Improved Analysis of UCRL2B

Ronan Fruit, Matteo Pirotta and Alessandro Lazaric

March 22, 2019

UCRL2B is a variant of UCRL2 (Jaksch et al., 2010) that construct confidence intervals based on the empirical Bernstein inequality (Audibert et al., 2007) rather than Hoeffding's inequality. Let

$$\beta_{p,k}^{sas'} := 2\sqrt{\frac{\widehat{\sigma}_{p,k}^2(s'|s,a)}{N_k^+(s,a)} \ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)} + \frac{6\ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)}{N_k^+(s,a)} \tag{1}$$

$$\beta_{r,k}^{sa} := 2\sqrt{\frac{\widehat{\sigma}_{r,k}^2(s,a)}{N_k^+(s,a)} \ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)} + \frac{6r_{\max}\ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)}{N_k^+(s,a)} \tag{2}$$

where $\widehat{\sigma}_{p,k}^2$ and $\widehat{\sigma}_{r,k}^2$ are the population variance of transition and reward function at episode k. then we define $\mathcal{M}_k := \{\mathcal{S}, \mathcal{A}, r_k, p_k\}$ to be the extended MDP defined by the confidence intervals

$$p_k(s'|s,a) \in B_p^k(s,a,s') := \left[\widehat{p}_k(s'|s,a) - \beta_{p,k}^{sas'}, \widehat{p}_k(s'|s,a) + \beta_{p,k}^{sas'}\right] \cap [0,1] \tag{3}$$

$$r_k(s,a) \in B_r^k(s,a) := \left[\widehat{r}_k(s,a) - \beta_{r,k}^{sa}, \widehat{r}_k(s,a) + \beta_{r,k}^{sa}\right] \cap [0, r_{\max}] \tag{4}$$

Then, see App. B.2:

$$\mathbb{P}\left(\exists k \geq 1, \text{ s.t. } M \notin \mathcal{M}_k\right) \leq \frac{\delta}{3}.$$

We can now provide the improved regret bound for UCRL2B

Theorem 1. *There exists a numerical constant $\beta > 0$ such that for* any *communicating MDP, with probability at least $1-\delta$, it holds that for all initial state distributions $\mu_1 \in \Delta_S$ and for all time horizons $T > 1$*

$$\begin{aligned} \Delta(\text{UCRL2B}, T) \leq & \beta \cdot r_{\max} \sqrt{D\left(\sum_{s,a}\Gamma(s,a)\right) T \ln\left(\frac{T}{\delta}\right) \ln(T)} \\ & + \beta \cdot r_{\max} D^2 S^2 A \ln\left(\frac{T}{\delta}\right) \ln(T) \end{aligned} \tag{5}$$

where $\Gamma(s,a) := \|p(\cdot|s,a)\|_0$.

We now report the standard regret decomposition (e.g., Fruit et al., 2018)

$$
\begin{aligned}
R(T, \text{UCRL2B}) &\leq \sum_{k=1}^{k_T} \sum_{s} \nu_k(s) \left(g^\star_{M^\star} - \sum_{a} \pi_k(s,a) r(s,a) \right) + 2 r_{\max} \sqrt{T \ln\left(\frac{5T}{\delta}\right)} \\
&= \sum_{k=1}^{k_T} \Delta_k + 2 r_{\max} \sqrt{T \ln\left(\frac{5T}{\delta}\right)}
\end{aligned}
$$

where $k_T = \sup\{k \geq 1 : t \geq t_k\}$ and we further decompose Δ_k as

$$
\Delta_k \leq \Delta_k^p + \Delta_k^r + \frac{3\varepsilon_k}{2} \sum_{s \in \mathcal{S}} \nu_k(s)
$$

with

$$
\begin{aligned}
\Delta_k^p = \alpha & \underbrace{\sum_{s,a,s'} \nu_k(s) \pi_k(s,a) \Big(p_k(s'|s,a) - p(s'|s,a) \Big) h_k(s')}_{:= \Delta_k^{p1}} \\
& + \underbrace{\alpha \sum_{s} \nu_k(s) \left(\sum_{a,s'} \pi_k(s,a) p(s'|s,a) h_k(s') - h_k(s) \right)}_{:= \Delta_k^{p2}}
\end{aligned}
\tag{6}
$$

where $\alpha \in]0,1]$ is the coefficient of the *aperiodicity transformation* applied to extended MDP $\mathcal{M}_k$ (in most cases, this coefficient can be taken equal to 1 but we include it for the sake of generality). We also consider the general case where the optimistic policy π_k can be *stochastic* (in most cases this is not necessary).
Finally we define the event $E^C = \left\{ \exists T > 0, \exists k > 0, \ s.t. \ M^\star \notin \mathcal{M}_k \right\}$. It is easy to show that the probability of this event is small:

$$
\mathbb{P}(E^C) \leq \frac{\delta}{3}
$$

1 Improved regret analysis for UCRL2B

We will now prove Thm. 1. In order to improve the dependency of the regret bound in D (i.e., replace D by $\sqrt{D}$), we refine our analysis with three key improvements:

1. We leverage on *Freedman's inequality* (Freedman, 1975) instead of Azuma's inequality to bound the MDS. We recall this inequality in Prop. 2 below.
2. We use a *tighter bound* than Hölder's inequality to upper-bound the sum $\sum_{k=1}^{k_T} \Delta_k^{p3}$.
3. We shift the optimistic bias h_{k_t} by a different constant *at every time step* $t \geq 1$ rather than only at every episode $k \geq 1$. More precisely, the optimistic bias is shifted by a different constant for every episode $k \geq 1$ and for every visited state $s \in \mathcal{S}$.

