u} N_t(s,a) \sum_{x\in\mathcal{S}} \Big|\widehat{p}_t(\sigma_{s,a,t}(x)|s,a) - p(\sigma_{s,a}(x)|s,a)\Big| \\ &\leqslant \frac{1}{n_t(u)} \sum_{(s,a)\in u} N_t(s,a) \|\widehat{p}_t(\sigma_{s,a,t}(\cdot)|s,a) - p(\sigma_{s,a}(\cdot)|s,a)\|_1 \\ &\leqslant \frac{1}{n_t(u)} \sum_{(s,a)\in u} N_t(s,a) \|\widehat{p}_t(\cdot|s,a) - p(\cdot|s,a)\|_1\,,\end{aligned}$$

where we have used Lemma 7 as well as the non-expansive property of the norm function. Hence, under the event $\mathcal{E}$,

$$\|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot|u) - p^{\boldsymbol{\sigma}_u}(\cdot|u)\|_1 \leqslant \frac{1}{n_t(u)} \sum_{(s,a)\in u} N_t(s,a)\beta_{N_t(s,a)}\big(\tfrac{\delta}{SA}\big) = \varepsilon_{u,t}\,. \tag{1}$$

A similar argument yields $A_2 \leqslant \varepsilon_{v,t}$ under $\mathcal{E}$. Putting these together verifies (i').

Proof of (ii')**.** The proof of (ii') is quite similar to that of (i'), hence omitted.

Proof of (iii')**.** Consider $u \in \mathcal{S} \times \mathcal{A}$ and $\ell \in u \cup v$. We have

$$\begin{aligned}\|\widehat{p}_t(\sigma_{\ell,t}(\cdot)|\ell) - \widehat{p}_t^{\boldsymbol{\sigma}_{u\cup v,t}}(\cdot|u\cup v)\|_1 &\leqslant \|p(\sigma_\ell(\cdot)|\ell) - p^{\boldsymbol{\sigma}_{u\cup v}}(\cdot|u\cup v)\|_1 + \|\widehat{p}_t(\sigma_{\ell,t}(\cdot)|\ell) - p(\sigma_\ell(\cdot)|\ell)\| \\ &\quad + \|\widehat{p}^{\boldsymbol{\sigma}_{u\cup v}}(\cdot|u\cup v) - \widehat{p}_t^{\boldsymbol{\sigma}_{u\cup v,t}}(\cdot|u\cup v)\|_1 \\ &\leqslant d(\{\ell\}, u\cup v) + \|\widehat{p}_t(\cdot|\ell) - p(\cdot|\ell)\|_1 + \|p^{\boldsymbol{\sigma}_{u\cup v}}(\cdot|u\cup v) - \widehat{p}_t^{\boldsymbol{\sigma}_{u\cup v,t}}(\cdot|u\cup v)\|_1\,,\end{aligned}$$

where we have used Lemma 7 and the non-expansive property of the norm function. The third term in the right-hand side is bounded as follows:

$$\begin{aligned}\|p^{\boldsymbol{\sigma}_{u\cup v}}(\cdot|u\cup v) - \widehat{p}_t^{\boldsymbol{\sigma}_{u\cup v,t}}(\cdot|u\cup v)\|_1 &\leqslant \sum_{(s,a)\in u\cup v} \frac{N_t(s,a)}{n_t(u)+n_t(v)} \sum_{x\in\mathcal{S}} \Big|p(\sigma_{s,a}(x)|x) - \widehat{p}_t(\sigma_{s,a,t}(x)|s,a)\Big| \\ &= \sum_{(s,a)\in u\cup v} \frac{N_t(s,a)}{n_t(u)+n_t(v)} \|\widehat{p}_t(\sigma_{s,a,t}(\cdot)|s,a) - p(\sigma_{s,a}(\cdot)|s,a)\|_1 \\ &\leqslant \sum_{(s,a)\in u\cup v} \frac{N_t(s,a)}{n_t(u)+n_t(v)} \|\widehat{p}_t(\cdot|s,a) - p(\cdot|s,a)\|_1\,.\end{aligned}$$

Hence, when $\mathcal{E}$ occurs, $\left\|p^{\boldsymbol{\sigma}_{u\cup v}}(\cdot|u\cup v) - \widehat{p}_t^{\boldsymbol{\sigma}_{u\cup v,t}}(\cdot|u\cup v)\right\|_1 \leqslant \varepsilon_{u\cup v,t}$, so that

$$\left\|\widehat{p}_t(\sigma_{\ell,t}(\cdot)|\ell) - \widehat{p}_t^{\boldsymbol{\sigma}_{u\cup v,t}}(\cdot|u\cup v)\right\|_1 \leqslant d(\{\ell\}, u\cup v) + \beta_{N_t(\ell)}\big(\tfrac{\delta}{SA}\big) + \varepsilon_{u\cup v,t}\,.$$

■

We are now ready to prove the proposition.

Proof (of Proposition 13) Fix $t \geqslant 1$, and consider $\alpha \to \infty$ (the choice $\alpha \geqslant \frac{t}{\max\{1, f^{-1}(\Delta)\}}$ suffices). Assume that $\min_{s,a} N_t(s,a) > f^{-1}(\Delta)$, and that $\mathcal{E}$ holds. By Lemma 1, we have that at any round of the algorithm, the set of PAC Neighbors of a given $v \in \mathcal{S}\times\mathcal{A}$ maintained by the algorithm contains some $u \in \mathcal{S}\times\mathcal{A}$ belonging to the same class as v.

