Near-optimal Regret Bounds for Reinforcement Learning in Factored MDPs

Ian Osband, Benjamin Van Roy

Stanford University

May 11, 2014

Abstract

Any learning algorithm over Markov decision processes (MDPs) will have worst-case regret $\Omega(\sqrt{SAT})$ where $T$ is the elapsed time and $S$ and $A$ are the cardinalities of the state and action spaces. In many settings of interest $S$ and $A$ may be so huge that it is impossible to guarantee good performance for an arbitrary MDP on any practical timeframe $T$. We show that, if we know the true system can be represented as a factored MDP, we can obtain regret bounds which scale polynomially in the number of parameters of the MDP, which may be exponentially smaller than $S$ or $A$. Assuming an algorithm for approximate planning and knowledge of the graphical structure of the underlying MDP, we demonstrate that posterior sampling reinforcement learning (PSRL) and an algorithm based upon optimism in the face of uncertainty (UCRL-Factored) both satisfy near-optimal regret bounds.

1 Introduction

The classic reinforcement learning problem considers an agent who must make sequential decisions within its environment while trying to maximize total reward accumulated over time [1, 2]. The environment is modeled as a Markov decision process (MDP) but the agent is uncertain of the true dynamics of the MDP. The agent must plan actions to maximize rewards based upon its imperfect knowledge, but also learns about its environment through experience. Efficient reinforcement learning manages this tradeoff between exploration and exploitation in such a way that the deviation from the optimal policy given perfect information is controlled.

Factored MDPs [3] allow us to represent large structured MDPs compactly. A state is described by a selection of state variables, whose transitions can be represented by a dynamic Bayesian network (DBN) [4]. This is particularly beneficial when the transition of a state variable depends only on a small subset of other variables. For example, consider a large production line with $m$ machines in sequence, each with $K$ possible states. We write $s = (s_1, \ldots, s_m)$ with each $s_i \in \{1, \ldots, K\}$. It may be that, over a single time-step, machine $i$ can only be influenced by the states of $i-1, i$ and $i+1$. If so, any single $s_i$ can still influence the entire system eventually, but the dimensionality of the learning problem is reduced exponentially from $O(K^m)$ to $O(mK^3)$.

There has been some success in establishing efficient reinforcement learning in factored MDPs (FMDPs). Kearns and Koller extend the $E^3$ algorithm [5, 6] to exploit the DBN structure and obtain probably approximately correct (PAC) bounds with polynomial sample complexity. There are similar results available for the $R_{max}$ algorithm [7, 8] and even the greedy policy given an optimistic initialization [9]. These algorithms require the graph structure of the FMDP as a fixed prior. Some algorithms do seek to learn this structure from experience [10], but we will assume this structure is known.

Another form of efficiency guarantees for reinforcement learning are given by regret bounds. These bound the difference in accumulated rewards of a learning algorithm and the optimal policy over $T$ steps [11]. Regret bounds naturally give rise to PAC bounds as a corollary, but also give a guarantee on the algorithm’s performance during the learning phase. Jaksch et al. [12] present UCRL2, which attains near-optimal regret of $O(S \sqrt{AT})$ with high probability. Recently Osband et al. [13] analyze PSRL, which also provides bounds on the expected regret of $O(S \sqrt{AT})$. Unlike the algorithms mentioned so far, PSRL does not use “optimism in the face of uncertainty” (OFU) to guide exploration, but instead the variance in posterior sampling. There has been no algorithm with efficient regret bounds in FMDPs so far.
We present two algorithms, PSRL and UCRL-Factored, with efficient regret bounds for FMDPs. These algorithms are described fully in Section 6. UCRL-Factored is a minor modification to UCRL2 that allows us to exploit the DBN structure while PSRL is unchanged. UCRL-Factored is guided by the OFU principle whereas PSRL is guided by posterior (also known as Thompson) sampling [14]. Posterior sampling has already been shown to perform extremely well in bandits [15, 16, 17] and non-factored MDPs [13]. We believe that posterior sampling will be simpler to implement, computationally cheaper and statistically more efficient than the optimistic alternative in FMDPs as well.

Both algorithms make use of approximate FMDP planner in internal steps. However, even where an FMDP is able to represented concisely, solving for the optimal policy may still be intractable in the most general case [18]. Our focus in this paper is upon the statistical aspect of the learning problem and like earlier discussions [5] we do not specify which computational methods are used. Our results serve as a reduction of the reinforcement learning problem to finding an approximate solution for a given FMDP. In many cases of interest, effective approximate planning methods for FMDPs do exist. Investigating and extending these methods are an ongoing subject of research [19, 20, 21, 22].

We believe that dimensionality reduction in large MDPs is essential for practical reinforcement learning. Factored MDPs are an approach with successful applications in many fields [3] but they are not the only one. There are regret bounds for continuous state spaces where the underlying reward and transition functions are known to belong to some function class, such as Höder continuous [24] or linear quadratic control [24]. These results are interesting, but each have some undesirable properties, the former has regret bounds which approach $O(T)$ for high dimensions and the latter retains an exponential dependence on the dimension. Perhaps the most popular approach in the literature is to assume the value function can be well-approximated by a low-dimensional (usually linear) representation of basis functions. Value-based approaches typically struggle to plan efficient exploration and so cannot obtain efficient learning guarantees, although there has been interesting progress in this field as well [25].

2 Problem formulation

We consider the problem of learning to optimize a random finite horizon MDP $M = (\mathcal{S}, \mathcal{A}, R^M, P^M, \tau, \rho)$ in repeated finite episodes of interaction. This is the same formulation as earlier work [13], which we reproduce here for completeness. $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $R^M(s, a)$ is a probability distribution over $\mathbb{R}$ when selecting action $a$ while in state $s$, $P^M(s'|s, a)$ is the probability of transitioning to state $s'$ if action $a$ is selected while at state $s$, $\tau$ is the time horizon, and $\rho$ the initial state distribution. We define the MDP and all other random variables we will consider with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

A deterministic policy $\mu$ is a function mapping each state $s \in \mathcal{S}$ and $i = 1, \ldots, \tau$ to an action $a \in \mathcal{A}$. For each MDP $M = (\mathcal{S}, \mathcal{A}, R^M, P^M, \tau, \rho)$ and policy $\mu$, we define a value function

$$V^{M}_{\mu,i}(s) := E_{M,\mu} \left[ \sum_{j=i}^{\tau} R^M(s_j, a_j) \bigg| s_i = s \right],$$

where $R^M(s, a)$ denotes the expected reward realized when action $a$ is selected while in state $s$, and the subscripts of the expectation operator indicate that $a_j = \mu(s_j, j)$, and $s_{j+1} \sim P^M(s_{j+1}, a_{j+1})$ for $j = i, \ldots, \tau$. A policy $\mu$ is said to be optimal for MDP $M$ if $V^{M}_{\mu,i}(s) = \max_{a_{i+1}} V^{M}_{\mu,i+1}(s)$ for all $s \in \mathcal{S}$ and $i = 1, \ldots, \tau$. We will associate with each MDP $M$ a policy $\mu^M$ that is optimal for $M$.

