Upper Confidence Reinforcement Learning exploiting state-action equivalence

Odalric-Ambrym Maillard, Mahsa Asadi

▶ **To cite this version:**

Odalric-Ambrym Maillard, Mahsa Asadi. Upper Confidence Reinforcement Learning exploiting state-action equivalence. 2018. hal-01945034

HAL Id: hal-01945034
https://hal.science/hal-01945034v1

Preprint submitted on 5 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Upper Confidence Reinforcement Learning exploiting state-action equivalence

Mahsa Asadi
Inria Lille – Nord Europe
Villeneuve d'Ascq, France
mahsa.asadi@inria.fr

Odalric-Ambrym Maillard
Inria Lille – Nord Europe
Villeneuve d'Ascq, France
odalricambrym.maillard@inria.fr

Abstract

Leveraging an equivalence property on the set of states of state-action pairs in an Markov Decision Process (MDP) has been suggested by many authors. We take the study of equivalence classes to the reinforcement learning (RL) setup, when transition distributions are no longer assumed to be known, in a discrete MDP with average reward criterion and no reset. We study powerful similarities between state-action pairs related to optimal transport. We first analyze a variant of the UCRL2 algorithm called C-UCRL2, which highlights the clear benefit of leveraging this equivalence structure when it is known ahead of time: the regret bound scales as $\tilde{O}(D\sqrt{KCT})$ where C is the number of classes of equivalent state-action pairs and K bounds the size of the support of the transitions. A non trivial question is whether this benefit can still be observed when the structure is unknown and must be learned while minimizing the regret. We propose a sound clustering technique that provably learn the unknown classes, but show that its natural combination with UCRL2 empirically fails. Our findings suggests this is due to the ad-hoc criterion for stopping the episodes in UCRL2. We replace it with hypothesis testing, which in turns considerably improves all strategies. It is then empirically validated that learning the structure can be beneficial in a full-blown RL problem.

1 Introduction

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \nu)$ be an (undiscounted) MDP where $\mathcal{S}$ denotes the discrete state space, $\mathcal{A}$ the discrete action space, p the transition kernel such that $p(s'|s,a)$ denotes the probability of transiting to state s', starting from state s and executing action a, and ν is a reward distribution function on $\mathcal{R} = [0,1]$ with mean function denoted μ.
Many MDPs exhibit strong similarities between different transitions. This is typically the case in a grid-world MDP when taking action Up from state s or right from state s' when both are far from any wall results in similar transitions (typically, move to the target state with probability p and stay still or transit to other neighbors with the remaining probability), see Figure 1.

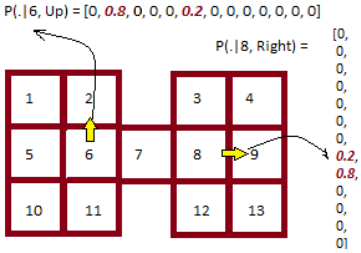

Figure 1: A grid-world MDP showing similar transitions from state-action pairs (6,up) and (8,right).

Leveraging an equivalence structure is popular in the MDP literature (see Abel et al. (2016), Li et al. (2006) or Ravindran & Barto (2004)). However, most notions are unfortunately not well adapted to a reinforcement learning (RL) setup when the transition probabilities are unknown to the learner.

Preprint. Work in progress.

We further want not only to compute a near-optimal policy but to minimize the cumulative regret incurred while interacting with the MDP, in a *never ending single stream* of observations, against an optimal policy for the *average regret* criterion. This problem has received a lot of attention in the recent literature after the seminal work of Jaksch et al. (2010) introduced the UCRL2 algorithm that uses the *optimism in face of uncertainty* principle coming from the multi-armed bandit literature Auer et al. (2002) presents a regret guarantee of $\tilde{O}(DS\sqrt{AT})$ after T steps for any unknown MDP with S states, A actions per state, and diameter D. Obviously, the regret is dependent on the number of states and actions and increases as these parameters grow. Our goal in this paper is to leverage similarity structures to reduce this dependency. We consider to that end the following key definition:

Definition 1 (Similar state-action pairs) *The pair (s', a') is ε-similar to the pair (s, a), for $\varepsilon = (\varepsilon_p, \varepsilon_\mu) \in \mathbb{R}^2_+$, if* $\|p(\sigma_{s,a}(\cdot)|s,a) - p(\sigma_{s',a'}(\cdot)|s',a')\|_1 \leqslant \varepsilon_p$, *(similar profile)*
and $|\mu(s,a) - \mu(s',a')| \leqslant \varepsilon_\mu$, *(similar rewards)*
where $\sigma_{s,a} : \{1,\dots,S\} \to \mathcal{S}$ indexes a permutation of states such that $p(\sigma_{s,a}(1)|s,a) \geqslant p(\sigma_{s,a}(2)|s,a) \geqslant \dots \geqslant p(\sigma_{s,a}(S)|s,a)$. We call it a profile *mapping.*

Remark 1 *The similarity is not only stated about states, but about state-action pairs.*

Remark 2 *$(0,0)$-similarity is an equivalence relation. It thus induces a canonical partition of $\mathcal{S} \times \mathcal{A}$, which we denote $\mathcal{C}$. In appendix A, we show that in typical grid-world MDPs, the number of classes of state-action pairs using Definition 1 stays small even for large SA. This is no longer the case without the ordering. Also the ordering makes the notion more robust, see Lemma 2.*

Remark 3 *The profile mapping $\sigma_{s,a}$ is not unique in general, especially if distributions are sparse. We assume in the sequel that the restriction of $\sigma_{s,a}$ to the support of $p(\cdot|s,a)$ is uniquely defined.*

The goal of this paper is to adapt the UCRL2 strategy to take advantage of this structure which outperforms it when acting in MDPs with few number of classes and large number of state and actions. We do so by aggregating the information of similar state-action pairs when estimating MDP's transitions.

Literature overview The literature relevant to our standpoint can be categorized into two main parts. First, the rich literature on state-aggregation and state-abstraction. We refer to Abel et al. (2016) for a good survey of recent approaches, Li et al. (2006) on earlier methods, Ravindran & Barto (2004) that introduces interesting definitions but with no algorithm or regret analysis, and Anand et al. (2015), similar to our work in that it considers state-action equivalence, but unlike us does not consider orderings, transition estimation errors or regret analysis; Interesting works revolving around complementary RL questions include the work on selection amongst different state representations in Ortner et al. (2014) or on state-aliasing from Hallak et al. (2013). The most relevant works to our approach are the work of Ortner (2013) on aggregation of states (but not of pairs, and with no ordering) based on concentration inequalities, a path that we follow, and of Dean et al. (1997) considering partitions of the state-space. Further, Ferns et al. (2004) and Ferns et al. (2006) work on the bi-simulation metrics suggests to resort to optimal transport, which is intimately linked with Definition 1 when using the Hamming metric $c(i,j) = \mathbb{I}\{i \neq j\}$ defined on $i, j \in \{1,\dots,S\}$ to define the transport cost (see details in Appendix B). However they consider an MDP, not RL setup. Second, the articles regarding UCRL2-style algorithms for discrete MDPs following Jaksch et al. (2010) and inspired from multi-armed bandits. Let us mention the Regal algorithm Bartlett & Tewari (2009), KL-UCRL that replaces Weissman concentration with a better-behaved Kullback Leibler concentration Filippi et al. (2010), another powerful replacement of Weissman bounds suggested in Maillard et al. (2014), and a recent Thompson sampling approach inspired from bandits Agrawal & Jia (2017). Some analysis have tried to reduce the regret dependency on the number of states, such as in Azar et al. (2017) (restricted to a fixed, known and small horizon), or in Dann et al. (2017), using refined time-uniform concentration inequalities similar to that of Lemma 1 below.
Although the concept of equivalence is well-studied in MDPs, no work seems to have investigated the possibility of defining an aggregation first based on *state-action* pairs instead of states only for *reinforcement learning* problems, and second using optimal transportation maps combined with statistical tests. Especially, the use of *profile* maps seems novel and we show it is also effective.

Outline and contribution We first introduce a similarity measure of state-action pairs based on equivalence of *profile* distributions, see Definition 1. To our knowledge, while other notions of equivalence have been introduced, it is the first time profile (ordering) of distributions is explicitly used in a discrete reinforcement learning (as opposed to MDP) setup.

Section 3, studies the potential benefit of exploiting this structure. We introduce **C**-UCRL2$(\mathcal{C}, \sigma)$, a natural variant of UCRL2 that has access to the equivalence classes. We prove in Theorem 1 that its

regret scales as that of UCRL2, except by replacing a $\sqrt{SA}$ factor with $\sqrt{C}$, where C is the number of classes. More convincingly we provide numerical illustration that this improvement can be massive, reducing by one to several order of magnitudes the accumulated regret.

In Section 4, we move to the more realistic situation when the profiles σ are unknown. We modify **C**-UCRL2$(\mathcal{C}, \sigma)$ to estimate the mappings, provide illustration as well as a novel theoretical regret guarantees thanks to a non-expensive property of the ordering operator (Lemma 2).

Section 5 then deals with the fully agnostic and most challenging scenario, when the partition must be learned from data. While it is intuitive clear that an important regret reduction is achievable when knowing $\mathcal{C}$ and $C \ll SA$, we ask the following non-trivial question: can such an improvement still be observed *without* any prior knowledge on the structure? The cost of learning C could indeed be prohibitive. To answer this question, we first provide an online clustering-based algorithm (agnostic to the number of classes) that provably guarantees valid cluster probability distribution estimations (see Lemma 3). We naturally modify **C**-UCRL2$(\mathcal{C}, \sigma)$ in a sound way and highlight the fact that this strategy still outperforms the vanilla UCRL2 algorithm on experiments, thus showing the advantage of considering the classes. However, we also note that the straightforward derivation of the algorithm from UCRL2 suffers from a much higher regret than its oracle counter part knowing $\mathcal{C}$. While it suggests the price for learning $\mathcal{C}$ is too high, we show the high regret is in fact caused by the stopping criterion of UCRL2, that was actually a trick, as acknowledged by the authors, and reveals to be sub-optimal when handling the unknown classes.

This leads us to the second main contribution of this paper, that is to revisit the stopping criterion used in UCRL2. The old intuition was to "recompute the policy when confidence bounds have changed a lot" following a simple "doubling trick" heuristics. We proposed instead a mathematically more rigorous stopping rule by testing whether "the optimistic model corresponding to the chosen policy is still correct". This very simple idea is illustrated on numerical experiments in section 6, where it proves to be *massively beneficial*, reducing the regret by several order of magnitudes again, even in the most challenging scenario (unknown $\mathcal{C}$). Interestingly, the modification also benefits UCRL2.

While the main goal of the paper is to numerically illustrate the benefit of these two powerful ideas, we provide some theoretical results for soundness of the approach, thus paving the way towards a sharper understanding of regret minimization in discrete structured Markov Decision Processes.

2 UCRL2 setup and notations

Let $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ denote a possibly stochastic policy and $\mu_\pi(s) = \mathbb{E}_{Z\sim\pi(s)}[\mu(s, Z)]$ denote the mean reward after following policy π in state s. Let $p(s'|s, \pi(s)) = \mathbb{E}_{Z\sim\pi(s)}[p(s'|s, Z)]$ and write $P_\pi f$ to denote the function $s \mapsto \sum_{s'\in\mathcal{S}} p(s'|s, \pi(s))f(s')$.

Definition 2 (Value) *The expected cumulative reward of policy π when run for T steps from initial state s_1 is denoted as:*

$$V_T^\pi(s_1) = \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t)\right] = \sum_{t=1}^{T} (P_\pi^{t-1}\mu_\pi)(s_1)\,.$$

where $a_t \sim \pi(s_t)$, $s_{t+1} \sim p(\cdot|s_t, a_t)$, and $r(s, a) \sim \nu(s, a)$.

Definition 3 (Average gain) *The average transition operator is $\overline{P}_\pi = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} P_\pi^{t-1}$. The average gain g_π is:*

$$g_\pi(s_1) \quad = \quad \lim_{T\to\infty} \frac{1}{T} R_{\pi,T}(s_1) = (\overline{P}_\pi \mu_\pi)(s_1)\,.$$

Definition 3 requires some mild assumption on the MDP for the limits to makes sense. It is shown (see, e.g. Puterman (2014)) that the average gain achieved by executing a stationary policy π in a communicating MDP M is well-defined and does not depend on the initial state, i.e., $g_\pi(s_1) = g_\pi$. For this reason, we restrict to such MDPs in the rest of this paper. Let $\star$ denote an optimal policy, that is such that[1] $g_\star = \max_\pi g_\pi$. The following definition captures the dominant term of the regret of any algorithm, as justified by the first steps of the regret analysis from Jaksch et al. (2010),

Definition 4 (Effective-regret) *we define the effective-regret of any learning algorithm $\mathbb{A}$ after T steps as:*

$$\mathfrak{R}(\mathbb{A}, T) := \sum_{t=1}^{T} g_\star(s_1) - \sum_{t=1}^{T} \mu(s_t, a_t) \quad \textit{where}$$

$a_t = \mathbb{A}(s_t, (\{s_{t'}, a_{t'}, r_{t'}\})_{t'<t})$ and s_1 is the initial state.

[1] The maximum is reached since there are only finitely many deterministic policies.