We prove the theorem by induction. First we show that the best case holds, that is in the first iteration of the algorithm, (i) the algorithm avoids grouping state-action pairs belonging to different classes; and (ii) the algorithm groups all the pairs in the same class. Initially, all the classes are singletons. So in the first iteration, the algorithm starts with the classes sorted according to a non-increasing order of number of samples, and then iteratively merges each class with its PAC Nearest Neighbor (see Definition 10). Recall that for a partition $\mathcal{C}$, $\mathsf{Near}(c,\mathcal{C})$ denotes the PAC Nearest Neighbor of $\mathcal{C}$: $\mathsf{Near}(c,\mathcal{C}) \in \operatorname{argmin}_{x\in\mathcal{N}(c)} \widehat{d}(c,x)$. In the first round the algorithm, if $i,j \in \mathcal{S}\times\mathcal{A}$ are combined, then $\widehat{d}(\{i\},\{j\}) \leqslant 0$. In view of the definition of $\widehat{d}(\cdot,\cdot)$, we deduce that

$$\left\|\widehat{p}_t(\sigma_{i,t}(\cdot)|i) - \widehat{p}_t(\sigma_{j,t}(\cdot)|j)\right\|_1 - \beta_{N_t(i)}\big(\tfrac{\delta}{SA}\big) - \beta_{N_t(j)}\big(\tfrac{\delta}{SA}\big) \leqslant 0\,. \tag{2}$$

In order to show that the algorithm makes no mistake, we need to show that $d(\{i\},\{j\}) = 0$. We have

$$\begin{aligned}
d(\{i\},\{j\}) &= \|p(\sigma_i(\cdot)|i) - p(\sigma_j(\cdot)|j)\|_1 \\
&\leqslant \|p(\sigma_i(\cdot)|i) - \widehat{p}_t(\sigma_i(\cdot)|i)\|_1 + \|p(\sigma_j(\cdot)|j) - \widehat{p}_t(\sigma_{j,t}(\cdot)|j)\|_1 + \|\widehat{p}_t(\sigma_{i,t}(\cdot)|i) - \widehat{p}_t(\sigma_{j,t}(\cdot)|j)\|_1 \\
&\leqslant \|p(\sigma_i(\cdot)|i) - \widehat{p}_t(\sigma_i(\cdot)|i)\|_1 + \|p(\sigma_j(\cdot)|j) - \widehat{p}_t(\sigma_j(\cdot)|j)\|_1 + \|\widehat{p}_t(\sigma_{i,t}(\cdot)|i) - \widehat{p}_t(\sigma_{j,t}(\cdot)|j)\|_1 \\
&= \|p(\cdot|i) - \widehat{p}_t(\cdot|i)\|_1 + \|p(\cdot|j) - \widehat{p}_t(\cdot|j)\|_1 + \|\widehat{p}_t(\sigma_{i,t}(\cdot)|i) - \widehat{p}_t(\sigma_{j,t}(\cdot)|j)\|_1\,,
\end{aligned}$$

where the first inequality follows from the sub-additivity of the norm function, and the second follows from Lemma 7. Hence, under the event $\mathcal{E}$, it holds that

$$d(\{i\},\{j\}) \leqslant \beta_{N_t(i)}\big(\tfrac{\delta}{SA}\big) + \beta_{N_t(j)}\big(\tfrac{\delta}{SA}\big) + \|\widehat{p}_t(\sigma_{i,t}(\cdot)|i) - \widehat{p}_t(\sigma_{j,t}(\cdot)|j)\|_1\,.$$

Combining this with (2), we have under $\mathcal{E}$,

$$d(\{i\},\{j\}) \leqslant 2\beta_{N_t(i)}\big(\tfrac{\delta}{SA}\big) + 2\beta_{N_t(j)}\big(\tfrac{\delta}{SA}\big)\,.$$

In view of the assumption $\min_{s,a} N_t(s,a) > f^{-1}(\Delta)$, and noting that $d(\{i\},\{j\}) \geqslant \Delta$, we deduce that $d(\{i\},\{j\}) \leqslant 0$, so that the base case holds.

Now assume that at the end of iteration m, the algorithm outputs a valid partition under $\mathcal{E}$, namely, it does not wrongly group pairs coming from different classes. We would like to show that the partition obtained in iteration $m+1$ is valid, too. To this end, consider

$u, v \in \mathcal{C}^m$ that are merged by the algorithm in round $m+1$, so that $u \cup v \in \mathcal{C}^{m+1}$. First note that by Lemma 1, for any $v \in \mathcal{C}^m$, there exists $u' \in \mathcal{N}(v)$ with $d(u', v) = 0$. We need to show that $d(u, v) = 0$. By construction, $u = \mathsf{Near}(v, \mathcal{C}^m)$, and so the following inequalities hold:

$$\begin{aligned}
&\big\|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot|u) - \widehat{p}_t^{\boldsymbol{\sigma}_{v,t}}(\cdot|v)\big\|_1 - \varepsilon_{u,t} - \varepsilon_{v,t} \leqslant 0\,;\\
&\big\|\widehat{p}_t(\sigma_{i,t}(\cdot)|i) - \widehat{p}_t(\sigma_{j,t}(\cdot)|j)\big\|_1 - \beta_{N_t(i)}\big(\tfrac{\delta}{SA}\big) - \beta_{N_t(j)}\big(\tfrac{\delta}{SA}\big) \leqslant 0\,, \quad \forall i \in u, \forall j \in v;\\
&\big\|\widehat{p}_t(\sigma_{\ell,t}(\cdot)|\ell) - \widehat{p}_t^{\boldsymbol{\sigma}_{u\cup v,t}}(\cdot|u \cup v)\big\|_1 - \beta_{N_t(\ell)}\big(\tfrac{\delta}{SA}\big) - \varepsilon_{u\cup v,t} \leqslant 0\,, \quad \forall \ell \in u \cup v.
\end{aligned}$$