The reinforcement learning agent interacts with the MDP over episodes that begin at times $t_k = (k-1)\tau + 1, k = 1, 2, \ldots$. At each time $t$, the agent selects an action $a_t$, observes a scalar reward $r_t$, and then transitions to $s_{t+1}$. If an agent follows a policy $\mu$ then when in state $s$ at time $t$ during episode $k$, it selects an action $a_t = \mu(s, t - t_k)$. Let $H_t = (s_1, a_1, r_1, \ldots, s_{t-1}, a_{t-1}, r_{t-1})$ denote the history of observations made prior to time $t$. A reinforcement learning algorithm is a deterministic sequence $\{\pi_k|k = 1, 2, \ldots \}$ of functions, each mapping $H_{t_k}$ to a probability distribution $\pi_k(H_{t_k})$ over policies. At the start of the $k$th episode, the algorithm samples a policy $\mu_k$ from the distribution $\pi_k(H_{t_k})$. The algorithm then selects actions $a_t = \mu_k(s_t, t - t_k)$ at times $t$ during the $k$th episode.

We define the regret incurred by a reinforcement learning algorithm $\pi$ up to time $T$ to be

$$\text{Regret}(T, \pi, M^*) := \sum_{k=1}^{\lceil T/\tau \rceil} \Delta_k,$$
where $\Delta_k$ denotes regret over the $k$th episode, defined with respect to the MDP $M^*$ by

$$\Delta_k := \sum_S \rho(s)(V^{M^*}_{\pi^*_k}(s) - V^{M^*}_{\mu_k}(s))$$

with $\mu^* = \mu^{M^*}$ and $\mu_k \sim \pi_k(H_{\mu_k})$. Note that regret is not deterministic since it can depend on the random MDP $M^*$, the algorithm’s internal random sampling and, through the history $H_{\mu_k}$, on previous random transitions and random rewards. We will assess and compare algorithm performance in terms of regret and its expectation.

## 3 Factored MDPs

To formalize our definition of a factored MDP we introduce some notation common to the literature [9].

**Definition 1** (Scope operation for factored sets $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_n$).

For any subset of indices $Z \subseteq \{1, 2, ..., n\}$ let us define the scope set $\mathcal{X}[Z] := \bigotimes_{i \in Z} \mathcal{X}_i$. Further, for any $x \in \mathcal{X}$ define the scope variable $x[Z] \in \mathcal{X}[Z]$ to be the value of the variables $x_i \in \mathcal{X}_i$ with indices $i \in Z$. For singleton sets $Z$ we will write $x[i]$ for $x[[i]]$ in the natural way.

Let $\mathcal{P}_{X,R}$ be the set of functions mapping elements of a finite set $\mathcal{X}$ to probability mass functions over a finite set $\mathcal{Y}$. $\mathcal{P}_{X,R}^{C,\sigma}$ will denote the set of functions mapping elements of a finite set $\mathcal{X}$ to $\sigma$-sub gaussian probability measures over $(\mathcal{R}, \mathcal{B}(\mathcal{R}))$ with mean bounded in $[0, C]$. We will consider factored reward and factored transition functions which are drawn from within these families.

**Definition 2** (Factored reward functions $R \in \mathcal{R} \subseteq \mathcal{P}_{X,R}^{C,\sigma}$).

The reward function class $\mathcal{R}$ is factored over $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_n$ with scopes $Z_1,..,Z_l$ if and only if, for all $R \in \mathcal{R}, x \in \mathcal{X}$ there exist functions $\{R_i \in \mathcal{P}_{\mathcal{X}[Z_i],\mathcal{R}}\}_{i=1}^l$ such that,

$$E[r] = \sum_{i=1}^l E[r_i]$$

where the observed reward $r \sim R(x)$ is equal to $\sum_{i=1}^l r_i$ with each $r_i \sim R_i(x[Z_i])$ and individually observed.

**Definition 3** (Factored transition functions $P \in \mathcal{P} \subseteq \mathcal{P}_{X,S}$).

The transition function class $\mathcal{P}$ is factored over $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_n$ and $\mathcal{S} = \mathcal{S}_1 \times ... \times \mathcal{S}_m$ with scopes $Z_1,..,Z_m$ if and only if, for all $P \in \mathcal{P}, x \in \mathcal{X}, s \in \mathcal{S}$ there exist some $\{P_i \in \mathcal{P}_{\mathcal{X}[Z_i],\mathcal{S}}\}_{i=1}^m$ such that,

$$P(s|x) = \prod_{i=1}^m P_i \left( s[i] \mid x[Z_i] \right)$$

A factored MDP (FMDP) is then defined to be an MDP with both factored rewards and factored transition functions. If we write $\mathcal{X} = \mathcal{S} \times \mathcal{A}$, then an FMDP is fully characterized by the tuple

$$M = (\{\mathcal{S}_i\}_{i=1}^m; \{\mathcal{X}_i\}_{i=1}^n; \{Z_i^R\}_{i=1}^l; \{R_i\}_{i=1}^l; \{Z_i^P\}_{i=1}^m; \{P_i\}_{i=1}^m; \tau; \rho),$$

where $Z_i^R$ and $Z_i^P$ are the scopes for the reward and transition functions respectively $\subseteq \{1,..,n\}$ which refer to $\mathcal{X}_i$. We assume that the size of all scopes $|Z_i| \leq \zeta \ll n$ and factors $|\mathcal{X}_i| \leq K$ so that the domains of $R_i$ and $P_i$ are of size at most $K^\zeta$.

## 4 Results

We present two algorithms, PSRL and UCRL-Factored with efficient regret bounds over factored MDPs. PSRL is guided by posterior sampling while UCRL-Factored uses optimism in the face of uncertainty. Full details of these algorithms are available in Section 6.

Our first result shows that we can bound the expected regret of PSRL.
Theorem 1 (Expected regret for PSRL in factored MDPs). Let $M^*$ be factored with graph structure $G = \{\mathcal{S}_i\}_{i=1}^m; \mathcal{X}_i, Z_i^R, Z_i^P; \tau\) if $\phi$ is the distribution of $M^*$ and $\Psi$ is the span of the optimal value function then we can bound the regret of PSRL:

$$E[\text{Regret}(T, \pi_{\tau}^{PS}, M^*)] \leq 4 + 2\sqrt{T} + \sum_{i=1}^l \left\{4(\tau C|X|Z_i^R|) + 1 + 8\sqrt{2|X|Z_i^R|T \log (4|X|Z_i^R|kT)}\right\}$$

$$+ E[\Psi \left(1 + \frac{4}{T-4}\right) \sum_{j=1}^m \left\{4(\tau|X|Z_j^P|) + 1 + 8\sqrt{2|X|Z_j^P|}|S_j|T \log (4m|X|Z_j^P|kT)}\right\}$$

We also show that using UCRL-Factored in a factored MDP we can bound the regret with high probability.

Theorem 2 (High probability regret for UCRL-Factored in factored MDPs). Let $M^*$ be factored with graph structure $G = \{\mathcal{S}_i\}_{i=1}^m; \mathcal{X}_i, Z_i^R, Z_i^P; \tau\) if $\phi$ is the distribution of $M^*$, then for any $M^*$ can bound the regret of UCRL-Factored:

$$\text{Regret}(T, \pi_{\tau}^{UC}, M^*) \leq C D \sqrt{2T \log (6/\delta)} + 2\sqrt{T} + \sum_{i=1}^l \left\{4(\tau C|X|Z_i^R|) + 1 + 8\sqrt{2|X|Z_i^R|T \log (12|X|Z_i^R|kT/\delta)}\right\}$$

$$+ C D \sum_{j=1}^m \left\{4(\tau|X|Z_j^P|) + 1 + 8\sqrt{2|X|Z_j^P|}|S_j|T \log (12m|X|Z_j^P|kT/\delta)}\right\}$$

with probability at least $1 - \delta$

For clarity, we present a symmetric problem instance for which we can produce a cleaner single-term upper bound. Let $Q$ be shorthand for the structure $G$ such that $l + 1 = m, C = \sigma = 1, |\mathcal{S}_i| = |\mathcal{X}_i| = K$ and $|Z_i^R| = |Z_i^P| = \zeta$ for all suitable $i$ and write $J = K^5$. In this case $\Psi, D \leq \tau$ trivially.