To the best of our knowledge, Thm. 1 and its proof are new although it is largely inspired by what is often referred to as *"variance reduction methods"* in the literature (Munos and Moore, 1999; Lattimore and Hutter, 2012, 2014; Azar et al., 2017; Kakade et al., 2018). Similar techniques are used by (Azar et al., 2017) to achieve a similar bound but in the *finite horizon setting*. This approach is also related to (Talebi and Maillard, 2018) and Maillard et al. (2014) (in the latter, the variance is called the distribution-norm instead of the variance).

Proposition 2 (Freedman's inequality)**.** *Let $(X_n, \mathcal{F}_n)_{n\in\mathbb{N}}$ be an MDS such that $|X_n| \leq a$ a.s. for all $n \in \mathbb{N}$. Then for all $\delta \in]0,1[$,*

$$\mathbb{P}\left(\forall n \geq 1, \quad \left|\sum_{i=1}^{n} X_i\right| \leq 2\sqrt{\left(\sum_{i=1}^{n} \mathbb{V}\left(X_i \middle| \mathcal{F}_{i-1}\right)\right) \cdot \ln\left(\frac{4n}{\delta}\right)} + 4a\ln\left(\frac{4n}{\delta}\right)\right) \geq 1-\delta$$

For any *vector* $u \in \mathbb{R}^S$, we slightly abuse notation and write $u^2 := u \circ u$ the *Hadamard product* of u with itself. For any probability distribution p over states $\mathcal{S}$ and any vector $u \in \mathbb{R}^S$ we define $\mathbb{V}_p(u) := p^\intercal u^2 - (p^\intercal u)^2 = \mathbb{E}_{X\sim p}[u(X)^2] - \left(\mathbb{E}_{X\sim p}[u(X)]\right)^2$ the *"variance"* of u with respect to p. For the sake of clarity we introduce new notations for the transition probabilities: $p_k(s'|s) := \sum_{a\in\mathcal{A}_s} \pi_k(s,a)p_k(s'|s,a)$, $\overline{p}_k(s'|s) := \sum_{a\in\mathcal{A}_s} \pi_k(s,a)p(s'|s,a)$ and $\widehat{p}_k(s'|s) := \sum_{a\in\mathcal{A}_s} \pi_k(s,a)\widehat{p}_k(s'|s,a)$, for every $s, s' \in \mathcal{S}$ and every $k \geq 1$.
We start with a new bound relating Δ_k^{p1}:

Lemma 3. *Under event E, with probability at least $1 - \frac{\delta}{6}$:*

$$\forall T \geq 1, \quad \sum_{k=1}^{k_T} \Delta_k^{p1} \leq \sum_{k=1}^{k_T} \Delta_k^{p3} + 4r_{\max}D\ln\left(\frac{24T}{\delta}\right)$$
$$+ 2\sqrt{S\ln\left(\frac{24T}{\delta}\right)}\left(\sqrt{\sum_{t=1}^{T} \mathbb{V}_{p_{k_t}(\cdot|s_t)}\left(\alpha h_{k_t}\right)} + \sqrt{\sum_{t=1}^{T} \mathbb{V}_{\overline{p}_{k_t}(\cdot|s_t)}\left(\alpha h_{k_t}\right)}\right) \tag{7}$$

Proof. We use a martingale argument and Prop. 2. □

We *refine* the upper-bound of Δ_k^{p3} derived by Jaksch et al. (2010). Instead of bounding the scalar product $(p_k(\cdot|s,a) - p(\cdot|s,a))^\intercal w_k$ by $\|p_k(\cdot|s,a) - p(\cdot|s,a)\|_1^\intercal \|w_k\|_\infty$ using Hölder's inequality, we bound it by $\sum_{s'} |p_k(s'|s,a) - p(s'|s,a)| \cdot |w_k(s')|$ using the triangle inequality. Since $\sum_{a,s'} p_k(s'|s,a) = \sum_{a,s'} p(s'|s,a) = 1$ we can shift h_k by an arbitrary scalar $\lambda_k^s \in \mathbb{R}$ for all $k \geq 1$ and all $s \in \mathcal{S}$, i.e., $w_k^s := h_k + \lambda_k^s e$. Unlike in UCRL2, we choose a *state-dependent* shift, namely $\lambda_k^s := -\sum_{a,s'} \widehat{p}_k(s'|s,a)\pi_k(s,a)h_k(s') = -\widehat{p}_k(\cdot|s)^\intercal h_k$. It is easy to see that $sp(w_k^s) = sp(h_k)$ and $\|w_k^s\|_\infty \leq sp(h_k)$ implying that under event E, $\|w_k^s\|_\infty \leq (r_{\max}D)/\alpha$.
Using the triangle inequality and the fact that $p_k(s,a) \in B_p^k(s,a)$ by construction and $p(s,a) \in B_p^k(s,a)$ under event E:

$$\left|p_k(s'|s,a) - p(s'|s,a)\right| \leq \left|p_k(s'|s,a) - \widehat{p}_k(s'|s,a)\right| + \left|\widehat{p}_k(s'|s,a) - p(s'|s,a)\right| \leq 2\beta_{p,k}^{sas'}$$