We now briefly present and revisit the UCRL2 algorithm from Jaksch et al. (2010). At a high level, UCRL2 follows the optimistic principle by trying to compute $\overline{\pi}_t^+ = \operatorname{argmax}_{\pi:\mathcal{S}\to\mathcal{A}} \max\{g_\pi^{\mathcal{M}} : \mathcal{M} \in \mathscr{M}_t\}$ where $g_\pi^{\mathcal{M}}$ is the average-gain for policy π in MDP $\mathcal{M}$, and

$$\mathscr{M}_t = \Big\{(\mathcal{S},\mathcal{A},\tilde{p},\tilde{\nu}) : \forall (s,a)\in\mathcal{S}\times\mathcal{A}, \quad |\mu_{N_t(s,a)}(s,a) - \tilde{\mu}(\cdot|s,a)| \leqslant \tilde{b}_t^H(s,a,\frac{\delta}{2SA})$$
$$\text{and} \quad \|p_{N_t(s,a)}(\cdot|s,a) - \tilde{p}(\cdot|s,a)\|_1 \leqslant \tilde{b}_t^W(s,a,\frac{\delta}{2SA})\Big\}.$$

This is achieved approximately by an Extended Value Iteration (EVI) algorithm that builds a near-optimistic policy π_t^+ and MDP $\mathcal{M}_t^+$ such that $g_{\pi_t^+}^{\mathcal{M}_t^+} \geqslant \max_{\pi,M\in\mathscr{M}_t} g_\pi^{\mathcal{M}} - \frac{1}{\sqrt{t}}$.

Here $\mu_n(s,a)$ denotes the empirical mean built using n i.i.d. rewards from $\nu(s,a)$, $p_n(\cdot|s,a)$ is the empirical distribution built using n i.i.d. observations from $p(\cdot|s,a)$, $N_t(s,a)$ is the total number of observations of state action pair (s,a) up to time t, and finally $\tilde{b}_t^H$ and $\tilde{b}_t^W$ are the two functions

$$\tilde{b}_t^H(s,a,\delta) = \sqrt{\frac{3.5\log(t/\delta)}{N_t(s,a)\vee 1}}, \quad \tilde{b}_t^W(s,a,\delta) = \sqrt{\frac{7S\log(t/S\delta)}{N_t(s,a)\vee 1}},$$

respectively based on a Hoeffding and Weissman inequality where $\vee$ represents max operator. Finally, UCRL2 does not recompute π_t^+ at each time step. Instead, it proceeds in internal episodes $k = 0,\dots$ and computes π_t^+ only at the starting time t_k of each episode, defined as $t_1 = 1$ and for all $k > 1$,

$$t_k = \min\Big\{t > t_{k-1}; \exists s,a\colon n_{t_k:t}(s,a) \geqslant \max\{N_{t_k}(s,a),1\}\Big\},$$

where $n_{t_1:t_2}(s,a)$ denotes the number of observations of state-action pair (s,a) between time t_1 and t_2. We provide the detailed code of UCRL2 in Appendix D for reference.

Remark 4 *The bounds $\tilde{b}^H$ and $\tilde{b}^W$ are obtained from simple union bounds with a slightly loose analysis. A more careful bound using similar arguments suggests to use*

$$b_t^H(s,a,\delta) = \sqrt{\frac{\log(2t^2(t+1)/\delta)}{2N_t(s,a)\vee 1}}, \quad b_t^W(s,a,\delta) = \sqrt{\frac{2\log(t^2(t+1)(2^K-2)/\delta)}{N_t(s,a)\vee 1}},$$

where $K \geqslant |Support(p(\cdot|s,a))|$, instead to insure that $\mathcal{M}\in\mathscr{M}_t$ holds w.p.h. than $1-\delta$ uniformly $\forall t$.

Tighter bounds An easy improvement is to further modify these naive bounds based on a union-bound argument over all time steps with proper time-uniform concentration bounds using self-normalization (following the Laplace method to replace optimization with integration see Peña et al. (2008); Abbasi-Yadkori et al. (2011)), that we provide below for completeness.

Lemma 1 (Time-uniform concentration) *$\forall (s,a) \in \mathcal{S}\times\mathcal{A}$ and any $[0,1]$-bounded distribution with mean $\mu(s,a)$:*

$$\mathbb{P}\Big(\forall t\in\mathbb{N} \quad |\mu_{N_t(s,a)}(s,a) - \mu(s,a)| \geqslant b^H_{N_t(s,a)}(\delta)\Big) \leqslant \delta, \text{ with } b_n^H(\delta) = \sqrt{\frac{(1+\frac{1}{n})\log(2\sqrt{n+1}/\delta)}{2n}}.$$

Further, for any discrete distribution $p(\cdot|s,a)$ on $\mathcal{S}$ with support[2] of size $K \leqslant |\mathcal{S}|$

$$\mathbb{P}\Big(\forall t\in\mathbb{N}\, \|p_{N_t(s,a)}(\cdot|s,a) - p(\cdot|s,a)\| \geqslant b^W_{N_t(s,a)}(\delta)\Big) \leqslant \delta, \text{ with } b_n^W(\delta) = \sqrt{\frac{2(1+\frac{1}{n})\log\left(\sqrt{n+1}\frac{2^K-2}{\delta}\right)}{n}}.$$

Hence we replace $b_t^H(s,a,\frac{\delta}{2SA})$ with $b^H_{N_t(s,a)}(\frac{\delta}{2SA})$, and $b_t^W(s,a,\frac{\delta}{2SA})$ with $b^W_{N_t(s,a)}(\frac{\delta}{2SA})$ in the definition of $\mathscr{M}_t$. As the bounds for rewards and transitions are similar, from now on, we assume the mean function μ is known, to simplify the presentation.

Remark 5 *In contrast a peeling argument would lead to a seemingly better* $\log\log$ *scaling in place of the* $\log$ *square-root scaling, but with increased multiplicative constants compared to* $(1+1/n)$. *The bound using the Laplace method is better in practice (for $n < 2.10^9$ when $\delta = 0.01$).*

3 Class-UCRL : known class and profiles

We now introduce the first natural modification of UCRL2, that takes into account the similarity of state-action pairs. We assume that an oracle provides us with a perfect knowledge of the equivalence classes $\mathcal{C}$ plus profile maps $\sigma = (\sigma_{s,a})_{s,a}$ of all state-action pairs. In this ideal situation, our goal is to illustrate the potential benefit of taking the similarity into account. We explain the modification of the algorithm, provide a bound on the regret and illustrate its behavior on numerical experiments.

[2] In all this paper, we consider $K = \mathcal{S}$, that is, we have no prior information on the support.

Class-UCRL2 The most obvious modification is to aggregate observations from all state-action pairs in the same class in order to build more accurate estimates. Formally (i) for a class $c \subset \mathcal{S} \times \mathcal{A}$, with $N_t(c) = \sum_{s,a \in c} N_t(s,a)$ many observations, we define for each index $\in \{1, \dots, S\}$,

$$p^{\sigma}_{N_t(c)}(i|c) \quad = \quad \frac{\sum_{s,a \in c} N_t(s,a) p_{N_t(s,a)}(\sigma_{s,a}(i)|s,a)}{N_t(c)} \,.$$

Then, (ii) for a partition $\mathcal{C}$ of $\mathcal{S} \times \mathcal{A}$ in equivalence classes,

$$\mathscr{M}_t(\mathcal{C}) = \left\{ (\mathcal{S}, \mathcal{A}, \tilde{p}, \tilde{\nu}) \colon \tilde{p} \in \mathrm{Pw}(\mathcal{C}), \forall c \in \mathcal{C} \forall (s,a) \in c, \|p^{\sigma}_{N_t(c)}(\sigma^{-1}_{s,a}(\cdot)|c) - \tilde{p}(\cdot|s,a)\|_1 \leqslant b^{W}_{N_t(c)}\Big(\frac{\delta}{2C}\Big) \right\} .$$

where $\mathrm{Pw}(\mathcal{C})$ denotes the state-transition functions that are piecewise constant on $\mathcal{C}$ ($\tilde{p}(\cdot|s,a)$ has same value for all $(s,a) \in c$). Finally (iii), we redefine the stopping criterion as

$$t_{k+1} = \min \Big\{ t > t_k ; \exists c \in \mathcal{C} \colon n_{t_k:t}(c) \geqslant \max\{N_{t_k}(c), 1\} \Big\} .$$

The precise modified steps of the algorithm are presented in Algorithm 1 in appendix D.

Definition 5 (Class-UCRL2) **C-*UCRL2*$(\mathcal{C}, \sigma)$** *is defined as* ***UCRL2*** *using modifications (i,ii,iii).*

Remark 6 *It is crucial to remark that the algorithm is not using classes as "meta" states (that is, replacing the states with classes); Rather, the classes are only used to group observations from different sources and build more refined estimates for each state-action: The plausible MDPs are built using the same underlying state $\mathcal{S}$ and action space $\mathcal{A}$, unlike e.g. in Ortner (2013).*

Regret guarantee A modification of the analysis from Jaksch et al. (2010) yields:

Theorem 1 (Regret of C-UCRL2$(\mathcal{C}, \sigma)$) *With probability higher than $1 - 2\delta$, uniformly over all time horizon T,*

$$\begin{aligned} \mathrm{Regret}(\textbf{C-}\textit{UCRL2}(\mathcal{C}, \sigma), T) \quad &\leqslant \quad \Big(D\sqrt{4 \ln\big(\frac{2C\sqrt{T+1}(2^K - 2)}{\delta}\big)} + 2 \Big)\Big(\sqrt{2} + 1\Big)\sqrt{CT} \\ &\quad + D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)} + DC \log_2(\frac{8T}{C}) \\ &\leqslant \quad 17 D \sqrt{CTK \ln(C\sqrt{T}/\delta)} \,, \end{aligned}$$

where K bounds the support of the transition maps, and the proof is provided in appendix F.

The regret of this algorithm that knows the structure (but not the distributions) thus scales with the number of classes C, hence achieving a massive reduction when $C \ll SA$. This is the case in many grid-worlds thanks to our definition using orderings, see Appendix A. Also in such grid-worlds where transitions are local, $K < 5$ is a small constant. A comparison of the regret accumulated after $T = 2 \times 10^4$ steps on a 2-room MDP with 25 states (see Section 7), shows that **C** $-$ UCRL2 has regret in 149.0 ± 33.8 while UCRL2 has regret in 19487 ± 0.0 Now on a 4-room MDP with 49 states, we get a regret in 179.6 ± 35.7 versus 97131.8 ± 1003.3 for UCRL2 .

4 Known classes, unknown profile mappings

In this section, we consider a more realistic setting when the oracle provides the classes $\mathcal{C}$ to the learner but none of the profile mappings ($\sigma_{s,a}$ functions) are available.
In this more challenging situation, the algorithm **C**-UCRL2$(\mathcal{C}, \sigma)$ must be amended to use estimates of the profile mappings $\sigma_{s,a}$ for each s, a; we call the resulting algorithm **C**-UCRL2$(\mathcal{C})$.

Modified aggregation Let $\widehat{\sigma}_{s,a}$ be any profile mapping σ such that at time t, $p_{N_t(s,a)}(\sigma(1)|s,a) \geqslant p_{N_t(s,a)}(\sigma(2)|s,a) \geqslant \dots \geqslant p_{N_t(s,a)}(\sigma(S)|s,a)$. We build the modified empirical estimate as:

$$\widehat{p}^{\widehat{\sigma}}_{N_t(c)}(i|c) \quad = \quad \frac{\sum_{s,a \in c} N_t(s,a) p_{N_t(s,a)}(\widehat{\sigma}_{s,a}(i)|s,a)}{N_t(c)} \,.$$

Modified set of plausible MDPs We then modify the definition of $\mathscr{M}_t(\mathcal{C})$ to use for $\tilde{p} \in \mathrm{Pw}(\mathcal{C})$,

$$\|p^{\widehat{\sigma}}_{N_t(c)}(\widehat{\sigma}^{-1}_{s,a}(\cdot)|c) - \tilde{p}(\cdot|s,a)\|_1 \leqslant b^{W}_{N_t(c)}\Big(\frac{\delta}{2C}\Big) \,.$$

The previous construction is justified by the following non-expansive property of the ordering operator, as it ensures the Weissman concentration inequality also applies to the ordered empirical distribution. This also ensures that Theorem 1 also applies to **C**-UCRL2$(\mathcal{C})$ with same regret bound.

Lemma 2 (Non-expansive ordering) *Let $p \in \mathcal{P}(\mathcal{S})$ with profile map σ. Let p_n be its empirical version built from n samples, with profile map σ_n. Then (see appendix G),*

$$\|p_n(\sigma_n(\cdot)) - p(\sigma(\cdot))\|_1 \leqslant \|p_n(\sigma(\cdot)) - p(\sigma(\cdot))\|_1 \,.$$

5 Unknown classes: clustering

We finally address the most challenging situation when both the classes and profile mappings are unknown to the learner. To this end, we first introduce an optimistic clustering procedure that groups distributions based on statistical tests and is provably consistent.

Then, we introduce a natural modification of **C**-UCRL2($\mathcal{C}$) that uses a clustering algorithm to estimate the classes; We call this algorithm simply **C**-UCRL2, in contrast with **C**-UCRL2($\mathcal{C}$) which knows the classes, and **C**-UCRL2($\mathcal{C}, \sigma$) that also knows the profile mappings.

Since the non-trivial question is whether a benefit can be observed in practice, we illustrate this strategy on numerical experiments (Figure 2 and Section 7). They show that this strategy is promising as it does outperform UCRL2. But also that it suffers a much higher regret than its oracle counterpart, and is thus unable to leverage the unknown structure of the MDP. We realize this does not come from the clustering part that is sound, but rather from the episodes that stop too late. We thus introduce in Section 6 a key modification of the stopping rule, replacing the original heuristics from UCRL2 with hypothesis testing. This considerably improves its behavior on several illustrative examples, and also benefits the vanilla UCRL2 strategy.