Corollary 1 (Clean bounds for PSRL in a symmetric problem). For an MDP with structure $Q$, if $\phi$ is the distribution of $M^*$ then we can bound the regret of PSRL:

$$E[\text{Regret}(T, \pi_{\tau}^{PS}, M^*)] \leq 15m\tau \sqrt{JKT \log (2mJT)}$$

Corollary 2 (Clean bounds for UCRL-Factored in a symmetric problem). For an MDP with structure $Q$, then for any $M^*$ we can bound the regret of UCRL-Factored:

$$\text{Regret}(T, \pi_{\tau}^{UC}, M^*) \leq 15m\tau \sqrt{JKT \log (12mJT/\delta)}$$

with probability at least $1 - \delta$.

These simply follow from the theorems above with loose upper bounds upon constant and logarithmic factors. The derivations are available in the Appendix B. Both algorithms satisfy bounds of $O(\tau m \sqrt{JKT})$ whereas a $Q$-naive algorithm gives $O(\tau \sqrt{JKT})$. We see that these new bounds are improved exponentially. These results are near optimal since for a factored MDP with $m$ independent components with $S$ states and $A$ actions we obtain regret bounds $O(mS \sqrt{A T})$, which is close to the lower bound of $\Omega(mS \sqrt{A T})$.

4.1 Interpreting regret bounds

The bounds for PSRL and UCRL-Factored are qualitatively similar and share much of the same analysis. For each algorithm, the regret is $\hat{O} \left(\Xi \sum_{j=1}^m \sqrt{|X|Z_j^P|}|S_j|T\right)$ where $\Xi$ is a measure of MDP connectedness, expected span $E[\Psi]$ for PSRL and scaled diameter $CD$ for UCRL-Factored.

The span of an MDP is defined $\Psi(M^*) := \max_{s, s'} \mu_{s, s'} \sum_{j=1}^m \sqrt{|X|Z_j^P|}|S_j|T\right)$ which is the maximum difference in expected value of any two states under the optimal policy. The diameter of an MDP $D(M^*) = \max_{s, s'} \min_{\mu_{s, s'}} \sum_{j=1}^m T_j^{s, s'}\right)$ where $T_j^{s, s'}$ is the expected number of steps to get from $s$ to $s'$ under policy $\mu$. It is always the case that $\Psi(M) \leq CD(M)$, otherwise one might improve the optimal policy from $s'$ to follow simply by taking the fastest policy to the $s$ with highest value. In some cases the span may be exponentially smaller than the diameter. In this sense PSRL satisfies a tighter bound.

However, UCRL-Factored has stronger probabilistic guarantees than PSRL since its bounds hold with high probability for any MDP $M^*$ not just in expectation. There is an optimistic algorithm REGAL [29] which formally
replaces the UCRL2 $D$ with $\Psi$ and retains the high probability guarantees. However, no practical implementation of that algorithm exists, even when given access to an MDP planner. An analogous extension to the analysis of REGAL-Factored is possible.

We should also note that UCRL2 was designed to obtain regret bounds even in MDPs without episodic reset. This is accomplished by imposing artificial episodes which end whenever the number of visits to a state-action pair is doubled [12]. Using a similar modification, it is possible to extend UCRL-Factored to this setting without trouble and retain similar regret bounds. However, this doubling trick in PSRL does not retain provable regret bounds, since the episode length is no longer independent of the sampled MDP. Nevertheless, there has been good empirical performance using this method for non-factored MDPs without episodic reset in simulation [13].

5 Confidence sets

Our analysis will rely upon the construction of confidence sets based around the empirical estimates for the underlying reward and transition functions. These confidence sets are chosen so that at the beginning of any episode $k$ the true and sampled functions are contained within the confidence set with high probability. We will then bound the deviation between the true function and elements in the confidence set by the maximal deviation within the confidence set. This technique is common to the literature and follows the same arguments as numerous previous papers on the subject [12, 26, 13].

Consider a family of functions $F \subseteq M_{X,Y}$ which takes $x \in X$ to a probability distribution over $(Y, \Sigma_Y)$ measurable space. We will write this as $M_{X,Y}$ unless we wish to stress the dependence on a particular $\sigma$-algebra which is not obvious.

Definition 4 (Set widths).

Let $X$ be a finite set, and let $(Y, \Sigma_Y)$ be a measurable space. The width of a set $F \in M_{X,Y}$ at $x \in X$ with respect to a norm $\| \cdot \|$ is

$$w_F(x) := \sup_{\hat{f} \in F} \| \hat{f} - f \|(x)$$

Our confidence set sequence $\{F_t \subseteq F : t \in \mathbb{N}\}$ is initialized with a set $F$. We adapt our confidence set to the observations $y_t \in Y$ which are drawn from the true function $f^* \in F$ at measurement points $x_t \in X$ so that $y_t \sim f^*(x_t)$. Each confidence set is then centered around an empirical estimate $\hat{f}_t \in M_{X,Y}$ at time $t$, defined by

$$\hat{f}_t(x) = \frac{1}{n_t(x)} \sum_{\tau < t : x_\tau = x} \delta_{y_\tau},$$

where $n_t(x)$ is the number of time $x$ appears in $(x_1, \ldots, x_{t-1})$ and $\delta_{y_\tau}$ is the probability mass function over $Y$ that assigns all probability to the outcome $y_\tau$. If at any time $t$, $n_t(x) = 0$ then we will let $\hat{f}_t(x)$ be any arbitrary function $\in M_{X,Y}$. Our sequence of confidence sets depends on our choice of norm $\| \cdot \|$ and a non-decreasing sequence $\{d_t : t \in \mathbb{N}\}$. For each $t$, the confidence set is defined by:

$$F_t = F_t(\| \cdot \|, x_1^{t-1}, d_t) := \left\{ f \in F : \| f - \hat{f}_t(x_i) \| \leq \sqrt{\frac{d_t}{n_t(x_i)}} \text{ for each } i = 1, \ldots, t-1 \right\}.$$

Where $x_1^{t-1}$ is shorthand for $(x_1, \ldots, x_{t-1})$ and we interpret $n_t(x_i) = 0$ as a null constraint which is satisfied $\forall f \in F$. The following result shows that we can bound the sum of confidence widths through time.

Theorem 3 (Bounding the sum of widths).