As a result we can write:

$$\begin{aligned}
\Delta_k^{p3} &\leq \alpha \sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s,a)\left|p_k(s'|s,a) - p(s'|s,a)\right| \cdot \left|w_k^s(s')\right| \\
&\leq 2\alpha \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \sum_{s'} \beta_{p,k}^{sas'} \cdot \left|w_k^s(s')\right| \\
&= 4\alpha \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a)\left[\sqrt{\frac{\ln(6SAT/\delta)}{N_k^+(s,a)}} \sum_{s'\in\mathcal{S}} \sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))w_k^s(s')^2}\right. \\
&\qquad\qquad \left. + \frac{3\ln(6SAT/\delta)}{N_k^+(s,a)} \sum_{s'} \underbrace{\left|w_k^{sa}(s')\right|}_{\leq (r_{\max}D)/\alpha}\right]
\end{aligned}$$

We denote by $V_k(s,a) := \alpha^2 \sum_{s'} \widehat{p}_k(s'|s,a) w_k^s(s')^2$. We can prove the following inequality:

Lemma 4. *It holds almost surely that for all $k \geq 1$ and for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:*

$$\alpha \sum_{s' \in \mathcal{S}} \sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a)) w_k^s(s')^2} \leq \sqrt{V_k(s,a) \cdot (\Gamma(s,a) - 1)} \tag{8}$$

Proof. Define $\mathcal{S}_k(s,a) = \{s' \in \mathcal{S} \ : \ \widehat{p}_k(s'|s,a) > 0\}$. Then, using Cauchy-Schartz inequality we have

$$\begin{aligned}
\sum_{s' \in \mathcal{S}} \sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a)) w_k(s')^2} &= \sum_{s' \in \mathcal{S}_k(s,a)} \sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a)) w_k(s')^2} \\
&\leq \sqrt{\left(\sum_{s' \in \mathcal{S}_k(s,a)} 1 - \widehat{p}_k(s'|s,a) \right) \cdot \left(\sum_{s' \in \mathcal{S}_k(s,a)} \widehat{p}_k(s'|s,a) w_k(s')^2 \right)} \\
&= \sqrt{\Big(\Gamma_k(s,a) - 1 \Big) \cdot \left(\sum_{s' \in \mathcal{S}} \widehat{p}_k(s'|s,a) w_k(s')^2 \right)} \leq \sqrt{\Gamma(s,a) \sum_{s' \in \mathcal{S}} \widehat{p}_k(s'|s,a) w_k(s')^2}
\end{aligned}$$

By definition, for all $s' \in \mathcal{S}$, $w_k(s') = h_k(s') - \mathbb{E}_{X \sim \widehat{p}_k(\cdot|s,a)}[h_k(X)]$ and so

$$\sum_{s' \in \mathcal{S}} \widehat{p}_k(s'|s,a) w_k(s')^2 = \mathbb{V}_{\widehat{p}_k(\cdot|s,a)}\left(h_k\right)$$

$\square$

As a consequence of Lem. 4,

$$\begin{aligned}
\sum_{k=1}^{k_T} \Delta_k^{p3} &\leq 4 \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \left[\sqrt{V_k(s,a) \frac{\Gamma(s,a)}{N_k^+(s,a)} \ln\left(\frac{6SAT}{\delta} \right)} + \frac{3 r_{\max} DS}{N_k^+(s,a)} \ln\left(\frac{6SAT}{\delta} \right) \right] \\
&= 4 \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \left[\sqrt{V_k(s_t,a_t) \frac{\Gamma(s_t,a_t)}{N_k^+(s_t,a_t)} \ln\left(\frac{6SAT}{\delta} \right)} + \frac{3 r_{\max} DS}{N_k^+(s_t,a_t)} \ln\left(\frac{6SAT}{\delta} \right) \right]
\end{aligned}$$

Applying Cauchy-Schwartz gives

$$\begin{aligned}
\sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \sqrt{V_k(s_t,a_t)} \sqrt{\frac{\Gamma(s_t,a_t)}{N_k^+(s_t,a_t)}} &\leq \sqrt{\sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \frac{\Gamma(s_t,a_t)}{N_k^+(s_t,a_t)} \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} V_k(s_t,a_t)} \\
&= \sqrt{\sum_{k=1}^{k_T} \sum_{s,a} \frac{\Gamma(s,a) \nu_k(s,a)}{N_k^+(s,a)} \sum_{t=1}^{T} V_{k_t}(s_t,a_t)}
\end{aligned}$$

Using Lem. 8, Jensen's inequality and the fact that $N_{k_T+1}^+(s,a) \leq T$, we can bound the first sum

$$\begin{aligned}
\sum_{s,a} \sum_{k=1}^{k_T} \frac{\Gamma(s,a) \nu_k(s,a)}{N_k^+(s,a)} &\leq 2 \sum_{s,a} \Gamma(s,a) \left(1 + \ln\left(N_{k_T+1}^+(s,a)\right)\right) \\
&\leq 2 \left(1 + \ln\left(\frac{\sum_{s,a} \Gamma(s,a) N_{k_T+1}^+(s,a)}{\sum_{s,a} \Gamma(s,a)} \right) \right) \sum_{s,a} \Gamma(s,a) \\
&\leq 2(1 + \ln(T)) \sum_{s,a} \Gamma(s,a)
\end{aligned}$$

To bound the second sum $\sum_{t=1}^{T} V_{k_t}(s_t, a_t)$, we rely on the following Lemma:

Lemma 5. *Under event E, with probability at least $1 - \frac{\delta}{6}$:*

$$\forall T \geq 1, \quad \sum_{t=1}^{T} V_{k_t}(s_t, a_t) \leq \sum_{t=1}^{T} \mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t)}\left(\alpha h_{k_t}\right) + (r_{\max} D)^2 \sqrt{2T \ln\left(\frac{T}{\delta}\right)} \tag{9}$$