Estimated clusters The clustering algorithm is inspired from Khaleghi et al. (2012) and does not require to know the number of clusters in advance. This approach provides sound estimations by using tight concentration bounds from Lemma 1. Let $\mathcal{J} = \{1, \dots, J\}$ be an indexation of $\mathcal{S} \times \mathcal{A}$, where $J = SA$. If index j corresponds to the pair (s, a), we write $N(j)$ for $N_t(s, a)$ and introduce $q(\cdot|j) = p^{\widehat{\sigma}}_{N_t(s,a)}(\widehat{\sigma}_{s,a}(\cdot)|s, a)$. We extend these notations to sets $c \subset \mathcal{J}$ with $N(c) = \sum_{j\in c} N(j)$ and $q(\cdot|c) = \sum_{j\in c} \frac{N(j)q(\cdot|j)}{N(c)}$. Now, starting from the trivial partition of $\mathcal{J}$ into singletons $\mathcal{C}_0 = \big\{\{1\}, \dots, \{J\}\big\}$, Algorithm 3 builds a coarser partition, by iteratively merging sets of the partitions. Two sets (clusters) are merged if and only if they are *statistically close*. For a given partition $\mathcal{C}$ of $\mathcal{J}$ and $c \in \mathcal{C}$, we define the *statistically closest* set $c_0 \in \mathcal{C}$ from c (when it exists) as:

$$\text{Near}(c, \mathcal{C}) = \operatorname{argmin} \left\{ d(c, c_0) : c_0 \in \mathcal{C} \setminus \{c\} \text{ s.t.} d(c, c_0) \leqslant 0 \text{ and } \forall j \in c, j_0 \in c_0 \; d(j, j_0) \leqslant 0 \right\},$$

where we introduced the penalized dissimilarities with the bound $\varepsilon_{\mathcal{C}}(n) = b^W_n(\frac{\delta}{2|\mathcal{C}|})$

$$\begin{aligned} d(c, c_0) &= \|q(\cdot|c) - q(\cdot|c_0)\|_1 - \varepsilon_{\mathcal{C}}(N(c)) - \varepsilon_{\mathcal{C}}(N(c_0)), \\ d(j, j_0) &= \|q(\cdot|j) - q(\cdot|j_0)\|_1 - \varepsilon_{\mathcal{J}}(N(j)) - \varepsilon_{\mathcal{J}}(N(j_0)) \,. \end{aligned}$$

The clustering algorithm then proceeds as follows: From $\mathcal{C}^0$, the sets $c \in \mathcal{C}^0$ are ordered in decreasing order of $N(c)$, so as to promote sets with tightest confidence intervals. Then, starting from c with largest $N(c)$, it finds $c' = \text{Near}(c, \mathcal{C}^0)$ and merge it with c, thus creating the new cluster $c \cup c'$ in the novel partition $\mathcal{C}^1$, and removing c and c' from $\mathcal{C}^0$. The algorithm continues this procedure with the next set in $\mathcal{C}^0$, until exhaustion, thus producing a novel partition $\mathcal{C}^1$ of $\mathcal{J}$. The algorithm iterates this refinement process until iteration i when $\mathcal{C}^{i+1} = \mathcal{C}^i$ (convergence). It finally outputs at time t_k the clustering $\mathcal{C}_{t_k} \overset{\text{def}}{=} \mathcal{C}^i$ built from all the observations for episode k. At each iteration, either two or more clusters are grouped or the algorithm stops. Thus, it takes at most $|\mathcal{J}| - 1 = SA - 1$ steps for the algorithm to converge. The clustering algorithm is presented in Algorithm 3 and further details about merging and convergence are provided in appendix H.

Remark 7 (Optimistic clustering) *Algorithm 3 thus produces a partition $\mathcal{C}_t$ of $\mathcal{J}$ that clusters the distributions whose grouping satisfies the confidence bounds. As a result, when $N(s, a)$ is small for all pairs, all distributions tend to be grouped; non-similar state-action pair are then separated only when more evidence is collected.*

The correctness of the clustering algorithm is ensured under the following required assumption:

Assumption 1 (Separation between classes) *There exists some $\Delta > 0$ such that*

$$\forall c \neq c' \in \mathcal{C} \quad \forall (s, a) \in c, (s', a') \in c', \quad \|p(\sigma_{s,a}(\cdot|s, a)) - p(\sigma_{s',a'}(\cdot|s', a'))\|_1 \geqslant \Delta \,.$$

Lemma 3 *Under assumption 1, provided that $\min_{s,a} N_t(s, a) > f^{-1}(\Delta)$, where $f : n \to 2b^W_n(\frac{\delta}{SA})$ the clustering algorithm outputs the correct partition $\mathcal{C}$ of state-action pairs with high probability.*

Having introduced a clustering mechanism to find classes, the rest of the procedure is similar to **C**-UCRL2($\mathcal{C}$) with a modified set of plausible MDPs (details in Algorithm 2 and Appendix D):

$$\mathscr{M}_t(\mathcal{C}_t) = \left\{ (\mathcal{S}, \mathcal{A}, \tilde{p}, \tilde{\nu}) : \tilde{p} \in \text{Pw}(\mathcal{C}_t), \forall c \in \mathcal{C}_t, (s, a) \in c, \|p^{\widehat{\sigma}}_{N_t(c)}(\widehat{\sigma}^{-1}_{s,a}(\cdot)|c) - \tilde{p}(\cdot|s, a)\|_1 \leqslant b^W_{N_t(c)}\Big(\frac{\delta}{2|\mathcal{C}_t|}\Big) \right\},$$

and a modified stopping condition coming from the fact we are using concentration bounds both for the classes and for the pairs when building an estimated clustering (and thus $\mathscr{M}_t(\mathcal{C}_t)$):

$$t_{k+1} = \min \Big\{ t > t_k ; \exists c \in \mathcal{C}_k : n_{t_k:t}(c) \geqslant \max\{N_{t_k}(c), 1\} \text{or } \exists s, a : n_{t_k:t}(s, a) \geqslant \max\{N_{t_k}(s, a), 1\} \Big\} \,.$$

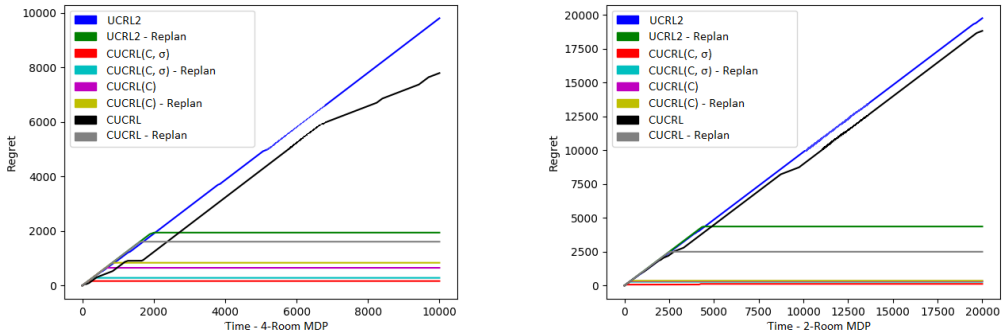

Figure 2: Regret as a function of time in an illustrative 4-room MDP with 25 states (left) and 2-room with 49 states (right). Note that UCRL2 gets sub-linear regret only after 1.2 and 2.5×10^5 time steps.

Remark 8 *Contrary to $\mathscr{M}_t(\mathcal{C})$, this set is* not guaranteed *to be optimistic due to the fact $\mathcal{C}_t$ may differ from $\mathcal{C}$. However, it provides a computationally cheap way to build a set of MDPs, compared to a more complex algorithm that would consider all plausible partitions $\tilde{\mathcal{C}}$.*

Figure 2 reveals that the clustering algorithm is improving over UCRL2, but is little able to handle unknown classes. One possible cause is that $\mathscr{M}_t(\mathcal{C}_t)$ may is not guarantee to contain $\mathcal{M}$ with high probability. However, extending $\mathscr{M}_t(\mathcal{C}_t)$ to handle all plausible clustering as well seems computationally heavy. Also, we observe numerically that the classes are correctly estimated. We thus call for another explanation.

6 A sound and effective stopping criterion using hypothesis testing

In this section, we revisit the criterion used to stop an episode in UCRL2.

A novel stopping criteria: inconsistency checking Since the plausible MDPs of UCRL2 algorithm are defined based on confidence intervals, the initial heuristic idea that guides the stopping time is to recompute a policy when these bounds have changed a lot. In other words, when number of observations is doubled for at least one state-action pair.

However, we suggest an alternative view, which is the second main contribution of this paper: since the optimistic policy in episode k corresponds to an optimistic MDP $\mathcal{M}_t^+ = (\mathcal{S}, \mathcal{A}, p_t^+, \nu_t^+)$ built at time t_k it is natural to check *whether this optimistic MDP is still consistent with the observations* at subsequent time steps $t > t_k$, and to stop when this is no longer the case. Thus, should we be able to detect the inconsistency of empirical estimation and the optimistic one, we would be able to detect this phenomenon sooner and update the policy instead of waiting for a long time. We actually observe that the episodes in **C**-UCRL2 last much longer even after novel observations show clear inconsistency of $\mathcal{M}_t^+$, which suggests this is the main reason for the observed regret.

A closer look at Extended Value Iteration A close look at Extended Value Iteration from Jaksch et al. (2010) reveals that it identifies a specific "optimistic" state s^+ with maximal optimistic value. This state is then used in order to build the optimistic transition model $p^+(\cdot|s,a)$, by putting maximal mass compatible with the empirical observations to transit to s^+. In grid-world MDPs, this often means that p^+ puts a large positive mass for transiting to a state s^+ from every (s,a), even when it is not reachable in one step from (s,a). This suggests detecting inconsistency of the optimistic model by looking at such specific transitions.

Modified stopping criterion Formally, we make use of concentration inequalities for a single entry $p(s'|s,a)$, and check if $p_{t_k}^+$ is compatible with the empirical transition at any time $t > t_k$, until the test fails. More precisely, for any pair s, a and s', it would make sense to use the test:

$$|p_{N_t(s,a)}(s'|s,a) - p_{t_k}^+(s'|s,a)| \leqslant (1+\varepsilon) b_{N_t(s,a)}^H \left(\frac{\delta}{S^2 A} \right),$$

where $\varepsilon > 0$ is a small margin. Indeed when $\varepsilon = 0$ and $p_{t_k}^+ = p$, the previous inequality is valid uniformly over all t, s, a, s' with probability higher than $1-\delta$. Considering $\varepsilon > 0$ allows to handle the case when $p_{t_k}^+$ differs a little from p. Further grouping the probabilities to transit to a target state s' on all pairs enables to get some refinement: For any $s' \in \mathcal{S}$,$g \subset \mathcal{S} \times \mathcal{A}$, and state-transition function q, we define the notation $q(s'|g) = \sum_{s,a \in g} q(s'|s,a)$. We then define $g_{t_k}(s_0) = \{s, a : s_0 \in \operatorname{argmax}_{s'} q(s'|s,a)\}$, for $q = p_{t_k}^+$. Since $p_{t_k}^+ \in \mathbf{Pw}(\mathcal{C}_{t_k})$, the set $\mathcal{G}_{t_k} = \{g_{t_k}(s_0) : s_0 \in \mathcal{S}\}$

contains at most $|\mathcal{C}_{t_k}|$ different groups. Thus, assuming that $|\mathcal{C}_{t_k}| \leqslant \overline{C}$ holds with high probability for some known constant $\overline{C}$, the following test is valid uniformly over all $t, s', g \in \mathcal{G}_{t_k}$

$$|p_{N_t(g)}(s'|g) - p^+_{t_k}(s'|g)| \leqslant (1+\varepsilon)|g| b^H_{N_t(g)}\Big(\frac{\delta}{S\overline{C}}\Big),$$

with high probability upon replacing $p^+_{t_k}$ with p and setting $\varepsilon = 0$.
This justifies to test at time $t+1$, after playing s_t, a_t, whether for the (random) state $s_{tg} \overset{\text{def}}{=} \operatorname{argmax}_{s,a} p^+_{t_k}(s_t, a_t)$ and the (random) group $g = g_{t_k}(s_{tg})$ it holds

$$|p_{N_t(g)}(s_{tg}|g) - p^+_{t_k}(s_{tg}|g)| \leqslant (1+\varepsilon)|g| b^H_{N_t(g)}\Big(\frac{\delta}{S\overline{C}}\Big), \tag{1}$$

leading to the following modified stopping time $\quad t_{k+1} = \min\{t > t_k : (1) \text{ fails}\}$.
We show in appendix I that a positive ε ensures that episodes for which $p^+_{t_k}$ is close to p, do not stop too early. Thus ensures a form of doubling stopping criterion (as for UCRL2) while the performed test enables an episode to terminate when an obvious mismatch between $p^+_{t_k}$ and p is detected. We use the value $\varepsilon = 1$ in the experiments, which is further discussed in appendix I.