Using similar steps as in the proof of Lemma 1, it follows that

$$d(u, v) \leqslant \|p^{\boldsymbol{\sigma}_u}(\cdot|u) - \widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot|u)\|_1 + \|p^{\boldsymbol{\sigma}_v}(\cdot|v) - \widehat{p}_t^{\boldsymbol{\sigma}_{v,t}}(\cdot|v)\|_1 + \|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot)|u) - \widehat{p}_t^{\boldsymbol{\sigma}_{v,t}}(\cdot|v)\|_1\,.$$

Using (1) in the proof of Lemma 1, we arrive at

$$\begin{aligned}
d(u, v) &\leqslant \varepsilon_{u,t} + \varepsilon_{v,t} + \big\|\widehat{p}_t^{\boldsymbol{\sigma}_{u,t}}(\cdot|u) - \widehat{p}_t^{\boldsymbol{\sigma}_{v,t}}(\cdot|v)\big\|_1\\
&\leqslant 2\varepsilon_{u,t} + 2\varepsilon_{v,t} < 4\beta_{f^{-1}(\Delta)}\big(\tfrac{\delta}{SA}\big)\,.
\end{aligned}$$

We thus deduce that $d(u, v) = 0$, which concludes the proof. ■

Appendix D. Regret Analysis of C-UCRL$(\mathcal{C}, \sigma)$: Proof of Theorem 13

In this section, we prove Theorem 13, which provides an upper bound on the regret of C-UCRL$(\mathcal{C}, \boldsymbol{\sigma})$. We provide the proof for the case when the reward function is unknown to the learner too. Our proof follows similar lines as in the proof of (Jaksch et al., 2010, Theorem 2). We first provide the following time-uniform concentration inequality to control a bounded martingale difference sequence, which follows from time-uniform Laplace concentration inequality:

Lemma 2 (Time-uniform Azuma-Hoeffding) *Let $(X_t)_{t\geqslant 1}$ be a martingale difference sequence bounded by b for some $b > 0$ (that is, $|X_t| \leqslant b$ for all t). Then, for all $\delta \in (0, 1)$,*

$$\mathbb{P}\Big(\exists T \in \mathbb{N} : \sum_{t=1}^{T} X_t \geqslant b\sqrt{2(T+1)\log\big(\sqrt{T+1}/\delta\big)}\Big) \leqslant \delta\,.$$

Proof (of Theorem 1) Let $\delta \in (0, 1)$. To simplify notations, we define the short-hand $J_k := J_{t_k}$ for various random variables that are fixed within a given episode k (for example $\mathcal{M}_k := \mathcal{M}_{t_k}$). Denote by $m(T)$ the number of episodes initiated by the algorithm up to time T. An application of Lemma 2 yields:

$$\mathfrak{R}(T) = \sum_{t=1}^{T} g_\star - \sum_{t=1}^{T} r_t(s_t, a_t) \leqslant \sum_{s,a} N_{m(T)}(s, a)(g_\star - \mu(s, a)) + \sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\,,$$

with probability at least $1 - \delta$. We have

$$\begin{aligned}
\sum_{s,a} N_{m(T)}(s, a)(g_\star - \mu(s, a)) &= \sum_{k=1}^{m(T)} \sum_{s,a} \sum_{t=t_k+1}^{t_{k+1}} \mathbb{I}\{s_t = s, a_t = a\}(g_\star - \mu(s, a))\\
&= \sum_{k=1}^{m(T)} \sum_{s,a} \nu_k(s, a)(g_\star - \mu(s, a))\,.
\end{aligned}$$

Defining $\nu_k(c) := \sum_{s,a} \nu_k(s,a)$ for $c \in \mathcal{C}$, we further obtain

$$\sum_{s,a} N_{m(T)}(s,a)(g_\star - \mu(s,a)) = \sum_{k=1}^{m(T)} \sum_{c \in \mathcal{C}} \nu_k(c)(g_\star - \mu(c))\,,$$

where we have used that $\mu(s,a)$ has constant value $\mu(c)$ for all $(s,a) \in c$. For $1 \leqslant k \leqslant m(T)$, we define the regret of episode k as $\Delta_k = \sum_{c \in \mathcal{C}} \nu_k(c)(g_\star - \mu(c))$. Hence, with probability at least $1-\delta$,

$$\mathfrak{R}(T) \leqslant \sum_{k=1}^{m(T)} \Delta_k + \sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta)}\,.$$

We say an episode is *good* if $M \in \mathcal{M}_k$ (that is, the set $\mathcal{M}_k$ of plausible MDPs contains the true model), and *bad* otherwise.

Control of the regret due to bad episodes ($M \notin \mathcal{M}_k$). Due to using time-uniform instead of time-instantaneous confidence bounds, we can show that with high probability, all episodes are good for $T \in \mathbb{N}$. More precisely, with probability higher than $1-2\delta$, for all T, bad episodes do not contribute to the regret:

$$\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \notin \mathcal{M}_k\} = 0\,.$$

Control of the regret due to good episodes ($M \in \mathcal{M}_k$). We closely follow (Jaksch et al., 2010) and decompose the regret to control the transition and reward functions. At a high level, we make two major modifications as follows. (i) We use the time-uniform bound stated in Lemma 2 to control the martingale difference sequence that appears; and (ii) as the stopping criterion of C-UCRL($\mathcal{C}, \boldsymbol{\sigma}$) slightly differs from that of UCRL2, we use the following lemma to control the number $m(T)$ of episodes:

Lemma 3 (Number of episodes) *The number $m(T)$ of episodes of C-UCRL($\mathcal{C}, \boldsymbol{\sigma}$) up to time $T \geqslant C$ is upper bounded by:*

$$m(T) \leqslant C \log_2(\tfrac{8T}{C})\,.$$

Consider a good episode k (hence, $M \in \mathcal{M}_k$). The EVI algorithm outputs a policy π_k^+ and $\widetilde{M}_k$ satisfying $g_{\pi_k^+}^{\widetilde{M}_k} \geqslant g_\star - \frac{1}{\sqrt{t_k}}$. Let us define $g_k := g_{\pi_k^+}^{\widetilde{M}_k}$. It then follows that

$$\Delta_k = \sum_{c \in \mathcal{C}} \nu_k(c)(g_\star - \mu(c)) \leqslant \sum_{c \in \mathcal{C}} \nu_k(c)(g_k - \mu(c)) + \sum_{c \in \mathcal{C}} \frac{\nu_k(c)}{\sqrt{t_k}}\,. \tag{3}$$

Using the same argument as in the proof of (Jaksch et al., 2010, Theorem 2), the value function $u_k^{(i)}$ computed by EVI at the last iteration i satisfies: $\max_s u_k^{(i)}(s) - \min_s u_k^{(i)}(s) \leqslant D$. Moreover, the convergence criterion of EVI implies

$$|u_k^{(i+1)}(s) - u_k^{(i)}(s) - g_k| \leqslant \frac{1}{\sqrt{t_k}}\,, \qquad \forall s \in \mathcal{S}\,. \tag{4}$$

By the design of EVI, we have $u_k^{(i+1)}(s) = \widetilde{\mu}_k(s,\pi_k^+(s)) + \sum_x \widetilde{p}_k(x|s,\pi_k^+(s))u_k^{(i)}(x)$. Substituting this into (4) gives

$$\left|\Big(g_k - \widetilde{\mu}_k(s,\pi_k^+(s))\Big) - \Big(\sum_x \widetilde{p}_k(x|s,\pi_k^+(s))u_k^{(i)}(x) - u_k^{(i)}(s)\Big)\right| \leqslant \frac{1}{\sqrt{t_k}}, \qquad \forall s \in \mathcal{S}.$$

Defining $\mathbf{g}_k = g_k\mathbf{1}$, $\widetilde{\boldsymbol{\mu}}_k := \big(\widetilde{\mu}_k(s,\pi_k^+(s))\big)_{s\in\mathcal{S}}$, $\widetilde{\mathbf{P}}_k := \big(\widetilde{p}_k(x|s,\pi_k^+(s))\big)_{s,x\in\mathcal{S}}$, and $\nu_k := \big(\nu_k(s,\pi_k^+(s))_{s\in\mathcal{S}}$, we can rewrite the above inequality as:

$$\left|\mathbf{g}_k - \widetilde{\boldsymbol{\mu}}_k - (\widetilde{\mathbf{P}}_k - \mathbf{I})u_k^{(i)}\right| \leqslant \frac{1}{\sqrt{t_k}}\mathbf{1}.$$

Combining this with (3) yields

$$\begin{aligned}
\Delta_k &\leqslant \sum_{s,a}\nu_k(s,a)(g_k - \mu(s,a)) + \sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{t_k}} \\
&= \sum_{s,a}\nu_k(s,a)(g_k - \widetilde{\mu}_k(s,a)) + \sum_{s,a}\nu_k(s,a)(\widetilde{\mu}_k(s,a) - \mu(s,a)) + \sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{t_k}} \\
&\leqslant \nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})u_k^{(i)} + \sum_{s,a}\nu_k(s,a)(\widetilde{\mu}_k(s,a) - \mu(s,a)) + 2\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{t_k}}.
\end{aligned}$$

Similarly to (Jaksch et al., 2010), we define $w_k(s) := u_k^{(i)}(s) - \frac{1}{2}(\min_s u_k^{(i)}(s) + \max_s u_k^{(i)}(s))$ for all $s \in \mathcal{S}$. Then, in view of the fact that $\widetilde{\mathbf{P}}_k$ is row-stochastic, we obtain

$$\Delta_k \leqslant \nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k + \sum_{s,a}\nu_k(s,a)(\widetilde{\mu}_k(s,a) - \mu(s,a)) + 2\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{t_k}}. \tag{5}$$

The second term in the right-hand side can be upper bounded as follows. Fix pair (s,a) and let $c_{s,a}$ denote the cluster to which (s,a) belongs. The fact $M \in \mathcal{M}_k$ implies

$$\begin{aligned}
\widetilde{\mu}_k(s,a) - \mu(s,a) &\leqslant |\widetilde{\mu}_k(s,a) - \widehat{\mu}_k(s,a)| + |\widehat{\mu}_k(s,a) - \mu(s,a)| \leqslant 2\beta'_{n_k(c_{s,a})}(\tfrac{\delta}{C}) \\
&= 2\sqrt{\frac{1}{2n_k(c_{s,a})}\Big(1+\frac{1}{n_k(c_{s,a})}\Big)\log\Big(C\sqrt{n_k(c_{s,a})+1}/\delta\Big)} \\
&\leqslant 2\sqrt{\frac{1}{n_k(c_{s,a})}\log\left(C\sqrt{T+1}/\delta\right)},
\end{aligned}$$

where we have used $1 \leqslant n_k(c_{s,a}) \leqslant T$ in the last inequality. Using this bound and noting that $t_k \geqslant n_k(c)$, we obtain

$$\Delta_k \leqslant \nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k + 2\Big(\sqrt{\log\left(C\sqrt{T+1}/\delta\right)} + 1\Big)\sum_{c\in\mathcal{C}}\frac{\nu_k(c)}{\sqrt{n_k(c)}}. \tag{6}$$