Let us write $F_k$ for $F_{t_k}$ and associate times within episodes of length $\tau$, $t = t_k + i$ for $i = 1, \ldots, \tau$ and $T = L \times \tau$. For all finite sets $X$, measurable spaces $(Y, \Sigma_Y)$, function classes $F \subseteq M_{X,Y}$ with uniformly bounded widths $w_F(x) \leq C_F \forall x \in X$ and non-decreasing sequences $\{d_t : t \in \mathbb{N}\}$:

$$\sum_{k=1}^{L} \sum_{i=1}^{\tau} w_{F_k}(x_{t_k+i}) \leq 4(C_F|X| + 1) + 4\sqrt{2d_T|X|T} \quad (5)$$

Proof. The proof follows from elementary considerations of $n_t(x)$ and the pigeonhole principle. We omit the details for brevity but refer the reader to Appendix A for a full derivation. \qed
6 Algorithms

Both algorithms require prior knowledge of \( \mathcal{G} = \{ \{ S_i \}_{i=1}^m; \{ X_i \}_{i=1}^n; \{ Z_i \}_{i=1}^l; \{ Z_i \}_{i=1}^m; \gamma \} \), the graphical structure of the FMDP. They also assume access to a “black box” that performs approximate dynamic programming for FMDPs. PSRL requires \( \Gamma(\cdot, \cdot) \) which takes a single MDP \( M \) and output an \( \epsilon \)-optimal policy for \( M \). UCRL-Factored requires \( \Gamma(\cdot, \cdot) \) which takes in a family of MDPs \( \mathcal{M} \) and outputs an \( \epsilon \)-optimal with respect to the most optimistic \( M \) in \( \mathcal{M} \). In general, it will be much more difficult to obtain an approximate solver \( \tilde{\Gamma} \) than \( \Gamma \).

PSRL remains identical to earlier treatment [27, 13] provided \( \phi \) is encoded in the prior \( \phi \). UCRL-Factored is essentially UCRL2 [12] modified to exploit \( \mathcal{G} \) in graph and episodic structure. We write \( \mathcal{R}_1^i(d_t^{R_i}) \) and \( \mathcal{P}_t^i(d_t^{P_i}) \) as shorthand for these confidence sets \( \mathcal{R}_1^i(|E[\cdot]|, x_1^{t-1}[Z_i^R], d_t^{R_i}) \) and \( \mathcal{P}_t^i(\| \cdot \|, x_1^{t-1}[Z_j^P], d_t^{P_i}) \) generated from initial sets \( \mathcal{R}_1 = \mathcal{P}_{\phi(t)} \) and \( \mathcal{P}_1 = \mathcal{P}_{\phi(0)} \).

Algorithm 1
PSRL (Posterior Sampling)

1: Input: Prior \( \phi \) encoding \( \mathcal{G}, t = 1 \)
2: for episodes \( k = 1, 2, \ldots \) do
3: sample \( M_k \sim \phi(H_k) \)
4: compute \( \mu_k = \Gamma(M_k, \sqrt{\tau/k}) \)
5: for timesteps \( j = 1, \ldots, \tau \) do
6: sample and apply \( a_t = \mu_k(s_t, j) \)
7: observe \( r_1^t, \ldots, r_t^t \) and \( s_{t+1}^t \)
8: \( t = t + 1 \)
9: end for
10: end for

Algorithm 2
UCRL-Factored (Optimism)

1: Input: Graph structure \( \mathcal{G} \), confidence \( \delta, t = 1 \)
2: for episodes \( k = 1, 2, \ldots \) do
3: \( d_t^{R_i} = 4\sigma^2 \log (4k|\mathcal{X}^R| |k/\delta|) \) for \( i = 1, \ldots, l \)
4: \( d_t^{P_i} = 4|\mathcal{S}_j| \log (4m|\mathcal{X}^P| |k/\delta|) \) for \( j = 1, \ldots, m \)
5: \( M_k = \{ M \in \mathcal{G}, R_t \in \mathcal{R}_1(d_t^{R_t}), P_t \in \mathcal{P}_t^{P_t} \} \forall i,j \}
6: compute \( \mu_k = \tilde{\Gamma}(M_k, \sqrt{\tau/k}) \)
7: for timesteps \( u = 1, \ldots, \tau \) do
8: sample and apply \( a_t = \mu_k(s_t, u) \)
9: observe \( r_1^t, \ldots, r_t^t \) and \( s_{t+1}^t \)
10: \( t = t + 1 \)
11: end for
12: end for

The parameters \( d_t^{R_i} \) and \( d_t^{P_i} \) are chosen to satisfy concentration inequalities so that the true MDP \( M^* \) lies within \( M_k \) for all \( k \) with high probability. Although PSRL makes no mention of confidence sets, \( M_k \) will also be useful in the analysis of PSRL.

7 Analysis

We will now piece together the necessary analysis for our main results. First we recap the analysis of PSRL and UCRL2 which allow us to the regret to the bellman error. Next we show that, for factored MDPs, it is possible to bound this estimation error by the error in each factored component separately. From here we will use concentration inequalities upon the individual factors to show that, with high probability, the true MDP \( M^* \) lies within \( M_k \) for all \( k \). The final results will then be obtained through an application of Theorem 3.

7.1 From regret to Bellman error

A key difficulty in providing regret bounds for reinforcement learning is that it depends upon the rewards of the optimal policy \( \mu^* \). For many reinforcement learning algorithms there is no clean way to relate the unknown optimal policy to the states and actions observed by the agent. Using the OFU principle, we can guarantee with high probability that the optimal rewards of the true MDP are upper bounded by the optimal rewards of the optimistic MDP [11]. In the case of posterior sampling, we make use of the posterior sampling lemma [17]

Lemma 1 (Posterior Sampling).
If \( \phi \) is the distribution of \( M^* \) then, for any \( \sigma(H_k) \)-measurable function \( g \),

\[
E[g(M^*)|H_k] = E[g(M_k)|H_k].
\]

(6)
Note that taking the expectation of \( \mathbb{E}[g(M^*)] = \mathbb{E}[g(M_{\hat{k}})] \) through the tower property. We introduce the Bellman operator \( \mathcal{T}_M^\mu \), which for any MDP \( M = (\mathcal{S}, \mathcal{A}, R^M, P^M, \tau, \rho) \), stationary policy \( \mu : \mathcal{S} \rightarrow \mathcal{A} \) and value function \( V : \mathcal{S} \rightarrow \mathbb{R} \), is defined by
\[
\mathcal{T}_M^\mu V(s) := R^M(s, \mu(s)) + \sum_{s' \in \mathcal{S}} P^M(s'|s, \mu(s))V(s').
\]
This returns the expected value of state \( s \) where we follow the policy \( \mu \) under the laws of \( M \), for one time step. The following lemma gives a concise form for the dynamic programming paradigm in terms of the Bellman operator.

**Lemma 2** (Dynamic programming equation).

For any MDP \( M = (\mathcal{S}, \mathcal{A}, R^M, P^M, \tau, \rho) \) and policy \( \mu : \mathcal{S} \times \{1, \ldots, \tau\} \rightarrow \mathcal{A} \), the value functions \( V^M_\mu \) satisfy
\[
V^M_{\mu, i} = \mathcal{T}^M_{\mu, i} V^M_{\mu, i+1}
\]
for \( i = 1 \ldots, \tau \), with \( V^M_{\mu, \tau+1} = 0 \).