Remark 9 *This novel stopping criterion is defined for an optimistic model p^+_t built at time $t = t_k$. Thus, it applies without further modification to $p^+_t \in \mathcal{M}_t$, $p^+_t \in \mathcal{M}_t(\mathcal{C})$ or $p^+_t \in \mathcal{M}_t(\mathcal{C}_t)$ and all algorithms including UCRL2 can be compared by replacing "doubling" stopping criterion.*

7 Numerical experiments: Empirical regret reduction in the agnostic case

We report in Figure 2 (see also appendix C) the outcome of running the algorithms on a few tasks (each averaged over 10 experiments). We consider: a) **Four-Room** MDP on a 5×5 grid world with four doors, and a total of $SA = 100$ pairs. and b) **Two-Room** MDP (wall in the middle) that is a 7×7 grid world, with one door. None of the algorithms is aware of the specific fact that the MDPs are grid-world, thus, instead of using support of $K = 4$, they all assume distributions with support $\mathcal{S}$. We compare the UCRL2 algorithm using Laplace bounds with the semi-oracles **C**-UCRL2$(\mathcal{C}, \sigma)$, **C**-UCRL2$(\mathcal{C})$ and the agnostic strategy **C**-UCRL2. Then, for each of these 4 algorithms, we consider their variants using the modified stopping criterion from section 6; they are denoted with suffix "replan".

Discussion We note that the agnostic **C**-UCRL2-replan is outperforming UCRL2 by a huge margin and is also outperforming UCRL2-replan. This clearly indicates that using the notion of similarity between state-action pairs can benefit learning even in the fully agnostic case, which answers our initial question in a non-trivial way.
In the two room environment (Figure 2), the algorithm has learned after about 2500 steps, given that there are about 50 states and 4 actions, this corresponds to about 12.5 observations per state-action pair on average, while UCRL2 is still suffering a linear regret event after 20000 steps (about 100 observations per state-action pair). UCRL2 indeed requires a much larger time horizon to converge on such experiments, of order 10^5 (see Appendix C).
The improvement is achieved while using the same tools as UCRL2, which opens interesting area of research, in view of the KL-UCRL (see Filippi et al. (2010)) and Thompson-Sampling alternatives.

Conclusion

While we believe it is more convincing to provide empirical evidence that one can leverage the structure of an MDP in the fully agnostic case to benefit regret minimization, our empirical findings obviously call for a theoretical regret analysis. We leave this intricate, but somehow secondary question for an extended version of this already dense article.
We introduced a state-action pair similarity, borrowed from optimal transport, in a reinforcement learning setup. This similarity is based on discrete optimal transport between distributions and produces large equivalence classes, thus effectively reducing the number of parameters to be learned. First, we have shown on numerical experiments and in theory that taking advantage of this structure can massively reduce the cumulative regret of a simple UCRL2 algorithm. In the challenging scenario when the partition is unknown, a non-trivial question is whether such an improvement can still be observed. Interestingly, our findings have led us to consider a second important idea: changing the stopping criterion of UCRL2 to take into account the possible mismatch between the optimistic environment and the observed one. This simple idea is illustrated on numerical experiments, where it proves to be massively beneficial, reducing the regret by several orders of magnitude. Whether this improvement can be seen on regret bounds (whose constants are typically loose in such setups) is an intriguing open question.
While the main goal of the paper is to illustrate the benefit of these two powerful ideas, we have provided several theoretical results regarding the soundness of the approach, thus paving the way towards a sharper understanding of regret guarantees in discrete Reinforcement Learning.

References

Abbasi-Yadkori, Yasin, Pál, Dávid, and Szepesvári, Csaba. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abel, David, Hershkowitz, David, and Littman, Michael. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, pp. 2915–2923, 2016.

Agrawal, Shipra and Jia, Randy. Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 1184–1194, 2017.

Anand, Ankit, Grover, Aditya, Singla, Parag, et al. Asap-uct: Abstraction of state-action pairs in uct. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Auer, Peter, Cesa-Bianchi, Nicolò, and Fischer, Paul. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Azar, Mohammad Gheshlaghi, Osband, Ian, and Munos, Rémi. Minimax regret bounds for reinforcement learning. In Precup, Doina and Teh, Yee Whye (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 263–272, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/azar17a.html`.

Bartlett, Peter L and Tewari, Ambuj. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 35–42, 2009.

Dann, Christoph, Lattimore, Tor, and Brunskill, Emma. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 5711–5721, 2017.

Dean, Thomas, Givan, Robert, and Leach, Sonia. Model reduction techniques for computing approximately optimal solutions for markov decision processes. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pp. 124–131. Morgan Kaufmann Publishers Inc., 1997.

Ferns, Norm, Panangaden, Prakash, and Precup, Doina. Metrics for finite markov decision processes. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 162–169. AUAI Press, 2004.

Ferns, Norm, Castro, Pablo Samuel, Precup, Doina, and Panangaden, Prakash. Methods for computing state similarity in markov decision processes. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 174–181. AUAI Press, 2006.

Filippi, Sarah, Cappé, Olivier, and Garivier, Aurélien. Optimism in reinforcement learning and Kullback-Leibler divergence. In *Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 115–122, 2010.

Hallak, Assaf, Di-Castro, Dotan, and Mannor, Shie. Model selection in markovian processes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 374–382. ACM, 2013.

Jaksch, Thomas, Ortner, Ronald, and Auer, Peter. Near-optimal regret bounds for reinforcement learning. *The Journal of Machine Learning Research*, 11:1563–1600, 2010.

Khaleghi, Azadeh, Ryabko, Daniil, Mary, Jérémie, and Preux, Philippe. Online clustering of processes. In *Artificial Intelligence and Statistics*, pp. 601–609, 2012.

Li, Lihong, Walsh, Thomas J, and Littman, Michael L. Towards a unified theory of state abstraction for mdps. In *ISAIM*, 2006.

Maillard, Odalric-Ambrym, Mann, Timothy A, and Mannor, Shie. How hard is my MDP? "the distribution-norm to the rescue". In *Advances in Neural Information Processing Systems 27 (NIPS)*, pp. 1835–1843, 2014.

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

Ortner, Ronald, Maillard, Odalric-Ambrym, and Ryabko, Daniil. Selecting near-optimal approximate state representations in reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 140–154. Springer, 2014.

Peña, Victor H, Lai, Tze Leung, and Shao, Qi-Man. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

Puterman, Martin L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Ravindran, Balaraman and Barto, Andrew G. Approximate homomorphisms: A framework for non-exact minimization in markov decision processes. 2004.

A Other examples of MDPs

We consider a grid-world MDP with four actions $a \in \{u, d, l, r\}$. 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 same probability 0.05 (it never goes down). When the resulting state is a wall, the distribution is modified: the probability mass is reported on the current state. Other actions have similar effect. Finally, the goal-state is put in the bottom-right corner of the MDP.

We now illustrate the scalability of the state-action pair similarity notion, of Definition 1 involving permutations. We show below four examples of grid-worlds defined according to the above scheme, with different number of state-action pairs.

Grid-world	Figure 3	Figure 4	Figure 5	Figure 6
SA	84	800	736	$\sim 10^4$
$\lvert\mathcal{C}\rvert$	6	6	7	7

Moreover, the number of state-action pairs in the introduced 4-room and 2-room MDPs changes as the grid state size grows, while the number of classes is fixed:

States	5*5	7*7	9*9	100*100
4Room-SA	100	196	324	$4 * 10^4$
4Room-$\lvert\mathcal{C}\rvert$	3	3	3	3
2Room-SA	100	196	324	$4 * 10^4$
2Room-$\lvert\mathcal{C}\rvert$	4	4	4	4

We note, in stark contrast that other notions 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, in case it satisfies

$$\forall s, s' \in \mathcal{S}_i, \forall a \in \mathcal{A}, \mu(s, a) = \mu(s', a) \text{ and } \forall j, \sum_{s'' \in \mathcal{S}_j} p(s''|s, a) = \sum_{s'' \in \mathcal{S}_j} p(s''|s', a)\,.$$

This readily prevents any two states s, s' such that $p(\cdot|s, a)$ and $p(\cdot|s', a)$ have disjoint support from being in the same set $\mathcal{S}_i$. Thus, since in grid-world MDP where transitions are local, the number of pairs with disjoint support is (about 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 instead of states only, thus slightly reducing the total number of classes.

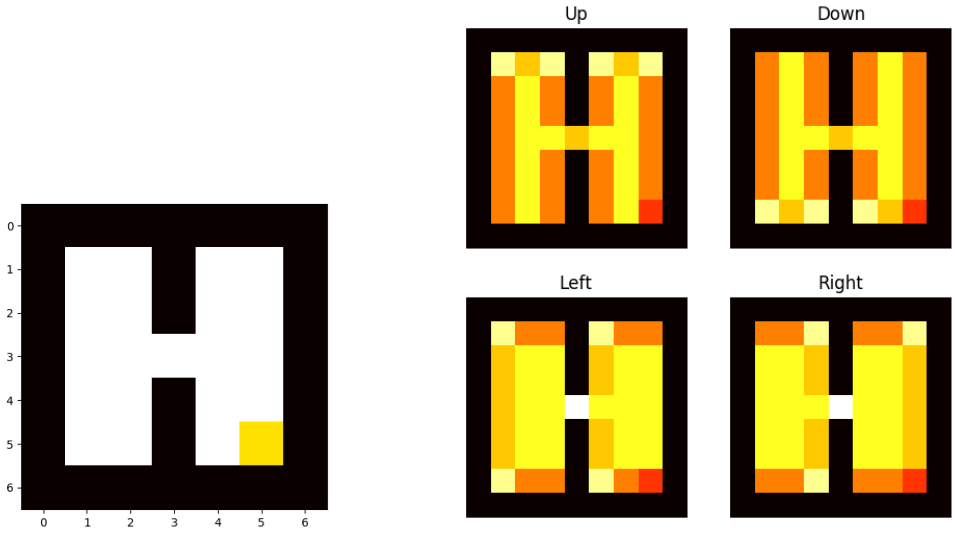

Figure 3: Left: Two-room grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

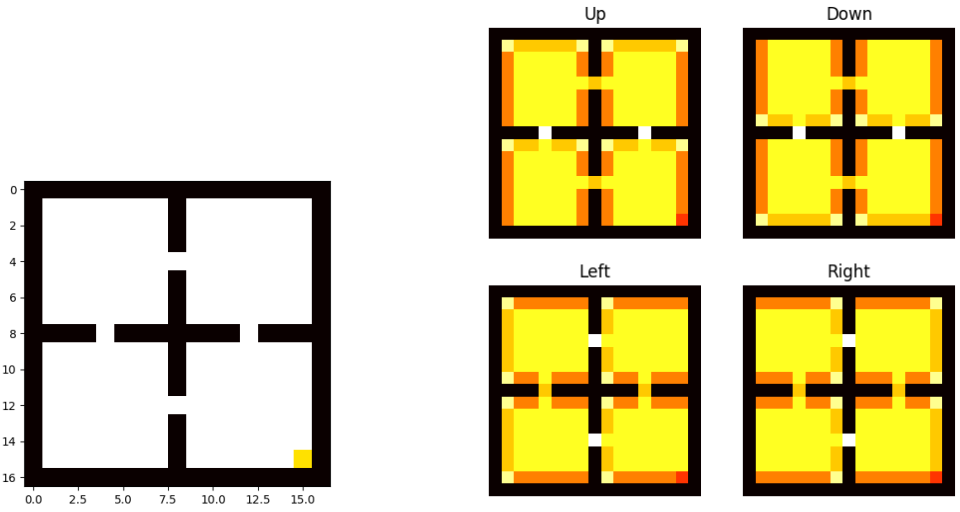

Figure 4: Left: Four-room grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

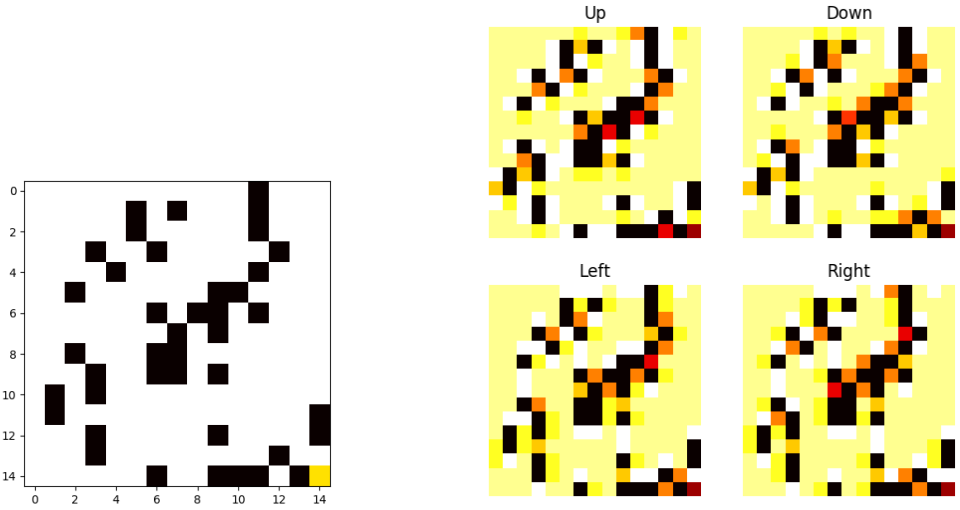

Figure 5: Left: A more complex grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

B Link with optimal transport

We now draw a connection between Definition 1 and optimal transport. Let us consider two discrete distributions p and q on $\mathcal{S}$ of size S. Let $\sigma : \{1, \dots, S\} \to \mathcal{S}$ be such that $p(\sigma(1)) \geqslant \dots \geqslant p(\sigma(S))$. Likewise, we consider a permutation τ such that $q(\tau(1)) \geqslant \dots \geqslant q(\tau(K))$, and define $y_i = \tau(i)$. Note that this can also be seen as considering the different *level sets* of the discrete distributions, a key notion when considering optimal transport.