In what follows, we derive an upper bound on $\nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k$. Similarly to (Jaksch et al., 2010), we consider the following decomposition:

$$\nu_k(\widetilde{\mathbf{P}}_k - \mathbf{I})w_k = \underbrace{\nu_k(\widetilde{\mathbf{P}}_k - \mathbf{P}_k)w_k}_{L_1(k)} + \underbrace{\nu_k(\mathbf{P}_k - \mathbf{I})w_k}_{L_2(k)}.$$

Noting that $\|w_k\|_\infty \leqslant \frac{D}{2}$, we upper bound $L_1(k)$ as follows:

$$\begin{aligned} L_1(k) &\leqslant \sum_{s,a} \nu_k(s,a)(\widetilde{p}_k(s'|s,a) - p(s'|s,a))w_k(s') \\ &\leqslant \sum_{s,a} \nu_k(s,a)\|\widetilde{p}_k(\cdot|s,a) - p(\cdot|s,a)\|_1 \|w_k\|_\infty \\ &\leqslant D\sum_{s,a} \nu_k(s,a)\beta_{n_k(c_{s,a})}(\tfrac{\delta}{C}) \\ &= D\sum_{c\in\mathcal{C}} \nu_k(c)\beta_{n_k(c)}(\tfrac{\delta}{C}) \\ &\leqslant 2D\sqrt{\log\left(C2^S\sqrt{T+1}/\delta\right)} \sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{n_k(c)}}\,. \end{aligned} \tag{7}$$

To upper bound $L_2(k)$, similarly to the proof of (Jaksch et al., 2010, Theorem 2), we define the sequence $(X_t)_{t\geqslant 1}$, with $X_t := (p(\cdot|s_t,a_t) - \mathbf{e}_{s_{t+1}})w_{k_t}\mathbb{I}\{M \in \mathcal{M}_{k_t}\}$, for all t, where k_t denotes the episode containing step t. Note that $\mathbb{E}[X_t|s_1,a_1,\ldots,s_t,a_t] = 0$, so $(X_t)_{t\geqslant 1}$ is martingale difference sequence. Furthermore, $|X_t| \leqslant D$: Indeed, for all t, by the Hölder inequality,

$$|X_t| \leqslant \|p(\cdot|s_t,a_t) - \mathbf{e}_{s_{t+1}}\|_1 \|w_{k(t)}\|_\infty \leqslant \Big(\|p(\cdot|s_t,a_t)\|_1 + \|\mathbf{e}_{s_{t+1}}\|_1\Big)\frac{D}{2} = D\,.$$

Using similar steps as in (Jaksch et al., 2010), for any k with $M \in \mathcal{M}_k$, we have that:

$$L_2(k) \leqslant \sum_{t=t_k}^{t_{k+1}-1} X_t + D\,,$$

so that $\sum_{k=1}^{m(T)} L_2(k) \leqslant \sum_{t=1}^{T} X_t + m(T)D$. Therefore, by Lemma 2, we deduce that with probability at least $1-\delta$,

$$\begin{aligned} \sum_{k=1}^{m(T)} L_2(k) &\leqslant D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)} + m(T)D \\ &\leqslant D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)} + DC\log_2(\tfrac{8T}{C})\,, \end{aligned} \tag{8}$$

where the last step follows from Lemma 3.

Final control. Combing (6)–(8) and summing over all episodes give:

$$\begin{aligned} \sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \in \mathcal{M}_k\} &\leqslant \sum_{k=1}^{m(T)} L_1(k) + \sum_{k=1}^{m(T)} L_2(k) + 2\Big(\sqrt{\log\left(C\sqrt{T+1}/\delta\right)} + 1\Big) \sum_{k=1}^{m(T)} \sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{n_k(c)}} \\ &\leqslant 2\Big(D\sqrt{\log\left(C2^S\sqrt{T+1}/\delta\right)} + \sqrt{\log\left(C\sqrt{T+1}/\delta\right)} + 1\Big) \sum_{k=1}^{m(T)} \sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{n_k(c)}} \\ &\quad + D\sqrt{2(T+1)\log\left(\sqrt{T+1}/\delta\right)} + DC\log_2(\tfrac{8T}{C})\,, \end{aligned} \tag{9}$$

with probability at least $1-\delta$. To upper bound the right-hand side, we recall the following lemma:

Lemma 4 ((Jaksch et al., 2010, Lemma 19)) *For any sequence of numbers $z_1, z_2, \dots, z_n$ with $0 \leqslant z_k \leqslant Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$,*

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \leqslant (\sqrt{2}+1)\sqrt{Z_n}\,.$$

Note that $n_k(c) = \sum_{k'<k} \nu_{k'}(c)$. Hence, applying Lemma 4 gives

$$\sum_{c\in\mathcal{C}} \sum_{k=1}^{m(T)} \frac{\nu_k(c)}{\sqrt{n_k(c)}} \leqslant \sum_{c\in\mathcal{C}} (\sqrt{2}+1)\sqrt{n_{m(T)}(c)} \leq (\sqrt{2}+1)\sqrt{CT}\,,$$

where the last step follows from Jensen's inequality and $\sum_c n_{m(T)}(c) = T$. Therefore,

$$\begin{aligned}\sum_{k=1}^{m(T)} \Delta_k \mathbb{I}\{M \in \mathcal{M}_k\} &\leqslant D\sqrt{2(T+1)\log\left(\sqrt{T+1}/\delta\right)} + DC\log_2(\tfrac{8T}{C}) \\ &\quad + 2(\sqrt{2}+1)\Big(D\sqrt{\log\left(C2^S\sqrt{T+1}/\delta\right)} + \sqrt{\log\left(C\sqrt{T+1}/\delta\right)} + 1\Big)\sqrt{CT}\,,\end{aligned}$$