In order to streamline our common analysis of PSRL and UCRL2 we will let \( \hat{M}_k \) refer generally to either the sampled MDP used in PSRL or the optimistic MDP chosen from \( \mathcal{M}_k \) with associated near-optimal policy \( \hat{\mu}_k \). We will streamline our discussion of \( P^M, R^M, V^M_{\hat{\mu}, i} \) and \( \mathcal{T}^M_{\mu, i} \) by simply writing \( \star \) in place of \( M^\star \) or \( \mu^\star \) and \( k \) in place of \( \hat{M}_k \) or \( \hat{\mu}_k \) where appropriate; for example \( V^M_{k, i} := V^M_{\hat{\mu}, i} \). We will also write \( x_{k, i} := (s_{k, i}, \mu(s_{k, i+1})) \).

We now break down the regret by adding and subtracting the \( \text{imagined} \) near optimal reward of policy \( \hat{\mu}_k \), which is known to the agent. For clarity of analysis we consider only the case of \( \rho(s') = 1\{s' = s\} \) but this changes nothing for our consideration of finite \( \mathcal{S} \).

\[
\Delta_k = V^*_{k, 1}(s) - V^*_{k, 1}(s) = \left( V^k_{k, 1}(s) - V^*_{k, 1}(s) \right) + \left( V^*_{k, 1}(s) - V^k_{k, 1}(s) \right)
\]

The second term \( V^*_{k, 1} - V^*_{k, 1} \) relates the optimal rewards of the imagined MDP \( M^\star \) to those near optimal for \( \hat{M}_k \). We can bound this difference by \( \sqrt{1/k} \) for PSRL in expectation, and for UCRL-Factored with high probability. This follows for PSRL by Lemma 1 and for UCRL-Factored whenever the true MDP lies within the confidence set \( \mathcal{M}_k \) by the principle of OFU and the approximate MDP planner \( \Gamma \).

We decompose the first term through repeated applications of the dynamic programming equation,
\[
(V^k_{k, 1} - V^*_{k, 1})(s_{k, 1}) = \sum_{i=1}^\tau (\mathcal{T}^k_{k, i} - \mathcal{T}^*_{k, i}) V^k_{k, i+1}(s_{k, i+1}) + \sum_{i=1}^\tau d_{k, i+1}.
\]
Where \( d_{k, i+1} := \sum_{s \in \mathcal{S}} \left\{ P^*(s|x_{k, i})(V^*_{k, i+1} - V^k_{k, i+1})(s) \right\} - (V^*_{k, i+1} - V^k_{k, i+1})(s_{k, i+1}) \). The second term captures the randomness in the transitions of the true MDP \( M^* \) under policy \( \mu \). The expected value of \( (V^k_{k, i+1} - V^*_{k, i+1})(s_{k, i+1}) \) given \( H_{k, i+1} \) is precisely \( \sum_{s \in \mathcal{S}} \left\{ P^*(s|x_{k, i})(V^*_{k, i+1} - V^k_{k, i+1})(s) \right\} \) so that \( \mathbb{E}[d_{k, i+1}] = 0 \) for all \( i \). To obtain high probability bounds for UCRL-Factored we note that \( d_{k, i+1} \) martingale difference bounded by \( \Psi_k \) equal to the span of \( V^k_{k, i} \). By the OFU principle we know that \( \Psi_k \leq CD \). We apply the Azuma-Hoeffding inequality to say that:
\[
P\left( \sum_{k=1}^m \sum_{i=1}^\tau d_{k, i+1} > CD \sqrt{2T \log(2/\delta)} \right) \leq \delta
\]

The remaining term is the one step Bellman error of the imagined MDP \( \hat{M}_k \). Crucially this term only depends on \( x_{k, i} \) which are actually observed. We can now use the Hölder inequality to bound
\[
\sum_{i=1}^\tau (\mathcal{T}^k_{k, i} - \mathcal{T}^*_{k, i}) V^k_{k, i+1}(s_{k, i+1}) \leq \sum_{i=1}^\tau |\mathcal{R}^k(x_{k, i}) - \mathcal{R}^*_{k, i})| + \frac{1}{2} \Psi_k \|P^k(\cdot|x_{k, i}) - P^*(\cdot|x_{k, i})\|_1
\]

**7.2 Factorization decomposition**

So far our analysis has made no mention of the factorized structure of the MDP. We will now show how we can further bound equation (1) by the sums of errors in each factor of the reward and transition functions. It is quite clear that we may upper bound the deviations of \( \mathcal{R}^k, \mathcal{R}^* \) by the sum of deviations of their factors using the triangle inequality. In fact, as we show in Lemma 3 we can also do this for the transition functions \( P^* \) and \( P^k \). This result really is the key insight for our results in this paper.
Lemma 3 (Bounding factored deviations).
Let the transition function class \( \mathcal{P} \subseteq \mathcal{P}_{\mathcal{X}, \mathcal{S}} \) be factored over \( \mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \) and \( \mathcal{S} = \mathcal{S}_1 \times \ldots \times \mathcal{S}_m \) with scopes \( Z_1, \ldots, Z_m \). Then, for any \( P, \hat{P} \in \mathcal{P} \) we may bound their L1 distance by the sum of the differences of their factorizations:

\[
\| P(x) - \hat{P}(x) \|_1 \leq \sum_{i=1}^{m} \| P_i(x[Z_i]) - \hat{P}_i(x[Z_i]) \|_1
\]

Proof. In order to prove this lemma we begin with the simple claim that for any \( \alpha_1, \alpha_2, \beta_1, \beta_2 \in (0, 1) \):

\[
|\alpha_1 \alpha_2 - \beta_1 \beta_2| = \alpha_2 \left| \alpha_1 - \frac{\beta_1 \beta_2}{\alpha_2} \right| \\
\leq \alpha_2 \left( |\alpha_1 - \beta_1| + \left| \frac{\beta_1 \beta_2}{\alpha_2} \right| \right) \\
\leq \alpha_2 |\alpha_1 - \beta_1| + \beta_1 |\alpha_2 - \beta_2|
\]

This result also holds for any \( \alpha_1, \alpha_2, \beta_1, \beta_2 \in [0, 1] \), where including 0 can be verified case by case.

We now consider the probability distributions \( p, \tilde{p} \) over \( \{1, \ldots, d_1\} \) and \( q, \tilde{q} \) over \( \{1, \ldots, d_2\} \). We let \( Q = p q^T, \tilde{Q} = \tilde{p} \tilde{q}^T \) be the joint probability distribution over \( \{1, \ldots, d_1\} \times \{1, \ldots, d_2\} \). Using the claim above we will be able to bound the L1 deviation \( \|Q - \tilde{Q}\|_1 \) by the deviations of their factors:

\[
\|Q - \tilde{Q}\|_1 = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |q_i q_j - \tilde{p}_i \tilde{q}_j| \\
\leq \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |q_i - \tilde{p}_i| + |\tilde{p}_i| - |\tilde{q}_j| \\
= \|p - \tilde{p}\|_1 + \|q - \tilde{q}\|_1
\]

Applying this result \( m \) times to the factored transition functions \( P \) and \( \hat{P} \) we recover our desired result. \(\square\)

7.3 Concentration guarantees for \( \mathcal{M}_k \)
We now want to show that the true MDP lies within \( \mathcal{M}_k \) with high probability. Note that posterior sampling will also allow us to then say that the sampled \( \mathcal{M}_k \) is within \( \mathcal{M}_k \) with high probability too. In order to show this, we first present a concentration result for the L1 deviation of empirical probabilities.