Proof. We notice that for all $k \geq 1$ and $s \in \mathcal{S}$, $\sum_a \pi_k(s,a) V_k(s,a) = \mathbb{V}_{\widehat{p}_k(\cdot|s)}\left(\alpha h_k\right)$. The concentration inequality then follows from a martingale argument and Azuma's inequality. $\square$

From Lem. 5 it follows that

$$\sum_{k=1}^{k_T} \Delta_k^{p3} \leq 4\sqrt{2\Big(1 + \ln(T)\Big) \ln\left(\frac{6SAT}{\delta}\right) \left(\sum_{s,a} \Gamma(s,a)\right) \left((r_{\max} D)^2 \sqrt{2T \ln\left(\frac{T}{\delta}\right)} + \sum_{t=1}^{T} \mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t)}\left(\alpha h_{k_t}\right)\right)}$$
$$+ 24 r_{\max} D S^2 A \ln\left(\frac{6SAT}{\delta}\right)(1 + \ln(T)) \tag{10}$$

It now remains to bound $\sum_{k=1}^{k_T} \Delta_k^{p2}$. As shown by (Jaksch et al., 2010; Fruit et al., 2018) using telescopic sum argument: $\sum_{k=1}^{k_T} \Delta_k^{p2} \leq \sum_{k=1}^{k_T} \Delta_k^{p4} + (r_{\max} D) k_T$ where

$$\Delta_k^{p4} = \alpha \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_{a,s'} \pi_k(s_t,a) p(s'|s,a) w_k(s') - w_k(s_{t+1})\right)$$

We bound $\sum_{k=1}^{k_T} \Delta_k^{p4}$ using Freedman's inequality instead of Azuma's.

Lemma 6. *Under event E, with probability at least $1 - \frac{\delta}{6}$:*

$$\forall T \geq 1, \quad \sum_{k=1}^{k_T} \Delta_k^{p4} \leq 2\sqrt{\left(\sum_{t=1}^{T} \mathbb{V}_{\overline{p}_{k_t}(\cdot|s_t)}\left(\alpha h_k\right)\right) \cdot \ln\left(\frac{24T}{\delta}\right)} + 4 r_{\max} D \ln\left(\frac{24T}{\delta}\right) \tag{11}$$

Proof. We use a martingale argument and Prop. 2 (see App. B.1 for further details). $\square$

1.1 Bounding the sum of variances

The main terms appearing respectively in (7), (10) and (11) all have the form of a *sum of variances over time* $\sum_{t=1}^{T} \mathbb{V}_{p_t}\left(\alpha h_{k_t}\right)$ with p_t a distribution over states (respectively $p_{k_t}(\cdot|s_t)$, $\overline{p}_{k_t}(\cdot|s_t)$ and $\widehat{p}_{k_t}(\cdot|s_t)$)[1], and h_{k_t} the optimistic bias of episode k_t. A first *naïve* upper bound of this sum can be derived using Popoviciu's inequality that we recall in Prop. 7.

Proposition 7 (Popoviciu's inequality on variances)**.** *Let M and m be upper and lower bounds on the values of a random variable X i.e., $\mathbb{P}m \leq X \leq M = 1$. Then $\mathbb{V}(X) \leq \frac{1}{4}(M - m)$.*

Using Popoviciu's inequality and under event E,

$$\mathbb{V}_{p_t}\left(\alpha h_{k_t}\right) \leq sp\left(\alpha h_k\right)^2/4 = \alpha^2 sp\left(h_k\right)^2/4 \leq (r_{\max} D)^2/4$$

and so $\sum_{t=1}^{T} \mathbb{V}_{p_t}\left(\alpha h_{k_t}\right) \leq (r_{\max} D)^2 T/4$. Unfortunately, this would result in a regret bound scaling as $\widetilde{\mathcal{O}}(r_{\max} D\sqrt{T})$ (ignoring all other terms like S, A, logarithmic terms, etc.) which is

[1] Recall that $\overline{p}_k(\cdot|s) := \sum_a \pi_k(s,a) p(s'|s,a)$.

not better than the classical bound of UCRL2. In this section, we show that the cumulative sum of variances only scales as $\widetilde{\mathcal{O}}(r_{\max}^2 DT + (r_{\max} D)^2\sqrt{T})$ resulting in a regret bound of order $\widetilde{\mathcal{O}}\left(r_{\max}\sqrt{DT} + r_{\max} DT^{1/4}\right)$ (ignoring all other terms).

We start by analyzing the variance term $\mathbb{V}_{\widehat{p}_k(\cdot|s_t)}(\alpha h_k)$. We will proceed similarly with the other variance terms $\mathbb{V}_{p_k(\cdot|s_t)}(\alpha h_k)$ and $\mathbb{V}_{\overline{p}_k(\cdot|s_t)}(\alpha h_k)$. We do the following decomposition:

$$
\begin{aligned}
\mathbb{V}_{\widehat{p}_k(\cdot|s_t)}(\alpha h_k) &= \alpha^2\left(\widehat{p}_k(\cdot|s_t)^\intercal h_k^2 - \left(\widehat{p}_k(\cdot|s_t)^\intercal h_k\right)^2\right)\\
&= \alpha^2\Big(\underbrace{\left(\widehat{p}_k(\cdot|s_t) - \overline{p}_k(\cdot|s_t)\right)^\intercal h_k^2}_{①} + \underbrace{\overline{p}_k(\cdot|s_t)^\intercal h_k^2 - h_k^2(s_{t+1})}_{②} + \underbrace{h_k^2(s_{t+1}) - \left(\widehat{p}_k(\cdot|s_t)^\intercal h_k\right)^2}_{③}\Big)
\end{aligned}
$$

Notice that for any r.v. X and any scalar $a\in\mathbb{R}$, $\mathbb{V}(X+a) = \mathbb{V}(X)$. Thus, the term $\mathbb{V}_{\widehat{p}_k(\cdot|s_t)}(\alpha h_k)$ remains unchanged when h_k is shifted by an arbitrary constant vector i.e., when h_k is replaced by $w_k := h_k + \lambda_k e$. As in UCRL2, we minimize the ℓ_∞-norm of w_k by choosing $\lambda_k = -\frac{1}{2}\left(\max_{s\in\mathcal{S}}\{h_k(s)\} + \min_{s\in\mathcal{S}}\{h_k(s)\}\right)$. We recall that under event E, $\|w_k\|_\infty \leq (r_{\max}D)/(2\alpha)$ and so $\|w_k^2\|_\infty \leq (r_{\max}D)^2/(4\alpha^2)$.