with probability of at least $1-\delta$. Finally, the regret of C-UCRL$(\mathcal{C}, \sigma)$ is controlled on an event of probability higher than $1 - 2\delta - \delta - \delta$, uniformly over all T, by

$$\begin{aligned}\mathfrak{R}(T) &\leqslant 2(\sqrt{2}+1)\Big(D\sqrt{\log\left(C2^S\sqrt{T+1}/\delta\right)} + \sqrt{\log\left(C\sqrt{T+1}/\delta\right)} + 1\Big)\sqrt{CT} \\ &\quad + D\sqrt{2(T+1)\log\left(\sqrt{T+1}/\delta\right)} + DC\log_2(\tfrac{8T}{C}) + \sqrt{\tfrac{1}{2}(T+1)\log\left(\sqrt{T+1}/\delta\right)} \\ &\leqslant 18\sqrt{CT(S + \log(C\sqrt{T+1}/\delta))} + DC\log_2(\tfrac{8T}{C})\,,\end{aligned}$$

thus completing the proof. We finally note that when the mean reward function is known, as in the main text, the above bound holds with a probability higher than $1-3\delta$. ■

D.1. Proof of Lemma 3

The proof uses similar steps as in the proof of Proposition 18 in (Jaksch et al., 2010).

Recall that given c, $N_T(c)$ and $\nu_k(c) := \nu_{t_k}(c)$ denote as the total number of state-action observations, up to step T and in episode k, respectively. For any c, let $K(c)$ denote the number of episodes where a state-action pair from c is sampled: $K(c) = \sum_{k=1}^{m(T)} \mathbb{I}\{\nu_k(c) > 0\}$. It is worth mentioning that if $n_k(c) > 0$ and $\nu_k(c) = n_k(c)$, by the design of the algorithm, $n_{k+1}(c) = 2n_k(c)$. Hence,

$$n_{m(T)}(c) = \sum_{k=1}^{m(T)} \nu_k(c) \geqslant 1 + \sum_{k:\nu_k(c)=n_k(c)} n_k(c) \geqslant 1 + \sum_{i=1}^{K(c)} 2^{i-1} = 2^{K(c)}\,.$$

If $n_{m(T)}(c) = 0$, then $K(c) = 0$, so that $n_{m(T)}(c) \geqslant 2^{K(c)} - 1$ for all c. Thus,

$$T = \sum_{c\in\mathcal{C}} n_{m(T)}(c) \geqslant \sum_{c\in\mathcal{C}} (2^{K(c)} - 1)$$

On the other hand, an episode has happened when either $n_k(c) = 0$ or $n_k(c) = \nu_k(c)$. Therefore, $m(T) \leqslant 1 + C + \sum_{c \in \mathcal{C}} K(c)$ and consequently, $\sum_{c \in \mathcal{C}} K(c) \geqslant m(T) - 1 - C$. Hence, by Jensen's inequality, we obtain

$$\sum_{c \in \mathcal{C}} 2^{K(c)} \geqslant C 2^{\sum_{c \in \mathcal{C}} \frac{K(c)}{C}} \geqslant C 2^{\frac{m(T)-1}{C} - 1}.$$

Putting together, we obtain $T \geqslant C\big(2^{\frac{m(T)-1}{C} - 1} - 1\big)$. Therefore,

$$m(T) \leqslant 1 + 2C + C \log_2\left(\tfrac{T}{C}\right) \leqslant 3C + C \log_2\left(\tfrac{T}{C}\right) \leqslant C \log_2\left(\tfrac{8T}{C}\right),$$

thus concluding the proof. ■

Appendix E. Environments Used in Numerical Experiments

In this section, we provide further details for the environments used in numerical experiments in Section 5.

E.1. RiverSwim and Ergodic RiverSwim

In the first set of experiments, we examined the performance of various algorithms in *RiverSwim* environments. Figures 2 and 1 respectively display the L-state *RiverSwim* and ergodic *RiverSwim* environments.

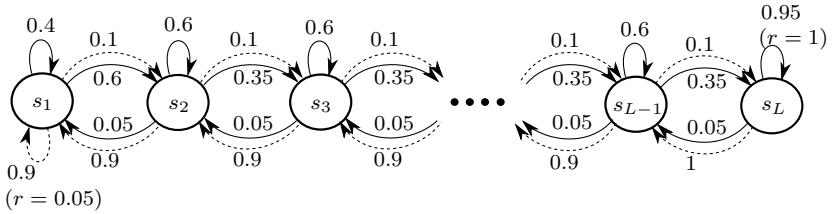

Figure 1: The L-state *Ergodic RiverSwim* MDP

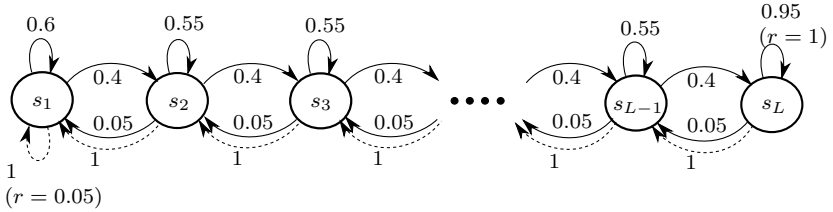

Figure 2: The L-state *RiverSwim* MDP

E.2. Grid-World

We conducted our last set of experiments in a 7×7 grid-world environment shown in Figure 3, which we refer to as the 4-room grid-world. This MDP comprises 20 states ($S = 20$). In

this environment, the initial state is the upper-left corner (shown in red). When the learner reaches the lower-right corner (shown in yellow), a reward of 1 is given, and the learner is sent back to the initial state. The learner can perform four actions ($A = 4$): Going up, left, down, or right. After playing a given action, the learner stays in the same state with probability 0.1, moves to the desired direction with probability 0.7 (for example, to the left, if the learner chooses to 'go left'), and moves to other possible directions with probability 0.2. Walls act as *reflectors*: When the next state is a wall, the transition probability of it is added to that of the current state.