An optimal transportation plan Γ between $p \circ \sigma$ and $q \circ \tau$ for a cost c minimizes over Γ

$$\sum_{i,j=1}^{S} c(i,j)\Gamma(i,j) \quad \text{where} \quad \sum_{j=1}^{S} \Gamma(i,j) = p \circ \sigma(i) \text{ and } \sum_{i=1}^{S} \Gamma(i,j) = q \circ \tau(j)\,.$$

When $c(i,j) = \mathbb{I}\{i \neq j\}$, the minimal value coincides with the total variation cost between $p \circ \sigma$ and $q \circ \tau$, that is $\|p \circ \sigma - q \circ \tau\|_1/2$. Introducing $s = \sigma(i)$ and $s' = \tau(j)$ explicitly, the optimal

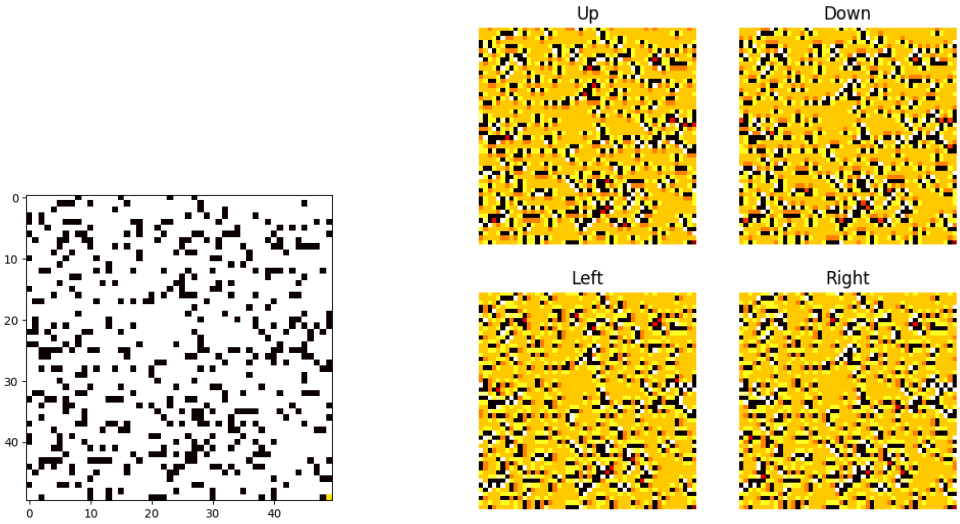

Figure 6: Left: A more complex grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

transportation problem now rewrites

$$\sum_{s,s' \in \mathcal{S}} c(\sigma^{-1}(s), \tau^{-1}(s'))\Gamma(\sigma^{-1}(s), \tau^{-1}(s'))$$

$$\text{where } \sum_{s' \in \mathcal{S}} \Gamma(\sigma^{-1}(s), \tau^{-1}(s')) = p(s) \text{ and } \sum_{s \in \mathcal{S}} \Gamma(\sigma^{-1}(s), \tau^{-1}(s')) = q(s')\,,$$

which corresponds to a cost $C(s, s') = \mathbb{I}\{\sigma^{-1}(s) \neq \tau^{-1}(s')\}$ defined on $\mathcal{S}$.

C Further experiments

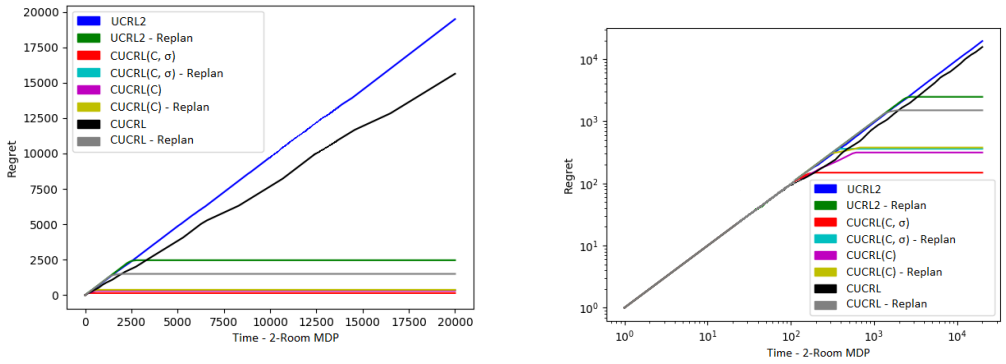

Figure 7: Regret as a function of time for an ilustrative for-room MDP, (standard, and loglog plot)

Let us also report below the number of observations required for `UCRL2` to converge in different MDPs:

2-Room, $5 \times 5 = 25$ states:	$8 * 10^4$	4-Room, $5 \times 5 = 25$ states:	$1.2 * 10^5$
2-Room, $7 \times 7 = 49$ states:	$2.5 * 10^5$	4-Room, $7 \times 7 = 49$ states:	$5 * 10^5$
2-Room, $9 \times 9 = 81$ states:	$9 * 10^5$	4-Room, $9 \times 9 = 81$ states:	10^6

D C-UCRL2 detailed pseudo-code

In this section, we provide further details about the implementation of $\mathbf{C}-$ UCRL2

Algorithm 1 Modified steps for Known Class & Mapping **C**-UCRL2$(\mathcal{C}, \sigma)$

In Initialize episode k at time t_k:
$\forall c \in \mathcal{C}$: set $\nu_k(c) := 0$ and compute
$N_{t_k}(c)$, $p^{\sigma}_{N_k(c)}(.|c)$, $\mu_{N_k(c)}(c)$.
In Compute Policy:
Apply Extended Value Iteration to $\mathscr{M}_{t_k}(\mathcal{C})$ to get compute (near) optimistic policy $\pi^+_{t_k}$.
In Execute policy $\pi^+_{t_k}$:
while $\nu_k(c_t) < \max\{1, N_{t_k}(c_t)\}$ where $c_t \in \mathcal{C}$ is the class containing (a_t, s_t) **do**
Choose $a_t = \pi^+_{t_k}(s_t)$,
Update c_t, and $\nu_k(c_t) := \nu_k(c_t) + 1$
Obtain reward r_t, observe next state s_{t+1}
t := t + 1
end while

Algorithm 2 Modified steps for Unknown Class & Mapping **C**-UCRL2

Find Similarity Classes and update:
$\widehat{p}_k \leftarrow \widehat{\sigma} \circ \widehat{p}_k$ {Reorder all probability distributions according to their permutation function}
$C_k \leftarrow$ **clustering** $\widehat{p}_k$
{An entity with all clustering information}
$r^{G_k}_N \leftarrow C_k.Centers_r$
$p^{G_k}_N \leftarrow C_k.Centers_p$
$N_k(c) := \#\{\tau < t_k : s_\tau = s, a_\tau = a, (s,a) \in c\}$
Compute Policy:
Let M_k be the set of all MDPs with states and actions as in M, and with transition probabilities $\bar{p}^+(.|c)$ close to $p^{G_k}_{N_k}(.|c)$ and rewards $\bar{r}^+(c) \in [0,1]$ close to $r^{G_k}_{N_k}(c)$, where $\forall (s,a) \exists c \in C_k.Centers : (s,a) \in G_C.Elements_c$ that is,
$\|\bar{p}^+(.|c) - p^{G_k}_{N_k}(.|c)\|_1 \leqslant b^W_{N_t(s,a)}(\delta)$
$|\tilde{r}(c) - \widehat{r}_{G_k}(c)| \leqslant b^W_n(\delta)$

E Time-uniform concentration inequalities

Theorem 2 (Hoeffding-Laplace concentration) *Let μ_n be the empirical mean built from n i.i.d. observations from a $[0,1]$-bounded distribution with mean μ. For any random stopping time τ with respect to the filtration of the past observations,*

$$\mathbb{P}\left(|\mu_\tau - \mu| \geqslant \sqrt{\frac{(1+\frac{1}{\tau})\log(2\sqrt{\tau+1}/\delta)}{2\tau}} \right) \leqslant \delta \,.$$

Corollary 1 (Weissman-Laplace concentration) *For any random stopping time τ with respect to the filtration of the past observations, and any discrete distribution p on $\mathcal{S}$ with support of size $K \leqslant |\mathcal{S}|$,*

$$\mathbb{P}\left(\|p_\tau - p\|_1 \geqslant \sqrt{\frac{2(1+\frac{1}{\tau})\log\left(\sqrt{\tau+1}\frac{2^K-2}{\delta}\right)}{\tau}} \right) \leqslant \delta \,.$$

Proof of Corollary 1:

For discrete measures, $\mathbb{P}\left(\|p_\tau - p\|_1 \geqslant \varepsilon \right) \leqslant \sum_{B \subset \mathcal{S}} \mathbb{P}\left(p_\tau(B) - p(B) \geqslant \frac{1}{2}\varepsilon \right)$. □

Algorithm 3 Confident Clustering

C $\leftarrow [p^{(1)}, \ldots, p^{(k)}]$
{Each sample is it's own cluster center}
N $\leftarrow [n_1, \ldots, n_k]$
$size \leftarrow [1, \ldots, 1]$ {k-element array of one}
$Changed \leftarrow$ **True**
while not Converged **and** Changed **do**
  $Changed \leftarrow$ **False**
  $Ordering \leftarrow argsort(N)$
  for all $i \in Ordering$ **do**
    if $n_i = 0$ **then**
      break
    end if
    k $\leftarrow$ **Near**(i, C) {Find the closest cluster to i}
    if $k = -1$ **then**
      Continue
    end if
    merge(k, i, C, N, size)
    $Changed \leftarrow$ **True**
  end for
end while

Algorithm 4 Combines two cluster centers

Function merge$(k, i, C, N, size)$:
  $C_k \leftarrow \frac{C_k * size_k + C_i * size_i}{size_k + size_i}$
  $N_k \leftarrow N_k + N_i$
  $size_k \leftarrow size_i + size_k$
  $size \leftarrow size - size_i$
  $N \leftarrow N - N_i$
  $C \leftarrow C - \{C_i\}$

Algorithm 5 Find Closest Cluster

Function Near$(index, C)$:
$k \leftarrow -1$
$dist_{min} \leftarrow \infty$
for all $i \in getIndex(C), i! = index$ **do**
  $dist = \|C_i - C_{index}\|_1 - \varepsilon_{C_i} - \varepsilon_{C_{index}}$
  if dist $<$ 0 **and** dist $< dist_{min}$ **and** isValid(i, index) **then**
    $dist_{min} \leftarrow dist$
    k = i
  end if
end for
return k

Algorithm 6 Checks both cluster points to have valid distant

Function isValid(i, j) :
$samples_i \leftarrow getSamples(i)$
{samples of i^{th} cluster}
$samples_j \leftarrow getSamples(j)$
for all $s_i \in samples_i$ **do**
  for all $s_j \in samples_j$ **do**
    if $\|s_i - s_j\|_1 - \varepsilon_i - \varepsilon_j > 0$ **then**
      Return False
    end if
  end for
end for
Return True

Corollary 2 *Let $N_t(s,a)$ be the number of observation of a state action pair (s,a) at time t. Then*

$$\mathbb{P}\Big(\exists t \in \mathbb{N}, \quad |\mu_{N_t(s,a)}(s,a) - \mu(s,a)| \geqslant \sqrt{\frac{(1+\frac{1}{N_t(s,a)})\log(2\sqrt{N_t(s,a)+1}/\delta)}{N_t(s,a)}}\Big) \quad \leqslant \quad \delta\,.$$

$$\mathbb{P}\Big(\exists t \in \mathbb{N}, \quad \|p_{N_t(s,a)}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geqslant \sqrt{\frac{2(1+\frac{1}{N_t(s,a)})\log\big(\sqrt{N_t(s,a)+1}\frac{2^K-2}{\delta}\big)}{N_t(s,a)}}\Big) \quad \leqslant \quad \delta\,.$$

Proof of Corollary 2:

$$\begin{aligned}
&\mathbb{P}\Big(\exists t : |\mu_{N_t(s,a)}(s,a) - \mu(s,a)| \leqslant \sqrt{\frac{(1+\frac{1}{N_t(s,a)})\log(2\sqrt{N_t(s,a)+1}/\delta)}{N_t(s,a)}}\Big) \\
&\quad\leqslant \quad \mathbb{P}\Big(\exists t, \exists n \leqslant t : |\mu_n(s,a) - \mu(s,a)| \leqslant \sqrt{\frac{(1+\frac{1}{n})\log(2\sqrt{n+1}/\delta)}{n}}\Big) \\
&\quad\leqslant \quad \mathbb{P}\Big(\exists n \in \mathbb{N} : |\mu_n(s,a) - \mu(s,a)| \leqslant \sqrt{\frac{(1+\frac{1}{n})\log(2\sqrt{n+1}/\delta)}{n}}\Big)\,,
\end{aligned}$$

Now, we recognize a uniform concentration inequality for the observations of the pair (s,a), which can be controlled by introducing the variable

$$\tau = \min\left\{n : |\mu_n(s,a) - \mu(s,a)| \leqslant \sqrt{\frac{(1+\frac{1}{n})\log(2\sqrt{n+1}/\delta)}{n}}\right\},$$

which is a stopping time with respect to the observations generated by the pair (s,a). □

F Effective regret analysis of $\mathcal{C}$-UCRL2$(\mathcal{C},\sigma)$

The analysis is based on the analysis of UCRL2 given by Jaksch et al. (2010). We are going to investigate "State-Aggregated UCRL2" by analyzing the regret of the algorithm. The effective regret naturally decomposes episode-wise

$$\begin{aligned}
\mathfrak{R}(\mathbb{A},T) &= \sum_{t=1}^{T} g_\star(s_1) - \sum_{t=1}^{T}\mu(s_t,a_t) \\
&= \sum_{k=1}^{m(T)} \sum_{s,a\in\mathcal{S}\times\mathcal{A}} \underbrace{\sum_{t=t_k+1}^{t_{k+1}} \mathbb{I}\{s_t=s, a_t=a\}}_{\nu_k(s,a)} \Big(g_\star - \mu(s,a)\Big) \\
&= \sum_{k=1}^{m(T)} \sum_{c\in\mathcal{C}} \underbrace{\sum_{t=t_k+1}^{t_{k+1}} \mathbb{I}\{(s_t,a_t)\in c\}}_{\nu_k(c)} \Big(g_\star - \mu(c)\Big) = \sum_{k=1}^{m(T)} \Delta_k\,.
\end{aligned}$$

where we used that $\mu(s,a)$ has constant value $\mu(c)$ for all $(s,a) \in c$ and introduced the effective regret in episode k

$$\Delta_k = \sum_{c\in\mathcal{C}} \nu_k(c)\Big(g_\star - \mu(c)\Big)\,.$$

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

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

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

where $m(T)$ is the number of episodes up to time T.