① The *first term* $\alpha^2\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\left(\widehat{p}_k(\cdot|s_t) - \overline{p}_k(\cdot|s_t)\right)^\intercal w_k^2$ is similar to $\sum_{k=1}^{k_T}\Delta_k^{p1}$ except that αw_k is replaced by $\alpha^2 w_k^2$ and $p_k(\cdot|s_t)$ is replaced by $\widehat{p}_k(\cdot|s_t)$. In the regret proof of UCRL2 we need to decompose $p_k(\cdot|s_t) - \overline{p}_k(\cdot|s_t)$ into the sum of $p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)$ and $\widehat{p}_k(\cdot|s_t) - \overline{p}_k(\cdot|s_t)$. Here we no longer need this decomposition and we can use the same derivation with $sp\left(\alpha^2 w_k^2\right) \leq (r_{\max}D)^2/4$ instead of $(r_{\max}D)/2$. Therefore, with probability at least $1-\frac{\delta}{6}$ (and under event E):

$$
\begin{aligned}
\alpha^2\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\left(\widehat{p}_k(\cdot|s_t) - \overline{p}_k(\cdot|s_t)\right)^\intercal w_k^2 \leq \frac{3}{2}(r_{\max}D)^2\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)T\ln\left(\frac{6SAT}{\delta}\right)} + (r_{\max}D)^2\sqrt{T\ln\left(\frac{5T}{\delta}\right)}\\
+3(r_{\max}D)^2 S^2 A\ln\left(\frac{6SAT}{\delta}\right)(1+\ln(T))
\end{aligned}
$$

② The *second term* $\alpha^2\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\overline{p}_k(\cdot|s_t)^\intercal w_k^2 - w_k^2(s_{t+1})$ is identical to $\sum_{k=1}^{k_T}\Delta_k^{p4}$ except that αw_k is replaced by $\alpha^2 w_k^2$. With probability at least $1-\frac{\delta}{6}$ (and under event E):

$$
\alpha^2\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\overline{p}_k(\cdot|s_t)^\intercal w_k^2 - w_k^2(s_{t+1}) \leq \frac{(r_{\max}D)^2}{2}\sqrt{T\ln\left(\frac{5T}{\delta}\right)}
$$

③ The *last term* $\alpha^2\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1} w_k^2(s_{t+1}) - \left(\widehat{p}_k(\cdot|s_t)^\intercal w_k\right)^2$ is the *dominant* one and requires more work. Unlike the first two terms, it scales *linearly* with T (instead of $\widetilde{\mathcal{O}}(\sqrt{T})$). We first notice that $\widehat{p}_k(\cdot|s_t)^\intercal w_k = w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)$. Using the fact that $(a+b)^2 = a^2 + b(2a+b)$ with $a = w_k(s_t)$ and $b = \widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)$ (and therefore $2a+b = w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k$) we obtain:

$$
\left(\widehat{p}_k(\cdot|s_t)^\intercal w_k\right)^2 = w_k^2(s_t) + \left(\widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)\right)\cdot\left(w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k\right)
$$

and so applying the *reverse triangle inequality*:

$$
\left(\widehat{p}_k(\cdot|s_t)^\intercal w_k\right)^2 \geq w_k^2(s_t) - \left|\widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)\right|\cdot\left|w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k\right| \tag{12}
$$

For all $k\geq 1$ and $s\in\mathcal{S}$, we define $r_k(s) := \sum_a \pi_k(s,a) r_k(s,a)$. Using the (near-)optimality equation we can write:

$$
\left|g_k - r_k(s_t) + \alpha\big(w_k(s_t) - p_k(\cdot|s_t)^\intercal w_k\big)\right| = \left|g_k - r_k(s_t) + \alpha\big(h_k(s_t) - p_k(\cdot|s_t)^\intercal h_k\big)\right| \leq \varepsilon_k
$$

Moreover, $\varepsilon_k = \frac{r_{\max}}{t_k} \leq r_{\max}$. As a result, since $\alpha > 0$:

$$\begin{aligned}
&\alpha\big|\widehat{p}_k(\cdot|s_t)^\intercal w_k - w_k(s_t)\big| \\
&\quad= \big|g_k - r_k(s_t) + \alpha\big(w_k(s_t) - p_k(\cdot|s_t)^\intercal w_k\big) - g_k + r_k(s_t) + \alpha\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\big| \\
&\quad\leq \underbrace{\big|g_k - r_k(s_t) + \alpha\big(w_k(s_t) - p_k(\cdot|s_t)^\intercal w_k\big)\big|}_{\leq r_{\max}} + \underbrace{\big|r_k(s_t) - g_k\big|}_{\leq r_{\max}} + \alpha\left|\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\right| \\
&\quad\leq 2r_{\max} + \alpha\left|\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\right|
\end{aligned}$$