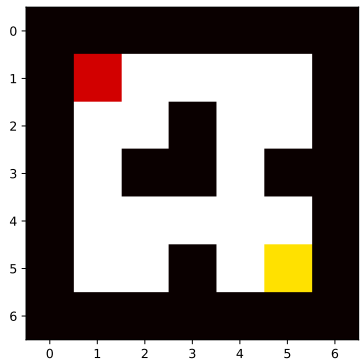

Figure 3: The 4-room grid-world MDP

Appendix F. Other Examples of MDPs

In this section, we examine the notion of similarity presented in Definition 1 on some grid-world environments.

For this purpose, we consider grid-world MDPs. The action-space is $\{u, d, l, r\}$. Playing action $a = u$ moves the current state 'up' with probability 0.8, does not change the current state with probability 0.1, and moves left or right with the same probability 0.05. Walls act as *reflectors*: When the next state is a wall, the transition probability of it is added to that of the current state. Other actions are defined in a similar way. Finally, the goal-state is put in the bottom-right corner of the MDP, where the learner is given a reward of 1.

Below, we show four examples of grid-world environments defined according to the above scheme, with different numbers of state-action pairs. The number of state-action pairs in the introduced 4-room and 2-room MDPs changes as the grid size grows, while keeping the number of classes almost fixed:

grid-world	Figure 4	Figure 5	Figure 6	Figure 7
SA	84	800	736	$\sim 10^4$
C	6	6	7	7

Environment	States	5×5	7×7	9×9	100×100
4-Room	SA	100	196	324	4×10^4
4-Room	C	3	3	3	3
2-Room	SA	100	196	324	4×10^4
2-Room	C	4	4	4	4

We stress that other notions of similarity from the RL literature do not scale well. For instance, in (Ortner, 2013), a partition $\mathcal{S}_1, \dots \mathcal{S}_n$ of the state-space $\mathcal{S}$ is considered to define an aggregated MDP, which satisfies: For all $i \in \{1, \dots, n\}$,

$$\begin{aligned} \forall s, s' \in \mathcal{S}_i, \forall a \in \mathcal{A}, \qquad & \mu(s,a) = \mu(s',a)\,, \\ \forall j, \qquad & \textstyle\sum_{s'' \in \mathcal{S}_j} p(s''|s,a) = \sum_{s'' \in \mathcal{S}_j} p(s''|s',a)\,. \end{aligned}$$

This readily prevents any two states s, s' such that $p(\cdot|s,a)$ and $p(\cdot|s',a)$ have disjoint supports from being in the same set $\mathcal{S}_i$. Thus, since in a grid-world MDP, where transitions are local, the number of pairs with disjoint support is (almost linearly) increasing with S, this implies a potentially large number of classes for grid-worlds with many states. A similar criticism can be formulated for (Anand et al., 2015), even though it considers sets of state-action pairs instead of states only, thus slightly reducing the total number of classes.