Control of the regret due to good episodes: $\mathcal{M} \in \mathscr{M}_{t_k}$

We closely follow Jaksch et al. (2010) and decompose the regret to control the transition and reward functions. At a high level, the only modifications that we do are 1) to use a time-uniform bound to control the martingale difference sequence that appears in the proof, using the following result:

Lemma 4 (Time-uniform Azuma Hoeffding) *For any martingale difference sequence $(X_t)_t$ bounded by D (that is, $|X_t| \leqslant D$ for all t) it comes, by application of time-uniform Laplace concentration inequality for bounded variables,*

$$\mathbb{P}\left(\exists T \in \mathbb{N}, \quad \sum_{t=1}^{T} X_t \geqslant D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)}\right) \leqslant \delta\,.$$

2) to control the number of episodes differently, due to the use of the partition $\mathcal{C}$

Lemma 5 (Number of episodes) *The number of episodes of* **C**-*UCRL2*$(\mathcal{C}, \sigma)$ *up to step $T \geqslant C$, denoted by m(T) is upper bounded by:*

$$m(T) \leqslant C \log_2(\frac{8T}{C})$$

Proof of lemma 5:

Having defined N_T and ν_k as the total number of state-action observations, up to step T and in episode k respectively, there is a state-action class(c) in each episode $k < m$ with $\nu_k(c) = N_{t_k}(c)$. Let $K(c)$ be the number of episodes with $\nu_k(c) = N_{t_k}(c)$ and $N_{t_k}(c) > 0$ for all c. It is worth mentioning that if $N_{t_k}(c) > 0$ and $\nu_k(c) = N_{t_k}(c)$, $N_{t_{k+1}}(c) = 2N_{t_k}(c)$, therefore:

$$N(c) = \sum_{k=1}^{m(T)} \nu_k(c) \geqslant 1 + \sum_{k:\nu_k(c)=N_{t_k}(c)} N_{t_k}(c) \geqslant 1 + \sum_{i=1}^{K(c)} 2^{i-1} = 2^{K(c)} \tag{2}$$

If $N(c) = 0$, $K(c) = 0$, therefore, $N(c) \geqslant 2^{K(c)} - 1$ for all pair classes. Thus,

$$T = \sum_{c\in\mathcal{C}} N(c) \geqslant \sum_{c\in\mathcal{C}} \left(2^{K(c)} - 1\right) \tag{3}$$

On the other hand, an episode has happened when either $N_{t_k}(c) = 0$ or $N_{t_k}(c) = \nu_k(c)$. Therefore, $m \leqslant 1 + C + \sum_{c\in\mathcal{C}} K(c)$ and consequently, $\sum_{c\in\mathcal{C}} K(c) \geqslant m - 1 - C$. As a result:

$$\sum_{c\in\mathcal{C}} 2^{K(c)} \geqslant C 2^{\sum_{c\in\mathcal{C}} \frac{K(c)}{C}} \geqslant C 2^{\frac{m-1}{C}-1} \tag{4}$$

Using the last two equations:

$$T \geqslant C\left(2^{\frac{m-1}{C}-1} - 1\right) \tag{5}$$

Therefore,

$$m \leqslant 1 + 2C + C\log_2(\frac{T}{C}) \leqslant 3C + C\log_2(\frac{T}{C}) \leqslant C\log_2(\frac{8T}{C}) \tag{6}$$

And the lemma is proved. □

Details Since $\mathcal{M} \in \mathscr{M}_{t_k}$ by assumption and by choosing $\pi^+_{t_k}$ and $\mathcal{M}^+_{t_k}$ by following the algorithm, we get that $g_k \overset{\text{def}}{=} g^{\mathcal{M}^+_{t_k}}_{\pi^+_{t_k}} \geqslant g_\star - \frac{1}{\sqrt{t_k}}$. Thus, it can be said that:

$$\Delta_k \leqslant \sum_{c\in\mathcal{C}} \nu_k(c)\big(g_\star - \mu(c)\big) \leqslant \sum_{c\in\mathcal{C}} \nu_k(c)\big(g_k - \mu(c)\big) + \sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{t_k}} \tag{7}$$

Besides, according to the proof of extended value iteration in UCRL2we also have, for the output value at iteration i

$$\max_s u_i(s) - \min_s u_i(s) \leqslant D \tag{8}$$

where D is the diameter of the MDP. Moreover, when the convergence criterion of extended-value iteration holds we have:

$$|u_{i+1}(s) - u_i(s) - g_k| \leqslant \frac{1}{\sqrt{t_k}} \forall s \in S \tag{9}$$

where g_k is the average reward of the policy $\pi^+_{t_k}$ chosen on the optimistic MDP $\mathcal{M}^+_{t_k}$. Using Bellman operator on the optimistic MDP:

$$u_{i+1}(s) = \mu^+_{t_k}(s, \pi^+_{t_k}(s)) + \sum_{s'} p^+_{t_k}(s'|s, \pi^+_{t_k}(s)).u_i(s')$$

which can be written as:

$$u_{i+1}(s) = \mu^+_{t_k}(c_{s,\pi^+_{t_k}(s)}) + \sum_{s'} p^+_{t_k}(s'|c_{s,\pi^+_{t_k}(s)}).u_i(s')$$

where $c_{s,a} \in C$ s.t. $\exists c \in \mathcal{C} : (s,a) \in c$. Employing the above equation alongside with 9 leads to:

$$|\Big(g_k - \mu^+_{t_k}(c_{s,\pi^+_{t_k}(s)})\Big) - \Big(\sum_{s'} p^+_{t_k}(s'|c_{s,\pi^+_{t_k}(s)}).u_i(s') - u_i(s)\Big)| \leqslant \frac{1}{\sqrt{t_k}}$$

By defining the column vector of $\mathbf{g}_k = (g_k(s))_s$, $\tilde{\mathbf{r}}_k := \big(\mu^+_{t_k}(c_{s,\pi^+_{t_k}(s)})\big)_s$ for $\pi^+_{t_k}$ policy, $\tilde{\mathbf{P}}_k := \big(p^+_{t_k}(s'|c_{s,\pi^+_{t_k}(s)})\big)_{s,s'}$ and $\upsilon_k := \big(\nu_k(c_{s,\pi^+_{t_k}(s)})\big)$ we can rewrite the above equation as:

$$|\Big(\mathbf{g}_k - \tilde{\mathbf{r}}_k\Big)_s - \Big((\tilde{\mathbf{P}}_k - \mathbf{I})u_i\Big)_s| \leqslant \frac{1}{\sqrt{t_k}}$$

Therefore, equation 7 can be rewritten as:

$$\begin{aligned}
\Delta_k &\leqslant \sum_{c\in\mathcal{C}} \nu_k(c)\big(g_k - \mu(c)\big) + \sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{t_k}} \\
&= \sum_{c\in\mathcal{C}} \nu_k(c)\big(g_k - \mu^+_{t_k}(c)\big) + \sum_{c\in\mathcal{C}} \nu_k(c)\big(\mu^+_{t_k}(c) - \mu(c)\big) + \sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{t_k}} \\
&\leqslant \upsilon_k(\tilde{\mathbf{P}}_k - \mathbf{I})u_i + \sum_{c\in\mathcal{C}} \nu_k(c)\big(\mu^+_{t_k}(c) - \mu(c)\big) + 2\sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{t_k}}
\end{aligned}$$

Since each row of $\tilde{\mathbf{P}}_k$ sums up to one, we can add a constant to u_i and still have the same equation. So, we are going to replace it with w_k:

$$w_k(s) := u_i(s) - \frac{\min_s u_i(s) + \max_s u_i(s)}{2}$$

At this point we are going to bound the reward part using reward confidence bound. Knowing that $\mathcal{M} \in \mathscr{M}_{t_k}$, then $\mu^+_{t_k}(c) - \mu(c) \leqslant |\mu^+_{t_k}(c) - \mu_{N_{t_k}(c)}(c)| + |\mu_{N_{t_k}(c)} - \mu(c)|$ can be bounded by 2. Therefore,

$$\begin{aligned}
\Delta_k &\leqslant \upsilon_{Ck}(\tilde{\mathbf{P}}_k - \mathbf{I})w_k + 2\sum_{c\in\mathcal{C}} \nu_k(c) b^H_{N_{t_k}(c)}(\frac{\delta}{2C}) + 2\sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{t_k}} \\
&\leqslant \upsilon_{Ck}(\tilde{\mathbf{P}}_k - \mathbf{I})w_k + 2\sum_{c\in\mathcal{C}} \nu_k(c)\sqrt{\frac{(1+\frac{1}{N_{t_k}(c)})\log(4C\sqrt{N_{t_k}(c)+1}/\delta)}{2N_{t_k}(c)}} + 2\sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{t_k}}
\end{aligned}$$

Since $\max\{1, N_{t_k}(c)\} \leqslant t_k \leqslant T$,

$$\Delta_k \leqslant \upsilon_{Ck}(\tilde{\mathbf{P}}_k - \mathbf{I})w_k + \Big(\sqrt{4\ln(\frac{2C\sqrt{T}}{\delta})} + 2\Big)\sum_{c\in\mathcal{C}} \frac{\nu_k(c)}{\sqrt{\max\{1, N_{t_k}(c)\}}} \tag{10}$$

Afterwards we have to bound the first part of the above equation, $\upsilon_k(\tilde{\mathbf{P}}_k - \mathbf{I})w_k$, which is going to be done by using the confidence bound on the transition probability distribution introduced in 2. Having defined $\mathbf{P}_k := \big(p(s'|c_{s,\pi^+_{t_k}(s)})\big)_{s,s'}$:

$$\upsilon_k(\tilde{\mathbf{P}}_k - \mathbf{I})w_k = \upsilon_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k + \mathbf{P}_k - \mathbf{I})w_k = \upsilon_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)w_k + \upsilon_k(\mathbf{P}_k - \mathbf{I})w_k \tag{11}$$

Assuming that $\mathcal{M} \in \mathscr{M}_{t_k}$ while knowing that $\|\mathbf{w}_k\| \leqslant \frac{D}{2}$ and taking advantage from holder inequality and 2:

$$\begin{aligned}
\upsilon_k(\tilde{\mathbf{P}}_k - \mathbf{P}_k)w_k &= \sum_s \sum_{s'} \nu_k(c_{s,\pi^+_{t_k}(s)})\Big(p^+_{t_k}(s'|c_{s,\pi^+_{t_k}(s)}) - p(s'|c_{s,\pi^+_{t_k}(s)})\Big)w_k(s') \\
&\leqslant \sum_s \nu_k(c_{s,\pi^+_{t_k}(s)})\|p^+_{t_k}(.|c_{s,\pi^+_{t_k}(s)}) - p(.|c_{s,\pi^+_{t_k}(s)})\|_1\|\mathbf{w}_k\|_\infty \\
&\leqslant \sum_s \nu_k(c_{s,\pi^+_{t_k}(s)})2b^W_{N_{t_k}(c_{s,\pi^+_{t_k}(s)})}\Big(\frac{\delta}{2C}\Big)\frac{D}{2} \\
&\leqslant \sum_s \nu_k(c_{s,\pi^+_{t_k}(s)})D\sqrt{\frac{2(1+\frac{1}{N_{t_k}(c_{s,\pi^+_{t_k}(s)})})\log\big(2C\sqrt{N_{t_k}(c_{s,\pi^+_{t_k}(s)})+1}\frac{2^S-2}{\delta}\big)}{\max(1, N_{t_k}(c_{s,\pi^+_{t_k}(s)}))}} \\
&\leqslant D\sqrt{4\ln\big(\frac{2C\sqrt{T+1}(2^{|S|}-2)}{\delta}\big)}\sum_c \frac{\nu_k(c)}{\max\{1, N_{t_k}(c)\}}
\end{aligned} \tag{12}$$

and to bound the second term in equation 11, we are going to use lemma 4 and lemma 5. Let us define $X_t := (p(.|c_t) - e_{s_{t+1}})w_k(t)\mathbb{1}_{\mathcal{M}\in\mathscr{M}_{t_k}}\,\forall t = 1, \ldots, T$. For any k with $\mathcal{M} \in \mathscr{M}_{t_k}$, we have that:

$$\upsilon_k(\mathbf{P}_k - \mathbf{I})w_k = \sum_{t=t_k}^{t_{k+1}-1}(p(.|c_t) - \mathbf{e}_{s_t})\mathbf{w}_k = \sum_{t=t_k}^{t_{k+1}-1}\Big(p(.|c_t) - \mathbf{e}_{s_{t+1}} + \mathbf{e}_{s_{t+1}} - \mathbf{e}_{s_t}\Big)\mathbf{w}_k$$

$$\sum_{t=t_k}^{t_{k+1}-1} X_t + w_k(s_{t+1}) - w_k(s_t) \leqslant \sum_{t=t_k}^{t_{k+1}-1} X_t + D$$

Knowing that $|w_k(t)|_\infty = \frac{D}{2}$ and using holder inequality, $|X_t| \leqslant |p(.|c_t) - e_{s_{t+1}}|_1\frac{D}{2} \leqslant \Big(|p(.|c_t)|_1 + |e_{s_{t+1}}|_1\Big)\frac{D}{2} = D$. So, X_t is bounded by D and also $\mathbb{E}\{X_t|s_1, a_1, \ldots, s_t, a_t\} = 0$, so that X_t is a sequence of martingale differences. Therefore by using lemma 4 we get:

$$\mathbb{P}(\exists T : \sum_{t=1}^{T} X_t \geqslant \underbrace{D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)}}_{\varepsilon_{p_2}}\Big) \leqslant \delta\,.$$

Using lemma 5, we know that $m \leqslant C\log_2\big(\frac{8T}{C}\big)$. By summing over all episodes we get:

$$\sum_{k=1}^{m(T)} \upsilon_k(\mathbf{P}_k - \mathbf{I})\mathbf{w}_k\mathbb{1}_{\mathcal{M}\in\mathscr{M}_{t_k}} \leqslant \sum_{t=1}^{T} X_t + m(T) \leqslant \varepsilon_{p_2} + m(T)D \leqslant \varepsilon_{p_2} + DC\log_2(\frac{8T}{C}) \tag{13}$$

with probability 1 - 2δ.