It is also immediate to see that $|w_k(s_t) + \widehat{p}_k(\cdot|s_t)^\intercal w_k| \leq 2\|w_k\|_\infty \leq (r_{\max}D)/\alpha$. Plugging these inequalities into (12) and adding $w_k^2(s_{t+1})$ we obtain:

$$\begin{aligned}
\alpha^2\left(w_k^2(s_{t+1}) - \left(\widehat{p}_k(\cdot|s_t)^\intercal w_k\right)^2\right) \leq &\left(2r_{\max} + \alpha\left|\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\right|\right)(r_{\max}D) \\
&+ \alpha^2\left(w_k^2(s_{t+1}) - w_k^2(s_t)\right)
\end{aligned} \tag{13}$$

It is easy to bound the telescopic sum

$$\alpha^2 \sum_{t=t_k}^{t_{k+1}-1} w_k^2(s_{t+1}) - w_k^2(s_t) = \alpha^2\left(w_k^2(s_{t_{k+1}}) - w_k^2(s_{t_k})\right) \leq \alpha^2 w_k^2(s_{t_{k+1}}) \leq (r_{\max}D)^2/4 \tag{14}$$

Finally, the sum $\alpha\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\left|\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\right|$ can be bounded in the exact same way as $\sum_{k=1}^{k_T} \Delta_k^{p1}$. With probability at least $1 - \frac{\delta}{6}$:

$$\begin{aligned}
\alpha\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\left|\left(p_k(\cdot|s_t) - \widehat{p}_k(\cdot|s_t)\right)^\intercal w_k\right| \leq &3r_{\max}D\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)T\ln\left(\frac{6SAT}{\delta}\right)} + 4r_{\max}D\sqrt{T\ln\left(\frac{5T}{\delta}\right)} \\
&+ 6r_{\max}DS^2A\ln\left(\frac{6SAT}{\delta}\right)(1 + \ln(T))
\end{aligned} \tag{15}$$

After gathering (14) and (15) into (13)) we conclude that with probability at least $1 - \frac{\delta}{6}$ (and under event E):

$$\alpha^2\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1} w_k^2(s_{t+1}) - \left(\widehat{p}_k(\cdot|s_t)^\intercal w_k\right)^2 \leq \underbrace{2r_{\max}^2DT}_{\text{main term}} + \frac{k_T(r_{\max}D)^2}{4} + \widetilde{\mathcal{O}}\left((r_{\max}D)^2\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)T}\right)$$

In conclusion, there exists an *absolute* numerical constant $\beta > 0$ (i.e., independent of the MDP instance) such that with probability at least $1 - \frac{5\delta}{6}$:

$$\sum_{t=1}^{T}\mathbb{V}_{\widehat{p}_{k_t}(\cdot|s_t)}(\alpha h_{k_t}) \leq \beta\cdot\left(r_{\max}^2DT + (r_{\max}D)^2\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)T\ln\left(\frac{T}{\delta}\right)} + (r_{\max}D)^2S^2A\ln\left(\frac{T}{\delta}\right)\ln(T)\right)$$

We can prove the same bound (possibly with a different multiplicative constant β) for $\sum_{t=1}^{T}\mathbb{V}_{\overline{p}_{k_t}(\cdot|s_t)}(\alpha h_{k_t})$ and $\sum_{t=1}^{T}\mathbb{V}_{p_{k_t}(\cdot|s_t)}(\alpha h_{k_t})$ using the same derivation.

1.2 Completing the regret bound of Thm. 1

After plugging the bound derived for the sum of variances in the previous section (Sec. 1.1) into (7), (10) and (11), we notice that (7) and (11) can be upper-bounded by (10) *up to a multiplicative numerical constant* ans so it is enough to restrict attention to (10). The dominant term that we obtain is (ignoring numerical constants):

$$r_{\max}\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)\ln\left(\frac{T}{\delta}\right)\ln(T)\left(DT+D^2\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)T\ln\left(\frac{T}{\delta}\right)}+D^2S^2A\ln\left(\frac{T}{\delta}\right)\ln(T)\right)}$$

Using the fact that $\sqrt{\sum_i a_i}\leq\sum_i\sqrt{a_i}$ for any $a_i\geq 0$, we can bound the above square-root term by three simpler terms:

(1) A $\sqrt{T}$-term (dominant): $r_{\max}\sqrt{D\left(\sum_{s,a}\Gamma(s,a)\right)T\ln\left(\frac{T}{\delta}\right)\ln(T)}$

(2) A $T^{1/4}$-term: $r_{\max}D\left(\sum_{s,a}\Gamma(s,a)\right)^{3/4}T^{1/4}\left(\ln\left(\frac{T}{\delta}\right)\right)^{3/4}\sqrt{\ln(T)}$

(3) A logarithmic term: $r_{\max}D\sqrt{S^2A\left(\sum_{s,a}\Gamma(s,a)\right)}\ln\left(\frac{T}{\delta}\right)\ln(T)\leq r_{\max}DS^2A\ln\left(\frac{T}{\delta}\right)\ln(T)$

When $T\geq D^2\left(\sum_{s,a}\Gamma(s,a)\right)\ln\left(\frac{T}{\delta}\right)$, we notice that the $T^{1/4}$-term (2) is actually upper-bounded by the $\sqrt{T}$-term (1), while for $T\leq D^2\left(\sum_{s,a}\Gamma(s,a)\right)\ln\left(\frac{T}{\delta}\right)$ we can use the trivial upper-bound $r_{\max}T$ on the regret:

$$R(T,M^\star,\text{UCRL2B})\leq r_{\max}T\leq r_{\max}D^2\left(\sum_{s,a}\Gamma(s,a)\right)\ln\left(\frac{T}{\delta}\right)\leq r_{\max}D^2S^2A\ln\left(\frac{T}{\delta}\right)$$