Lemma 4 (L1 bounds for the empirical transition function).
For all finite sets \( \mathcal{X} \), finite sets \( \mathcal{Y} \), function classes \( \mathcal{P} \subseteq \mathcal{P}_{\mathcal{X}, \mathcal{Y}} \) then for any \( x \in \mathcal{X}, \epsilon > 0 \):

\[
P \left( \| P^*(x) - \hat{P}_t(x) \|_1 \geq \epsilon \right) \leq \exp \left( |\mathcal{Y}| \log(2) - \frac{n_t(x) \epsilon^2}{2} \right)
\]

Proof. This is a relaxation of the result proved by Weissman [28]. \(\square\)

Lemma 4 ensures that for any \( x \in \mathcal{X}, j = 1, \ldots, m \) \( P \left( \| P_j^*(x) - \hat{P}_t(x) \|_1 \geq \sqrt{2|S| \log \left( \frac{2}{\delta'} \right)} \right) \leq \delta'. \) We then define \( d_{k,j}^j = 2|S| \log \left( \frac{2}{\delta'_{k,j}} \right) \) with \( \delta'_{k,j} = \delta/(2m|\mathcal{X}|^2 |Z_j^p| k^2) \). Now using a union bound, together with the fact that \( \sum_{n=1}^{\infty} 1/n^2 = \pi^2/6 < 2 \):

\[
P \left( P_j^* \in \mathcal{P}_j^j(d_{k,j}^j) \forall k \in \mathbb{N}, j = 1, \ldots, m \right) \geq 1 - \delta
\]

The proof for sub \( \sigma \)-gaussian random variables follows from the definition and resultant tail bounds. \( \epsilon \in \mathbb{R} \) is a sub \( \sigma \)-gaussian random variable \( \iff \forall t \in \mathbb{R}, \ E[\exp(\epsilon t)] \leq \exp \left( \frac{\epsilon^2 t^2}{2} \right) \).

Lemma 5 (Tail bounds for sub \( \sigma \)-gaussian random variables).
If \( \{\epsilon_i\} \) are all independent and sub \( \sigma \)-gaussian then \( \forall \beta \geq 0: \ P \left( \frac{1}{n} \sum_{i=1}^{n} |\epsilon_i| > \beta \right) \leq \exp \left( \log(2) - \frac{n \beta^2}{25 \pi^2} \right) \).
Lemma 5 ensures that for any $x \in X$, $i = 1, ..., l \ P\left( |\mathcal{R}_i(x) - \hat{\mathbb{E}}_i(x)| \geq \sqrt{\frac{2\pi^2}{m_i(x)} \log \left( \frac{2\pi^2}{m_i(x)} \right)} \right) \leq \delta'$. A similar argument from here ensures that: $P\left( \mathcal{R}_i \in \mathcal{R}_i(x) \forall k \in \mathbb{N}, i = 1, ..., l \right) \geq 1 - \delta$. This allows us to present our important result that

$$\mathbb{P}\left( M^* \in \mathcal{M}_k \forall k \in \mathbb{N} \right) \geq 1 - 2\delta$$

(12)

7.4 Regret bounds

We now have all the necessary intermediate results to complete our proof. We begin with the analysis of PSRL. Using equation (12) and the posterior sampling lemma we can say that $P(M^*, M_k \in \mathcal{M}_k \forall k \in \mathbb{N}) \geq 1 - 4\delta$. The contributions from near-optimal regret in planning function $\Gamma$ are bounded by $\sum_{k=1}^l \sqrt{\pi_k} \leq 2\sqrt{T}$. From here we take equation (11), Lemma 3 and Theorem 3 to say that for any $\delta > 0$:

$$E[\text{Regret}(T, \pi^*_T, M^*)] \leq 4\delta T + 2\sqrt{T} + \sum_{i=1}^l \left\{ 4(\tau C |X| Z^R_i) + 1 \right\} + 4\sqrt{2d^{R_i}_T |X| Z^R_i |T|}$$

+ $\sup_{k=1, ..., L} (E[\Psi | M_k, M^* \in \mathcal{M}_k]) \times \sum_{j=1}^m \left\{ 4(\tau |X| Z^R_j) + 1 \right\} + 4\sqrt{2d^{R_j}_T |X| Z^R_j |T|}$$

Let $A = \{ M^*, M_k \in \mathcal{M}_k \}$, since $\Psi_k \geq 0$ and $E[\Psi | k] = E[\Psi]$ via posterior sampling we can say that for all $k$:

$$E[\Psi | A] \leq P(A)^{-1} E[\Psi] \leq \left( 1 - \frac{4\delta}{k^2} \right)^{-1} E[\Psi] = \left( 1 + \frac{4\delta}{k^2 - 4\delta} \right) E[\Psi] \leq \left( 1 + \frac{4\delta}{1 - 4\delta} \right) E[\Psi]$$

Plugging in the values of $d^{R_i}_T$ and $d^{R_j}_T$ and setting $\delta = 1/T$ completes the proof of Theorem 1.

The analysis of UCRL-Factored and Theorem 2 follows in a similar manner. We use equation (10) to bound the contributions of $d_i$ with probability $1 - \delta$, and equation (12) to bound missed confidence regions. From here we take equation (11), Lemma 3 and Theorem 3 to say that, with probability $\geq 1 - 3\delta$:

$$\text{Regret}(T, \pi^*_T, M^*) \leq CD \sqrt{2T \log(2/\delta)} + 2\sqrt{T} + \sum_{i=1}^l \left\{ 4(\tau C |X| Z^R_i) + 1 \right\} + 4\sqrt{2d^{R_i}_T |X| Z^R_i |T|}$$

+ $CD \times \sum_{j=1}^m \left\{ 4(\tau |X| Z^R_j) + 1 \right\} + 4\sqrt{2d^{R_j}_T |X| Z^R_j |T|}$

and 2 The Corollaries 1 and 2 follow from simply plugging in these values in the symmetric case and upper bounding the constant and logarithmic terms. This is presented in more detail in Appendix 13.

References

[1] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov decision processes. Mathematics of Operations Research, 22(1):222–255, 1997.

[2] P. R. Kumar and P. Varaiya. Stochastic systems: estimation, identification and adaptive control. Prentice-Hall, Inc., 1986.

[3] Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. Stochastic dynamic programming with factored representations. Artificial Intelligence, 121(1):49–107, 2000.

[4] Zoubin Ghahramani. Learning dynamic bayesian networks. In Adaptive processing of sequences and data structures, pages 168–197. Springer, 1998.

[5] Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored MDPs. In IJCAI, volume 16, pages 740–747, 1999.

[6] M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. Machine Learning, 49(2-3):209–232, 2002.

[7] Alexander Strehl. Model-based reinforcement learning in factored-state MDPs. In Approximate Dynamic Programming and Reinforcement Learning, 2007. ADPRL 2007. IEEE International Symposium on, pages 103–110. IEEE, 2007.
[8] R. I. Brafman and M. Tennenholtz. R-max—a general polynomial time algorithm for near-optimal reinforcement learning. The Journal of Machine Learning Research, 3:213–231, 2003.

[9] István Szita and András Lörincz. Optimistic initialization and greediness lead to polynomial time learning in factored MDPs. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1001–1008. ACM, 2009.

[10] Alexander Strehl, Carlos Diuk, and Michael Littman. Efficient structure learning in factored-state MDPs. In AAAI, volume 7, pages 645–650, 2007.

[11] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. arXiv preprint cs/9605103, 1996.