Final control Using a union bound over the event that $\mathcal{M}$ is plausible at any time, and over the control of the martingale difference sequence $(X_t)_t$ appearing in the control of good episodes, we deduce that the regret is controlled with probability higher than $1 - 2\delta$, uniformly over all $T \in \mathbb{N}$.

More precisely, using 10, 11, 12 and 13 and summing over all episodes:

$$\begin{aligned}\sum_{k=1}^{m}\Delta_k \mathbb{1}_{\mathcal{M}\in\mathscr{M}_{t_k}} &\leqslant \sum_{k=1}^{m} v_k(\tilde{\mathbf{P}}_k-\mathbf{P}_k)w_k \mathbb{1}_{\mathcal{M}\in\mathscr{M}_{t_k}} + \sum_{k=1}^{m} v_k(\mathbf{P}_k-\mathbf{I})w_k \mathbb{1}_{\mathcal{M}\in\mathscr{M}_{t_k}} \\ &\quad + \sum_{k=1}^{m}\Big(\sqrt{4\ln(\frac{2C\sqrt{T+1}}{\delta})}+2\Big)\sum_{c\in\mathcal{C}}\frac{\nu_k(c)}{\max\{1,N_{t_k}(c)\}} \\ \leqslant \Big(D\sqrt{4\ln\big(\frac{2C\sqrt{T+1}(2^{|S|}-2)}{\delta}\big)}+\sqrt{4\ln(\frac{2C\sqrt{T+1}}{\delta})}+2\Big)\sum_{k=1}^{m}\sum_{c}\frac{\nu_k(c)}{\max\{1,N_{t_k}(c)\}} \\ +D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)}+DC\log_2\left(\frac{8T}{C}\right) \qquad (14)\end{aligned}$$

To bound the above equation, we are going to introduce the below lemma:

Lemma 6 *For any sequence of numbers $z_1, z_2, \ldots, z_n$ with $0 \leqslant z_k \leqslant Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$*

$$\sum_{k=1}^{n}\frac{z_k}{\sqrt{Z_{k-1}}} \leqslant \Big(\sqrt{2}+1\Big)\sqrt{Z_n} \qquad (15)$$

which is proved in Jaksch et al. (2010). Knowing that $N_T(c) := \sum_k \nu_k(c)$, $\sum_c N_T(c) = T$ and $N_{t_k}(c) = \sum_{i<k}\nu_i(c)$ and using the above lemma we get:

$$\sum_{c\in\mathcal{C}}\sum_{k=1}^{m}\frac{\nu_k(c)}{\max\{1,N_{t_k}(c)\}} \leqslant \sum_{c\in\mathcal{C}}\Big(\sqrt{2}+1\Big)\sqrt{N_T(c)}$$

On the other hand, using Jensen's inequality we get:

$$\Big(\sqrt{2}+1\Big)\sum_{c\in\mathcal{C}}\sqrt{N_T(c)} \leqslant \Big(\sqrt{2}+1\Big)C\frac{\sum_{c\in\mathcal{C}}\sqrt{N_T(c)}}{C} \leqslant \Big(\sqrt{2}+1\Big)\sqrt{CT}$$

Therefore,

$$\sum_{c\in\mathcal{C}}\sum_{k=1}^{m}\frac{\nu_k(c)}{\max\{1,N_{t_k}(c)\}} \leqslant \Big(\sqrt{2}+1\Big)\sqrt{CT}$$

And by using the above result in equation 14, we get:

$$\begin{aligned}\sum_{k=1}^{m(T)}\Delta_k \mathbb{1}_{\mathcal{M}\in\mathscr{M}_{t_k}} &\leqslant D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)}+DC\log_2(\frac{8T}{C}) \\ +\Big(D\sqrt{4\ln\big(\frac{2C\sqrt{T+1}(2^{|S|}-2)}{\delta}\big)}+\sqrt{4\ln(\frac{2C\sqrt{T+1}}{\delta})}+2\Big)\Big(\sqrt{2}+1\Big)\sqrt{CT} \qquad (16)\end{aligned}$$

with probability of at least $1-\delta-\delta$.
Finally, the regret of **C**-UCRL2$(\mathcal{C},\sigma)$ is controlled on an event of probability higher than $1-\delta$, uniformly over all time T, by

$$\begin{aligned}\mathfrak{R}(\mathbf{C}\text{-UCRL2}(\mathcal{C},\sigma),T) &\leqslant \Big(D\sqrt{4\ln\big(\frac{2C\sqrt{T+1}(2^{|S|}-2)}{\delta}\big)}+\sqrt{4\ln(\frac{2C\sqrt{T+1}}{\delta})}+2\Big)\Big(\sqrt{2}+1\Big)\sqrt{CT} \\ &\quad +D\sqrt{2(T+1)\log(\sqrt{T+1}/\delta)}+DC\log_2(\frac{8T}{C})\,.\end{aligned}$$

We conclude by polishing the bound to highlight the main terms of the regret:

$$\begin{aligned}\mathfrak{R}(\mathbf{C}\text{-UCRL2}(\mathcal{C},\sigma),T) &\leqslant [6(\sqrt{2}+1)+\sqrt{2}+1]D\sqrt{CTS\ln(C\sqrt{T}/\delta)} \\ &\leqslant 17D\sqrt{CTS\ln(C\sqrt{T}/\delta)}\,.\end{aligned}$$

The bound in the main body of the paper removes the $\sqrt{4\ln(\frac{2C\sqrt{T+1}}{\delta})}$ when assuming the mean rewards are known.

G Ordered Weissman

Proof of Lemma 2:

Let us consider the case when a switch occurs between index 1 and 2, that is $\sigma_n(1) = \sigma(2)$ and $\sigma_n(2) = \sigma(1)$. In this situation, we thus have $p(\sigma(1)) > p(\sigma(2))$ but $p(\sigma_n(1)) \leqslant p(\sigma_n(2))$. Then, we study $\sum_{i=1,2} |p(\sigma(i)) - p_n(\sigma_n(i))|$.

First, we note that if $p_n(\sigma_n(1)) < p(\sigma(1))$ and $p_n(\sigma_n(2)) < p(\sigma(2))$, then

$$\begin{aligned} |p(\sigma(1)) - p_n(\sigma_n(1))| + |p(\sigma(2)) - p_n(\sigma_n(2)))| &= p(\sigma(1)) - p_n(\sigma_n(2)) + p(\sigma(2)) - p_n(\sigma_n(1)) \\ &= |p(\sigma(1)) - p_n(\sigma(1))| + |p(\sigma(2)) - p_n(\sigma(2))| \,. \end{aligned}$$

Likewise, the same equality occurs if $p_n(\sigma_n(1)) > p(\sigma(1))$ and $p_n(\sigma_n(2)) > p(\sigma(2))$.

Now, in the remaining intermediate cases (that is $p_n(\sigma(1)) < p(\sigma(2)) < p_n(\sigma(2)) < p(\sigma(1))$, $p(\sigma(2)) < p_n(\sigma(1)) < p_n(\sigma(2)) < p(\sigma(1))$ and $p(\sigma(2)) < p_n(\sigma(1)) < p(\sigma(1)) < p_n(\sigma(2))$), it is immediate to check that

$$|p(\sigma(1)) - p_n(\sigma_n(1))| + |p(\sigma(2)) - p_n(\sigma_n(2)))| \leqslant |p(\sigma(1)) - p_n(\sigma(1))| + |p(\sigma(2)) - p_n(\sigma(2))| \,.$$

Thus, proceeding iteratively for all switch that occurs, and decomposing the permutations σ, σ_n into elementary switches, this shows that almost surely

$$\|p_n(\sigma_n(\cdot)) - p(\sigma(\cdot))\|_1 \leqslant \|p_n(\sigma(\cdot)) - p(\sigma(\cdot))\|_1 \,.$$

□

H Clustering Guarantees

H.1 Problem Definition

Consider the problem of having k different sequences $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(k)}$ generated using k probability distributions $Q^{(1)}, Q^{(2)}, \ldots, Q^{(k)}$ where all of them share the same $|\chi|$ alphabets and some of them may even have the same probability distribution. Therefore, one can say that they may have k' different classes of probability distributions where $k' \leqslant k$. In other words: $\forall \mathbf{x} \sim Q^{(i)} \in \{Q^{(1)}, \ldots, Q^{(k)}\} \exists k' < k : \mathbf{x} \sim Q^{(j)} \in \{Q^{(1)}, \ldots, Q^{(k')}\}$.
For instance, $\{\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \ldots, \mathbf{x}_n^{(1)}\}$ are elements of $\mathbf{x}^{(1)}$ which are drawn i.i.d. according to $Q^{(1)}$.
$P_{\mathbf{x}}$ represents the empirical probability distribution of a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and obviously, n is the length of the sequence. Therefore, we define $N = \{n_1, \ldots, n_k\}$ to represent the number of sampling steps our k sequences have. More concretely, $P_{\mathbf{x}}$ is a vector with size of the alphabet $|\chi|$ that gives the number of occurrences of each element divided by sequence length, i.e. $P_{\mathbf{x}}(a) = \frac{N(a|\mathbf{x})}{n}$.
In our problem setting, we receive a sample from one of the k different groups at each step. However, neither the probability distributions($Q^{(i)}$s) nor the number of different classes(k') is known.
We are going to propose a clustering approach so as to find this k' classes of probability distribution using the samples obtained from k initial distributions.

H.2 Confident Clustering (proof of Lemma 3)

Given $\{P_{\mathbf{x}^{(1)}}, \ldots, P_{\mathbf{x}^{(k)}}\}$ which would be called "samples" from now on and represented by $p^{(1)}, \ldots, p^{(k)}$ for simplicity. In this section, we are going to propose a clustering approach to group samples and find the k' distinct classes of probability distribution.
More concretely, we can say that for $k' \leqslant k$, $\forall Q^{(i)} \in \{Q^{(1)}, \ldots, Q^{(k)}\} \exists Q^{(j)} \in \{Q^{(1)}, \ldots, Q^{(k')}\}$ s.t. $Q^{(i)} \sim Q^{(j)}$. It is worth mentioning that we do not know the number of classes and probability distributions beforehand. We are provided with merely k different samples and we require to find an acceptable grouping in the last resort.

Initially, each sample is considered to be the center of a cluster and we have k clusters at the beginning. We sort the samples according to their corresponding value in N so as to guarantee that samples with higher confident are chosen sooner. Gaining advantage from the provided bounds, we search for the closest cluster assuring not only that the cluster centers are close enough, but also that all the cluster pairs satisfy 21 inequality. Thus, the clusters would be grouped when we have enough evidence to merge and would be left unchanged otherwise. The procedure is continued till the algorithm converges meaning that no cluster center alters. More exact procedure of the algorithm is demonstrated in 3.