To complete the regret bound of Thm. 1 we also need to take into consideration the *lower order terms* of (7), (10) and (11). It turns out that the only terms that are not already upper-bounded by (1), (2) and (3) (up to multiplicative numerical constants) sum as:

$$r_{\max}\sqrt{SAT\ln\left(\frac{T}{\delta}\right)}+r_{\max}SA\ln\left(\frac{T}{\delta}\right)\ln(T)+r_{\max}D^2S^2A\ln\left(\frac{T}{\delta}\right)\ln(T)$$

To conclude, we only need to *adjust* δ to obtain an event of probability at least $1-\delta$. This will *only* impact the multiplicative numerical constants of the above terms.

References

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2007). Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, pages 150–165, Berlin, Heidelberg. Springer Berlin Heidelberg.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19):1876–1902.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272, International Convention Centre, Sydney, Australia. PMLR.

Freedman, D. A. (1975). On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. *CoRR*, abs/1802.04020.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600.

Kakade, S., Wang, M., and Yang, L. F. (2018). Variance Reduction Methods for Sublinear Reinforcement Learning. *ArXiv e-prints*.

Lattimore, T. and Hutter, M. (2012). Pac bounds for discounted mdps. In *In Proc. 23rd International Conf. on Algorithmic Learning Theory (ALT'12), volume 7568 of LNAI*. Springer.

Lattimore, T. and Hutter, M. (2014). Near-optimal pac bounds for discounted mdps. *Theoretical Computer Science*, 558:125–143.

Lattimore, T. and Szepesvári, C. (2018). Bandit algorithms. Pre-publication version.

Maillard, O.-A., Mann, T. A., and Mannor, S. (2014). How hard is my mdp?" the distribution-norm to the rescue". In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, page 1835–1843. Curran Associates, Inc.

Munos, R. and Moore, A. (1999). Influence and variance of a markov chain: Application to adaptive discretization in optimal control. In *Proceedings: International Astronomical Union Transactions, v. 16B p*, pages 355–362.

Talebi, M. S. and Maillard, O. (2018). Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *ALT*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR.

A Additional Results

Lemma 8. *It holds almost surely that for all $k \geq 1$ and for all $(s,a) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:*

$$\sum_{k=1}^{k_T} \frac{\nu_k(s,a)}{\sqrt{N_k^+(s,a)}} \leq 3\sqrt{N_{k_T+1}(s,a)} \quad \textit{and} \quad \sum_{k=1}^{k_T} \frac{\nu_k(s,a)}{N_k^+(s,a)} \leq 2 + 2\ln\left(N_{k_T+1}^+(s,a)\right) \tag{16}$$

Proof. The proof follows from the rate of divergence of the series $\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \sim \sqrt{n}$ and $\sum_{i=1}^{n} \frac{1}{i} \sim \ln(n)$ respectively when $n \to +\infty$. $\square$

B MDS

For any $t \geq 0$, the σ-algebra induced by the past history of state-action pairs and rewards up to time t (included) is denoted $\mathcal{F}_t = \sigma(s_1, a_1, r_1, \dots, s_t, a_t, r_t, s_{t+1})$ where by convention $\mathcal{F}_0 = \sigma(\emptyset)$ and $\mathcal{F}_\infty := \cup_{t \geq 0} \mathcal{F}_t$. Trivially, for all $t \geq 0$, $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ and the filtration $(\mathcal{F}_t)_{t \geq 0}$ is denoted by $\mathbb{F}$. We recall that k_t is the integer-valued r.v. indexing the current episode at time t. It is immediate from the termination condition of episodes that for all $t \geq 1$, k_t is $\mathcal{F}_{t-1}$-measurable i.e., the past sequence $(s_1, a_1, r_1, \dots, s_{t-1}, a_{t-1}, r_{t-1}, s_t)$ fully determines the ongoing episode at time t. As a consequence, the stationary (randomized) policy π_{k_t} executed at time t is also $\mathcal{F}_{t-1}$-measurable.

B.1 Proof of Lemma 6

Let's define the stochastic process

$$X_t := \sum_{a,s'} \pi_{k_t}(s_t, a) p_{k_t}(s'|s_t, a) h_{k_t}(s') - \sum_{s'} p_{k_t}(s'|s_t, a_t) h_{k_t}(s')$$

Let's define $\lambda_t = -\sum_{a,s'} \pi_{k_t}(s_t, a) p_{k_t}(s'|s_t, a) h_{k_t}(s')$ and $w_t = h_{k_t} + \lambda_t e$. Since by definition $\sum_{s'} p_{k_t}(s'|s_t, a_t) = 1$, we have

$$X_t = -\sum_{s'} p_{k_t}(s'|s_t, a_t) w_t(s')$$

It is easy to verify that $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$ and so $(X_t, \mathcal{F}_t)_{t \geq 1}$ is an MDS. Moreover, $|X_t| \leq \|w_t\|_\infty \leq sp(h_{k_t}) \leq (r_{\max} D)$ and

$$\mathbb{V}\left(X_t \middle| \mathcal{F}_{t-1}\right) = \sum_a \pi_{k_t}(s_t, a) \left(\sum_{s'} p_{k_t}(s'|s_t, a) w_t(s') \right)^2$$

Proposition 9. *For any $n \geq 1$ and any n-tuple $(a_1, \dots, a_n) \in \mathbb{R}^n$, $\left(\sum_{i=1}^n a_i\right)^2 \leq n \left(\sum_{i=1}^n a_i^2\right)$.*