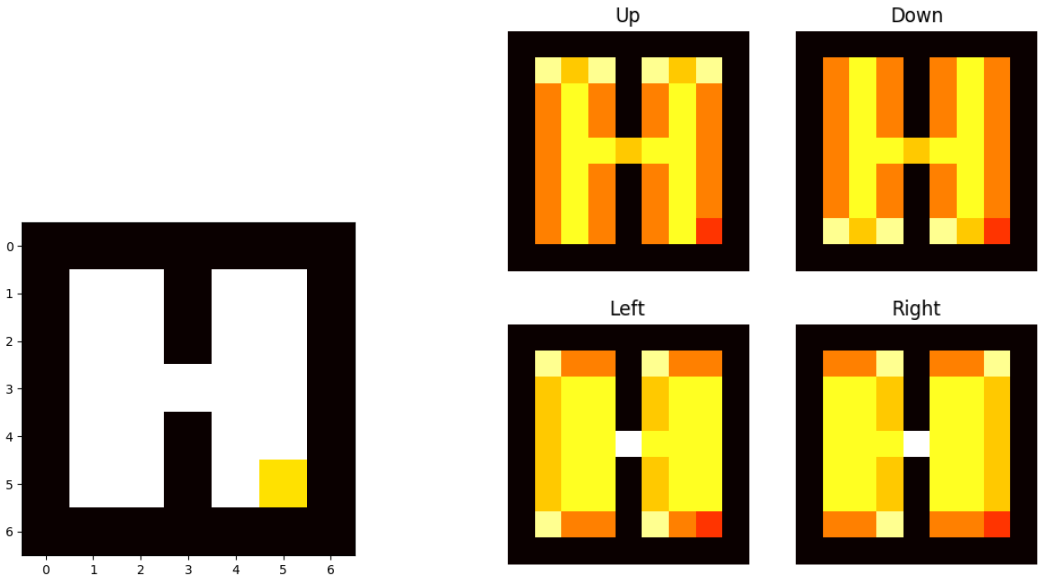

Figure 4: Left: Two-room grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

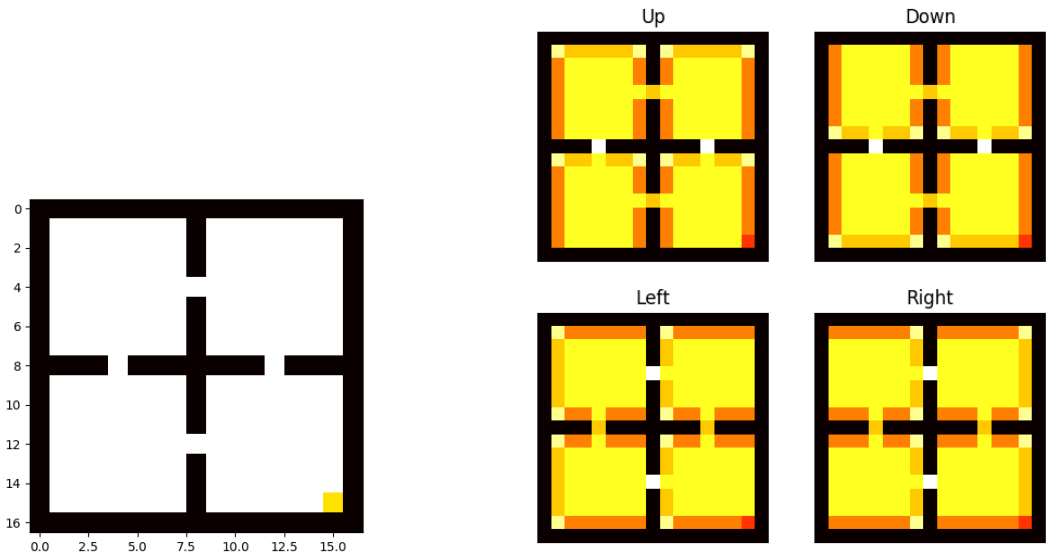

Figure 5: Left: Four-room grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

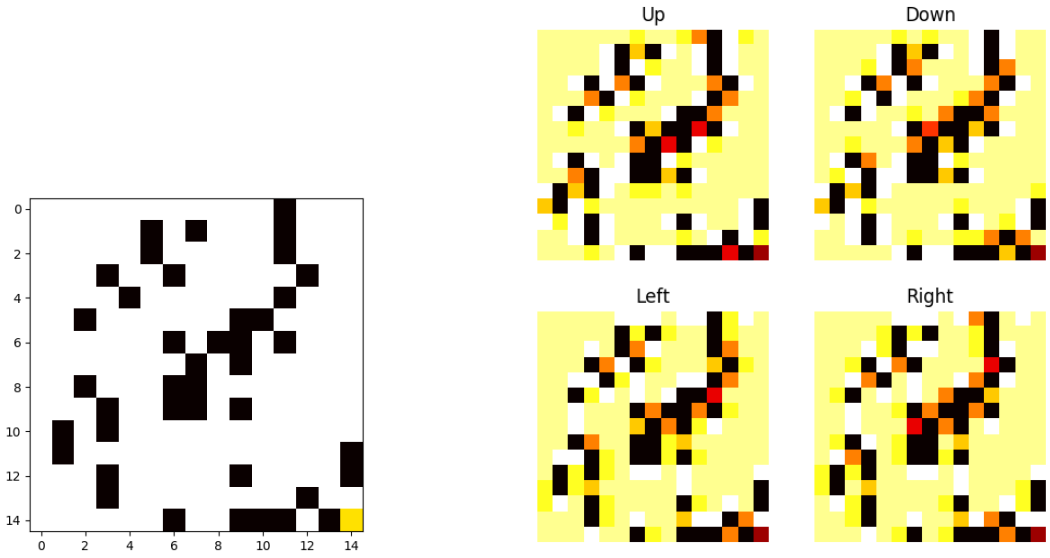

Figure 6: Left: A more complex grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

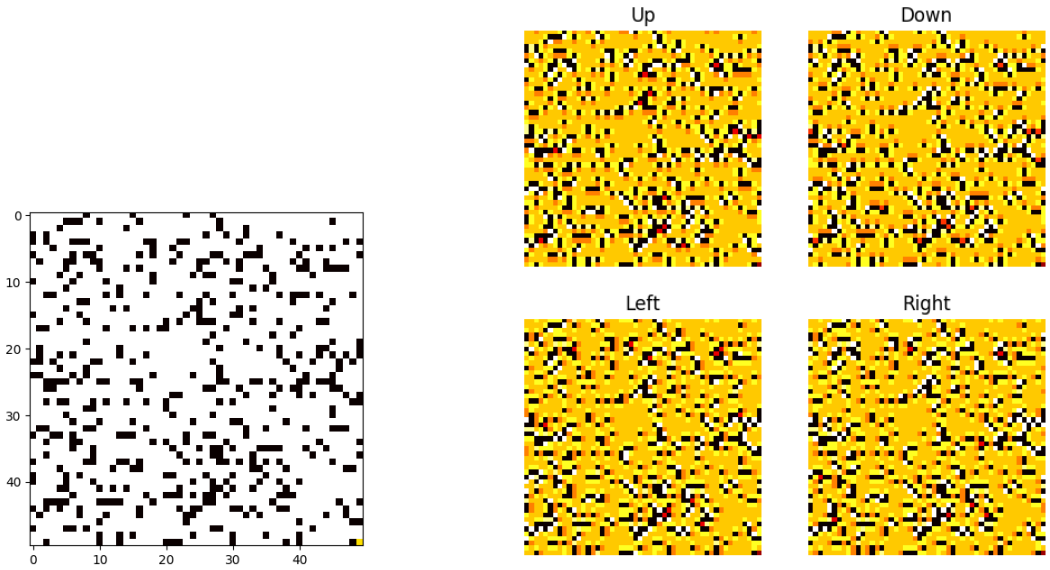

Figure 7: Left: A more complex grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

References

A. Anand, A. Grover, Mausam, and P. Singla. ASAP-UCT: Abstraction of state-action pairs in UCT. In *Proc. of IJCAI*, pages 1509–1515, 2015.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *JMLR*, 11:1563–1600, 2010.

R. Ortner. Adaptive aggregation for reinforcement learning in average reward Markov decision processes. *Annals of Operations Research*, 208(1):321–336, 2013.