Since the used samples for clustering are empirical probability distributions and there exists a deviation between empirical probability distributions and their corresponding true distribution, the measures are not exact and in order to bound error we are going to use Weissman concentration inequality measures to bound error.
To find whether two samples should be grouped or not, we have to find the distance of their probability distributions. So,

$$\begin{aligned} \|\tilde{p}^{(i)} - \tilde{p}^{(j)}\|_1 &\leqslant \|p^{(i)} - p^{(j)}\|_1 + \|\tilde{p}^{(i)} - p^{(i)}\|_1 + \|\tilde{p}^{(j)} - p^{(j)}\|_1 \\ &\leqslant \|p^{(i)} - p^{(j)}\|_1 + \varepsilon_i + \varepsilon_j \end{aligned}$$

In case that two samples have the same probability distribution, the difference of their true probability distribution is zero:

$$\|\tilde{p}^{(i)} - \tilde{p}^{(j)}\|_1 - \varepsilon_i - \varepsilon_j \leqslant 0 \tag{17}$$

Using Laplace inequality and knowing that it should hold for all state-action pairs of a cluster i, we can write:
$\forall (s,a) \in C_i$:

$$\mathbb{P}(||\tilde{p}(.|s,a) - p(.|s,a)||_1 \geqslant \varepsilon) \leqslant \frac{\delta}{2k} \tag{18}$$

Moreover, we also have to test the cluster centers to satisfy time uniform laplace inequality. Knowing that the number of clusters either reduces at least by one at each step or the clustering algorithm is converged, therefore the number of clusters for $iteration^{th}$ step would be at most "k - iteration + 1". So, $\forall C_i \in C$:

$$\begin{aligned} &\mathbb{P}(||\tilde{p}(.|C_i) - p(.|C_i)||_1 \geqslant \varepsilon) \\ &\leqslant \frac{\delta}{2(k - iteration + 1)} \end{aligned} \tag{19}$$

Therefore, using the above equation alongside with union bounds:

$$\begin{aligned} &\mathbb{P}(\forall (s,a) \in C_i : ||\tilde{p}(.|s,a) - p(.|s,a)||_1 \geqslant \varepsilon) \\ &+\mathbb{P}(\forall C_i \in C : ||\tilde{p}(.|C_i) - p(.|C_i)||_1 \geqslant \varepsilon) \\ &= \sum_{(s,a)\in C_i} \mathbb{P}(||\tilde{p}(.|s,a) - p(.|s,a)||_1 \geqslant \varepsilon) \\ &\quad + \sum_{C_i \in C} \mathbb{P}(||\tilde{p}(.|C_i) - p(.|C_i)||_1 \geqslant \varepsilon) \\ &\leqslant \sum_{(s,a)\in C_i} \frac{\delta}{2k} + \sum_{C_i \in C} \frac{\delta}{2(k - iteration + 1)} \\ &\leqslant \sum_{(s,a)\in C_i} \frac{\delta}{2k} + \sum_{C_i \in C} \frac{\delta}{2(k - iteration + 1)} \\ &\quad = \frac{|C_i|}{2k}\delta + \frac{|C|}{2(k - iteration + 1)}\delta \\ &\quad \leqslant \frac{k}{2k}\delta + \frac{k - iteration + 1}{2(k - iteration + 1)}\delta = \delta \end{aligned} \tag{20}$$

The last step of the above equation is due to the fact that C_i and number of elements in C_i are both a random variables and we do not know their exact value. So, we can replace them by the worst case

scenario. However, this bound is still not accurate and can be improved knowing the fact that the worst case for both parts can not happen simultaneously. For instance, when a cluster contains k elements, the number of clusters will not be k - iteration and would be 1.
Consequently, we can say that i and j should be grouped with high probability if inequality 17 is satisfied having used 18 to compute ε_i and ε_j:

$$\varepsilon_i = \sqrt{\frac{2}{n_i} \ln\Big(\frac{2^{|\chi|}-2}{\frac{\delta}{2k}}\Big)} \tag{21}$$

where T is the maximum possible number of samples for a sequence.
Following the same procedure, we can compare the distance of two clusters, C_i and C_j, using:

$$\varepsilon_{C_i} = \sqrt{\frac{2}{N_{C_i}} \ln\Big(\frac{2^{|\chi|}-2}{\frac{\delta}{2(k-iteration+1)}}\Big)} \tag{22}$$

where $N_{C_i} = \sum_{j \in C_i} n_j$ is the number of samples included in cluster i and iteration is the number of current iteration of clustering algorithm.

Moreover, following the same procedure for the reward part we would reach:

$$||\tilde{r}^{(i)} - \tilde{r}^{(j)}||_1 - \varepsilon_i^r - \varepsilon_j^r \leqslant 0 \tag{23}$$

And taking advantage from time uniform inequality for all state-action pairs we can find:

$$\mathbb{P}\big\{\forall (s,a) \in C_i : \tilde{r}(s,a) - r(s,a)||_1 \geqslant \varepsilon_i^r\big\} \leqslant \frac{\delta_r}{2k} \tag{24}$$

Considering the same condition as for probability distributions:

$$\mathbb{P}(\forall C_i \in C : ||\tilde{r}(C_i) - r(C_i)||_1 \geqslant \varepsilon_i^r) \leqslant \frac{\delta}{2(k - iteration + 1)} \tag{25}$$

Consequently, we can find

$$\varepsilon_i^r = \sqrt{\frac{1}{2n_i} \ln\Big(\frac{2k}{\delta}\Big)} \tag{26}$$

and

$$\varepsilon_{C_i}^r = \sqrt{\frac{1}{2N_{C_i}} \ln\Big(\frac{2(k - iteration + 1)}{\delta}\Big)} \tag{27}$$

and the idea is complete.
It is worth mentioning that this approach is only suitable when different class distributions are not overlapping which is the case of our reinforcement learning problem since in case of overlapping, they are having the same behavior and should be grouped consequently.

I A modified stopping criterion

The goal of the statistical test is to detect early enough a possible mismatch between the optimistic MDP $\mathcal{M}_{t_k}^+$ identified at time t_k and the correct MDP $\mathcal{M}$.

For a state $s_0 \in \mathcal{S}$ and a group of state-action pairs $g \subset \mathcal{S} \times \mathcal{A}$ and any state-transition distribution q, we denote by $q(s_0|g) = \sum_{s,a \in g} q(s_0|s,a)$ the probability of reaching s_0 from any of the pair $s, a \in g$. Likewise, we denote $N(g) = \sum_{s,a} N(s,a)$ for a counter N.

Proof :

Let $g_{t_k}(s_0) = \{s, a : s_0 \in \text{argmax}_{s'}\, q(s'|s,a)\}$, for $q = p_{t_k}^+$. Then, $\mathcal{G}_{t_k} = \{g_{t_k}(s_0) : s_0 \in \mathcal{S}\}$ contains at most $|\mathcal{C}_{t_k}|$ different groups.

At each time $t+1$, after playing s_t, a_t, the test checks whether for the (random) state $s_{tg} \stackrel{\text{def}}{=} \operatorname{argmax}_{s,a} p^+_{t_k}(s_t, a_t)$ and the (random) group $g = g_{t_k}(s_{tg})$ it holds

$$|p_{N_t(g)}(s_{tg}|g) - p^+_{t_k}(s_{tg}|g)| \leqslant (1+\varepsilon)|g| b^H_{N_t(g)}\Big(\frac{\delta}{S\overline{C}}\Big)\,, \tag{28}$$

We say episode k is good if there exists some (s_0, g) such that $|p^+_{t_k}(s_0|g) - p(s_0|g)| \leqslant |g| G_g$, and bad otherwise, for G_g to be defined later.

Closeness of $\widehat{p}_t$ and p. We first show that for all $t > t_k$ for the (random) state $\tilde{s}_t \stackrel{\text{def}}{=} \operatorname{argmax}_{s,a} p^+_{t_k}(s'|s_t, a_t)$ and the (random) group $\tilde{g}_t = g_{t_k}(\tilde{s}_t) \in \mathcal{G}_{t_k}$,

$$\mathbb{P}\bigg(\exists k, t_k : \exists t > t_k |p(\tilde{s}_t|\tilde{g}_t) - p_{N_t(\tilde{g}_t)}(\tilde{s}_t|\tilde{g}_t)| \geqslant |\tilde{g}_t| b^H_{N_t(\tilde{g}_t)}\Big(\frac{\delta}{S\overline{C}}\Big)\bigg) \leqslant \delta\,.$$

Indeed, considering all the possible states $\tilde{s}_t \in \mathcal{S}$ and groups $\tilde{g}_t \in \mathcal{G}_{t_k}$, we get

$$\begin{aligned}
&\mathbb{P}\bigg(\exists k, t_k : \exists t > t_k |p(\tilde{s}_t|\tilde{g}_t) - p_{N_t(\tilde{g}_t)}(\tilde{s}_t|\tilde{g}_t)| \geqslant |\tilde{g}_t| b^H_{N_t(\tilde{g}_t)}\Big(\frac{\delta}{S\overline{C}}\Big)\bigg) \\
&\leqslant \quad \mathbb{P}\bigg(\forall k, t_k : \exists s' \in \mathcal{S}, g \in \mathcal{G}_{t_k} : |p(s'|g) - p_{N_t(g)}(s'|g)| \geqslant |g| b^H_{N_t(g)}\Big(\frac{\delta}{S\overline{C}}\Big)\bigg)\,.
\end{aligned}$$

Since there are at most $\overline{C}$ many groups and S states, we simply use a union-bound argument over them. We then conclude by application of a time-uniform concentration bound for bounded distribution

$$\mathbb{P}\bigg(\forall k, t_k : |p(s'|g) - p_{N_t(g)}(s'|g)| \geqslant |g| b^H_{N_t(g)}\Big(\frac{\delta}{S\overline{C}}\Big)\bigg) \leqslant \frac{\delta}{S\overline{C}}\,.$$

Bad episodes The test (1) is satisfied at the last time before the end of the episode ($t = t_{k+1} - 1$). Thus, by closeness of $\widehat{p}_t$ and p at that time and definition of the test being passed, we deduce that at the last time before the end of the episode,

$$|p^+_{t_k}(\tilde{s}_t|\tilde{g}_t) - p(\tilde{s}_t|\tilde{g}_t)| \leqslant (2+\varepsilon)|\tilde{g}_t| b^H_{N_t(\tilde{g}_t)}\Big(\frac{\delta}{S\overline{C}}\Big)\,.$$

In particular, on an event of high probability, episode k is bad only if

$$(2+\varepsilon) b^H_{N_t(\tilde{g}_t)}\Big(\frac{\delta}{S\overline{C}}\Big) \geqslant G_{\tilde{g}_t}\,,$$

that is if $N_t(\tilde{g}_t)$ is small at the last time before the episode stops; typically $N_t(\tilde{g}_t) = \tilde{O}\Big(\frac{(2+\varepsilon)^2}{G_{\tilde{g}_t}}\Big)$. This ensures that we stop the episode early enough and recompute a better model in that case.

Good episodes Let us consider now a good episode such that $|p^+_{t_k}(s_0|g) - p(s_0|g)| \leqslant |g| G_g$ for all s_0, g. We want to show that such an episode is not stopped too early. Indeed in that case, we have on an event of high probability,

$$|p^+_{t_k}(\tilde{s}_t|\tilde{g}_t) - p_{N_t(\tilde{g}_t)}(\tilde{s}_t|\tilde{g}_t)| \leqslant |\tilde{g}_t| G_{\tilde{g}_t} + |\tilde{g}_t| b^H_{N_t(\tilde{g}_t)}\Big(\frac{\delta}{S\overline{C}}\Big)\,.$$

Thus, the episode does not stop unless

$$G_{\tilde{g}_t} > \varepsilon b^H_{N_t(\tilde{g}_t)}\Big(\frac{\delta}{S\overline{C}}\Big)\,.$$

Tuning Let $G_g = \beta b^H_{N_{t_k}(g)}\Big(\frac{\delta}{S\overline{C}}\Big)$. This ensures that a good episode stops only when

$$\beta b^H_{N_{t_k}(g)}\Big(\frac{\delta}{S\overline{C}}\Big) > \varepsilon b^H_{N_t(\tilde{g}_t)}\Big(\frac{\delta}{S\overline{C}}\Big)$$

thus typically when $N_t(g) \gtrsim \frac{\varepsilon^2}{\beta^2} N_{t_k}(g)$ for some g.

In particular, the specific value $\varepsilon = \sqrt{2}$ implements a criterion similar to the doubling stopping criterion of UCRL2 for good episodes.

On the other, an episode k is bad only if

$$(2+\varepsilon) b^H_{N_t(\tilde{g}_t)}\Big(\frac{\delta}{S\overline{C}}\Big) \geqslant \beta b^H_{N_{t_k}(\tilde{g}_t)}\Big(\frac{\delta}{S\overline{C}}\Big)$$

thus typically if $\frac{(2+\varepsilon)^2}{\beta^2} N_{t_k}(g) \gtrsim N_t(g)$ for all g.

Thus, we deduce that a "good" episode stops when $N_t(g) \gtrsim \frac{\varepsilon^2}{\beta^2} N_{t_k}(g)$ for some g, while a "bad" episode stops when $N_t(g) \gtrsim \frac{(2+\varepsilon)^2}{\beta^2} N_{t_k}(g)$ for some g.

For L1 norm, we can further show that for $g \in \mathcal{G}_{t_k}$ it holds

$$|p^+_{t_k}(s_0|g) - p(s_0|g)| \leqslant \|p^+_{t_k}(\cdot|g) - p(\cdot|g)\|_1 \leqslant 2|g| b^W_{N_{t_k}(g)}\Big(\frac{\delta}{S\overline{C}}\Big)$$

This leads to the natural choice to rule out all bad episodes with high probability:

$$G = \frac{2 b^W_{N_{t_k}(g)}\Big(\frac{\delta}{S\overline{C}}\Big)}{b^H_{N_{t_k}(g)}\Big(\frac{\delta}{S\overline{C}}\Big)} b^H_{N_{t_k}(g)}\Big(\frac{\delta}{S\overline{C}}\Big).$$

This in turns suggests to choose

$$\varepsilon = \frac{2\sqrt{2} b^W_{N_{t_k}(g)}\Big(\frac{\delta}{S\overline{C}}\Big)}{b^H_{N_{t_k}(g)}\Big(\frac{\delta}{S\overline{C}}\Big)}.$$

□

Number of episodes To conclude, it thus remains to compute a bound on the number of episodes and ensures that is stays logarithmic in the time horizon. This is done similarly to the count of episodes in UCRL2, since in each episode at least one group g gets a total number of observations multiplied by a constant factor.