[12] Thomas Jakobch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. The Journal of Machine Learning Research, 9:1563–1600, 2010.

[13] Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) Efficient Reinforcement Learning via Posterior Sampling. Advances in Neural Information Processing Systems, 2013.

[14] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3/4):285–294, 1933.

[15] O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In Neural Information Processing Systems (NIPS), 2011.

[16] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: an asymptotically optimal finite time analysis. In International Conference on Algorithmic Learning Theory, 2012.

[17] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. CoRR, abs/1301.2609, 2013.

[18] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored MDPs. J. Artif. Intell. Res.(JAIR), 19:399–468, 2003.

[19] Daphne Koller and Ronald Parr. Policy iteration for factored MDPs. In Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, pages 326–334. Morgan Kaufmann Publishers Inc., 2000.

[20] Carlos Guestrin, Daphne Koller, and Ronald Parr. Max-norm projections for factored MDPs. In IJCAI, volume 1, pages 673–682, 2001.

[21] Karina Valdivia Delgado, Scott Sanner, and Leliane Nunes De Barros. Efficient solutions to factored MDPs with imprecise transition probabilities. Artificial Intelligence, 175(9):1498–1527, 2011.

[22] Scott Sanner and Craig Boutiller. Approximate linear programming for first-order MDPs. arXiv preprint arXiv:1207.1415, 2012.

[23] Ronald Ortner, Danil Ryabko, et al. Online regret bounds for undiscounted continuous reinforcement learning. In NIPS, pages 1772–1780, 2012.

[24] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. Journal of Machine Learning Research-Proceedings Track, 19:1–26, 2011.

[25] Benjamin Van Roy and Zheng Wen. Generalization and exploration via randomized value functions. arXiv preprint arXiv:1402.0635, 2014.

[26] P. L. Bartlett and A. Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pages 35–42. AUAI Press, 2009.

[27] M. Strens. A Bayesian framework for reinforcement learning. In Proceedings of the 17th International Conference on Machine Learning, pages 943–950, 2000.

[28] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L1 deviation of the empirical distribution. Hewlett-Packard Labs, Tech. Rep, 2003.
A Bounding the widths of confidence sets

We present elementary arguments which culminate in a proof of Theorem 3.

Lemma 6 (Concentration results for $\sqrt{d_T/n_t(x)}$).
For all finite sets $\mathcal{X}$ and any $d_T, \epsilon \geq 0$:

$$\sum_{t=1}^T 1\{\sqrt{d_T/n_t(x_t)} > h(d_T, \epsilon)\} \leq \sum_{t=1}^T 1\{\sqrt{d_T/n_t(x_t)} > \epsilon\} + |\mathcal{X}|,$$

Where $h(d_T, \epsilon) := \sqrt{d_T \epsilon^2/(d_T + \epsilon^2)}$.

Proof. Let $(x_{s_1}, \ldots, x_{s_K})$ be the largest subsequence of $x_t^T$ such that $\sqrt{d_T/n_t(x_t)} \in (h(d_T, \epsilon), \epsilon) \forall i$. Now for any $x \in \mathcal{X}$, let $T_x = \{s_i | x_{s_i} = x\}$. Suppose there exist two distinct elements $\sigma, \rho \in T_x$ with $\sigma < \rho$ so that $n_\rho(x) \geq n_\sigma(x) + 1$. We note that for any $n \in \mathbb{R}_+$, $h(d_T, \sqrt{d_T/n}) = \sqrt{d_T/(n+1)}$ so that:

$$\epsilon \geq \sqrt{d_T/n_\sigma(x)} \implies h(d_T, \epsilon) \geq \sqrt{d_T/(n_\sigma(x) + 1)} \geq \sqrt{d_T/n_\rho(x)}$$

This contradicts our assumption $\sqrt{d_T/n_\rho(x)} \in (h(d_T, \epsilon), \epsilon)$ and so we must conclude that $|T_x| \leq 1$ for all $x \in \mathcal{X}$. This means that $(x_{s_1}, \ldots, x_{s_K})$ forms a subsequence of unique elements in $\mathcal{X}$, the total length of which must be bounded by $|\mathcal{X}|$. \qed

We now provide a corollary of this result which allows for episodic delays in updating visit counts $n_t(x)$. We imagine that we will only update our counts every $\tau$ steps.

Corollary 3 (Concentration results for $\sqrt{d_T/n_t(x)}$ in the episodic setting).
Let us associate times within episodes of length $\tau$, $t = t_k + i$ for $i = 1, \ldots, \tau$ and $T = M \times \tau$. For all finite sets $\mathcal{X}$ and any $d_T, \epsilon \geq 0$:

$$\sum_{k=1}^M \sum_{i=1}^\tau 1\{\sqrt{d_T/n_t(x_{t_k+i})} > h^{(\tau)}(d_T, \epsilon))\} \leq \sum_{k=1}^M \sum_{i=1}^\tau 1\{\sqrt{d_T/n_t(x_{t_k+i})} > \epsilon\} + 2\tau|\mathcal{X}|,$$

Where $h^{(\tau)}(d_T, \epsilon)$ is the $\tau$-fold composition of $h(d_T, \cdot)$ acting on $\epsilon$.

Proof. By an argument of visiting times similar to lemma 6 we can see that the worst case scenario for the episodic case $\sum_{k=1}^M \sum_{i=1}^\tau 1\{\sqrt{d_T/n_t(x_{t_k+i})} > h^{(\tau)}(d_T, \epsilon))\}$ is to visit each $x$ exactly $\tau - 1$ times before the start of an episode, and then spend the entirety of the following episode within the state. Here we have upper bounded $2\tau - 1$ by $2\tau$ and $|\mathcal{X}| - 1$ by $|\mathcal{X}|$ to complete our result. \qed

It will be useful to define notion of radius for each confidence set at each $x \in \mathcal{X}$, $r_{F_t}(x) := \sup_{f \in F_t} \|(f - \hat{f}_t)(x)\|$. By the triangle inequality, we have $w_{F_t}(x) \leq 2r_{F_t}(x)$ for all $x \in \mathcal{X}$.

Lemma 7 (Bounding the number of large radii).
Let us write $F_t$ for $F_{t_k}$ and associate times within episodes of length $\tau$, $t = t_k + i$ for $i = 1, \ldots, \tau$ and $T = M \times \tau$. For all finite sets $\mathcal{X}$, measurable spaces $(\mathcal{Y}, \mathcal{S}_y)$, function classes $\mathcal{F} \subseteq \mathcal{M}_{\mathcal{X}, \mathcal{Y}}$, non-decreasing sequences $\{d_t : t \in \mathbb{N}\}$, any $T \in \mathbb{N}$ and $\epsilon > 0$:

$$\sum_{k=1}^M \sum_{i=1}^\tau 1\{r_{F_t}(x_{t_k+i}) > \epsilon\} < \left(\frac{d_T}{2\epsilon^2} + 1\right)2\tau|\mathcal{X}|$$

Proof. By construction of $F_t$ and noting that $d_t$ is non-decreasing in $t$, we can say that $r_{F_t}(x_t) \leq \sqrt{d_T/n_t(x_t)}$ for all $t = 1, \ldots, T$ so that:

$$\sum_{k=1}^M \sum_{i=1}^\tau 1\{r_{F_t}(x_{t_k+i}) > \epsilon\} \leq \sum_{k=1}^M \sum_{i=1}^\tau 1\{\sqrt{d_T/n_t(x_{t_k+i})} > \epsilon\}.$$ 

Now let $g(\epsilon) = \sqrt{d_T \epsilon^2/(d_T - \tau \epsilon^2)}$ be the $\epsilon$-inverse of $h^{(\tau)}(d_T, \epsilon)$ such that $g(h^{(\tau)}(d_T, \epsilon)) = \epsilon$. Applying Corollary 3 to our expression $n$ times repeatedly we can say:

$$\sum_{k=1}^M \sum_{i=1}^\tau 1\{\sqrt{d_T/n_t(x_{t_k+i})} > \epsilon\} \leq \sum_{k=1}^M \sum_{i=1}^\tau 1\{\sqrt{d_T/n_t(x_{t_k+i})} > g^{(\tau)}(\epsilon)\} + 2\tau|\mathcal{X}|.$$ 

Where $g^{(\tau)}(\epsilon)$ denotes the composition of $g(\cdot)$ $n$-times acting on $\epsilon$. If we take $n$ to be the lowest integer such that $g^{(n)}(\epsilon) > \sqrt{d_T/\tau}$ then $\sum_{k=1}^M \sum_{i=1}^\tau 1\{\sqrt{d_T/n_t(x_{t_k+i})} > g^{(n)}(\epsilon)\} \leq 2\tau|\mathcal{X}|$ so that the whole expression is bounded by $(n + 1)2\tau|\mathcal{X}|$. Note that for all $N \in \mathbb{R}_+$, $g(\sqrt{d_T/N}) = \sqrt{d_T/(N - \tau)}$, if we write $\epsilon = \sqrt{d_T/N}$ then $n \leq N_1/\tau = \frac{d_T}{\epsilon^2}$, which completes the proof. \qed
Using these results we are finally able to complete our proof of Theorem 3. We first note that, via the triangle inequality

\[ \sum_{k=1}^{M} \sum_{t=1}^{T} w_{k,t} (x_{k,t+1}) \leq 2 \sum_{k=1}^{M} \sum_{t=1}^{T} r_{k,t} (x_{k,t+1}). \]

We streamline our notation by letting \( r_{k,i} = r_{k,i} (x_{k_{t+1}}) \). Reordering the sequence \( (r_{1,1}, ..., r_{n,T}) \rightarrow (r_{1,1}, ..., r_{n,T}) \) such \( i \geq r_{i,n} \geq \tau_{i,n} \) we have that:

\[ \sum_{k=1}^{M} \sum_{t=1}^{T} r_{k,t} (x_{k,t+1}) = \sum_{t=1}^{T} r_{i,t} \leq 1 + \sum_{i=1}^{T} r_{i,t} \{ r_{i,t} \geq T^{-1} \}. \]

We can see that \( r_{i,t} > \epsilon \geq T^{-1} \iff \sum_{i=1}^{T} r_{i,t} \geq \epsilon \geq t \). From Lemma 7 this means that \( t \leq \left( \frac{d_{r}}{\tau_{r}} + 1 \right) 2 \tau_{r} |X| \), so that \( \epsilon \leq \sqrt{\frac{2 |X| d_{r}}{\tau_{r} |X|}} \). This means that \( r_{i,t} \leq \min\{C_{r}, \sqrt{\frac{2 |X| d_{r}}{\tau_{r} |X|}} \} \). Therefore,

\[ \sum_{i=1}^{T} r_{i,t} \{ r_{i,t} \geq T^{-1} \} \leq 2 \tau_{r} |X| + \sum_{t=2}^{T} \sqrt{\frac{2 d_{r} |X|}{t - \tau_{r} |X|}} \leq 2 \tau_{r} |X| + 2 \sqrt{2 d_{r} |X| T} \]

Which completes the proof of Theorem 3.

B Clean bounds for the symmetric problem

We now provide concrete clean upper bounds for Theorems 1 and 2 in the simple symmetric case \( t + 1 = m, C = \sigma = 1, |S_i| = |X| = K \) and \( |Z^*_i| = |Z^*_T| = \zeta \) for all suitable \( i \) and write \( J = K^C \). For a non-trivial problem setting we assume that \( K \geq 2, m \geq 2, \tau \geq 2 \).

From Section 7, we have that

\[ \mathbb{E} \left[ \text{Regret}(T, \pi_{r}^{\text{PS}}, M^*) \right] \leq 4 + 2 \sqrt{T} + m \left\{ 4 (r J + 1) + 4 \sqrt{8 \log(4 m J T^2 / \tau)} \right\} \mathbb{E} [\Psi] \left( 1 + \frac{4}{T - 4} \right) \left\{ 4 (r J + 1) + 4 \sqrt{8 K \log(4 m J T^2 / \tau)} \right\} \]

Through looking at the constant term we know that the bounds are trivially satisfied for all \( T \leq 56 \), from here we can certainly upper bound \( 4/(T - 4) \leq 1/13 \). From here we can say that:

\[ \mathbb{E} \left[ \text{Regret}(T, \pi_{r}^{\text{UC}, M^*}) \right] \leq \left\{ 4 + 4 m \left( 1 + \frac{14}{13} \mathbb{E} [\Psi] \right) (r J + 1) \right\} + \sqrt{T} \left\{ 2 + 4 \sqrt{8 J \log(4 m J T^2 / \tau)} + 4 \sqrt{8 K \log(4 m J T^2 / \tau)} \right\} \]

\[ \leq 5 \left( 1 + \mathbb{E} [\Psi] \right) m \tau J + J \log(2 m J T) + 2 \sqrt{8 J \log(4 m J T^2 / \tau)} \log(2 m J T) \]

\[ \leq 5 \left( 1 + \mathbb{E} [\Psi] \right) m \tau J + 12 m (1 + \mathbb{E} [\Psi] \sqrt{K}) \sqrt{J T \log(2 m J T)} \]

\[ \leq \min(5 m \tau J, J T) + 2 m \tau J \sqrt{f K T \log(2 m J T)} \]

\[ \leq 15 m \tau J K T \log(2 m J T) \]

Where in the last steps we have used that \( \Psi \leq \tau \) and \( \min(a, b) \leq \sqrt{ab} \). We now repeat a similar procedure of upper bounds for UCRL-Factored, immediately replicating \( D \) by \( \tau \) in our analysis to say that with probability \( \geq 1 - 3 \delta \):

\[ \text{Regret}(T, \pi_{r}^{\text{UC}, M^*}) \leq \tau \sqrt{2 T \log(2 / \delta)} + 2 \sqrt{T} + m \left\{ 4 (r J + 1) + 4 \sqrt{8 \log(4 m J T / \delta)} \right\} \]

\[ + \tau m \left\{ 4 (r J + 1) + 4 \sqrt{8 K \log(4 m J T / \delta)} \right\} \]

\[ \leq (1 + \tau) m \left( 4 (r J + 1) + \sqrt{T} \right) \left\{ 2 \log(2 / \delta) + 2 + m \sqrt{8 \log(4 m J T / \delta)} + \tau m \sqrt{8 \log(4 m J T / \delta)} \right\} \]

\[ \leq 5 (1 + \tau) m \tau J + 12 m (1 + \tau \sqrt{K}) \sqrt{J T \log(4 m J T / \delta)} \]

\[ \leq 15 m \tau J K T \log(4 m J T / \delta) \]

Where in the last step we used a similar argument.