Proof. The statement is trivially true for $n = 1$. For $n = 2$ we have $(a_1 - a_2)^2 = a_1^2 + a_2^2 - 2a_1 a_2 \geq 0$ implying that $2a_1 a_2 \leq a_1^2 + a_2^2$. Therefore, $(a_1 + a_2)^2 = a_1^2 + a_2^2 + 2a_1 a_2 \leq 2(a_1^2 + a_2^2)$ and so the result holds. We prove the result for $n \geq 2$ by induction. Assumed that it is true for any $n \geq 2$. Then we have:

$$\begin{aligned}
\left(\sum_{i=1}^{n+1} a_i \right)^2 &= \underbrace{\left(\sum_{i=1}^{n} a_i \right)^2}_{\leq n\left(\sum_{i=1}^n a_i^2\right)} + a_{n+1}^2 + 2a_{n+1} \sum_{i=1}^n a_i \\
&\leq n \left(\sum_{i=1}^n a_i^2 \right) + a_{n+1}^2 + \sum_{i=1}^n \underbrace{2a_i a_{n+1}}_{\leq a_i^2 + a_{n+1}^2} \leq (n+1) \cdot \left(\sum_{i=1}^{n+1} a_i^2 \right)
\end{aligned}$$

where the first inequality follows from the induction hypothesis and the second inequality follows from the inequality for $n = 2$ that we proved. This concludes the proof. $\square$

For the sake of clarity we will now use the notation $p_k(s'|s) := \sum_{a \in \mathcal{A}_s} \pi_k(s, a) p_k(s'|s, a)$ for every

$s, s' \in \mathcal{S}$ and every $k \geq 1$. Using Prop. 9 we have that

$$\mathbb{V}\left(X_t \middle| \mathcal{F}_{t-1}\right) \leq S \sum_{a,s'} \pi_{k_t}(s_t, a) \underbrace{p_{k_t}(s'|s_t, a)^2}_{\leq p_{k_t}(s'|s_t,a)} w_{k_t}(s')^2$$
$$\leq S \sum_{a,s'} \pi_{k_t}(s_t, a) p_{k_t}(s'|s_t, a) w_{k_t}(s')^2 = S \cdot \mathbb{V}_{p_{k_t}(\cdot|s_t)}\left(h_{k_t}\right)$$

After applying Freedman's inequality (Prop. 2) to the MDS $(X_t, \mathcal{F}_t)_{t\geq 1}$ we obtain that with probability at least $1 - \frac{\delta}{6}$, for all $T \geq 1$:

$$\sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s) \pi_k(s,a) p_k(s'|s,a) h_k(s') \leq \sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s,a) p_k(s'|s,a) h_k(s') + 2(r_{\max} D) \ln\left(\frac{24T}{\delta}\right)$$
$$+ 2\sqrt{S \ln\left(\frac{24T}{\delta}\right) \sum_{t=1}^{T} \mathbb{V}_{p_{k_t}(\cdot|s_t)}\left(h_{k_t}\right)} \tag{17}$$

We can do exactly the same analysis with the stochastic process

$$X_t := \sum_{a,s'} \pi_{k_t}(s_t, a) p(s'|s_t, a) h_{k_t}(s') - \sum_{s'} p(s'|s_t, a_t) h_{k_t}(s')$$

i.e., with p instead of p_{k_t} and we obtain that with probability at least $1 - \frac{\delta}{6}$, for all $T \geq 1$:

$$-\sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s) \pi_k(s,a) p(s'|s,a) h_k(s') \leq -\sum_{k=1}^{k_T} \sum_{s,a,s'} \nu_k(s,a) p(s'|s,a) h_k(s') + 2(r_{\max} D) \ln\left(\frac{24T}{\delta}\right)$$
$$+ 2\sqrt{S \ln\left(\frac{24T}{\delta}\right) \sum_{t=1}^{T} \mathbb{V}_{\overline{p}_{k_t}(\cdot|s_t)}\left(h_{k_t}\right)} \tag{18}$$

with the notation $\overline{p}_k(s'|s) := \sum_{a \in \mathcal{A}_s} \pi_k(s,a) p(s'|s,a)$ for every $s, s' \in \mathcal{S}$ and $k \geq 1$.

B.2 Definition of The Confidence Intervalsd

Theorem 10. *The probability that there exists $k \geq 1$ s.t. the true MDP M does not belong to the extended MDP $\mathcal{M}_k$ defined by Eq. 3 and 4 is at most $\frac{\delta}{3}$, that is*

$$\mathbb{P}\left(\exists k \geq 1, \ \ s.t. \ M \notin \mathcal{M}_k\right) \leq \frac{\delta}{3}.$$

Proof. We want to bound the probability of event $E := \bigcup_{k=1}^{+\infty} \{M \notin \mathcal{M}_k\}$. As explained by Lattimore and Szepesvári (2018, Section 4.4), when (s,a) is visited for the n-th times, the reward that we observe is the n-th element of an infinite sequence of i.i.d. r.v. lying in $[0, r_{\max}]$ with expected value $r(s,a)$. Similarly, the next state that we observe is the n-th element of an infinite sequence of i.i.d. r.v. lying in $\mathcal{S}$ with probability density function (pdf) $p(\cdot|s,a)$. In UCRL2, we defined the sample means $\widehat{p}_k$ and $\widehat{r}_k$, and the confidence intervals B_p^k and B_r^k (Eq. 3 and 4) as depending on k. Actually, this quantities depends only on the first $N_k(s,a)$ elements of the infinite i.i.d. sequences that we just mentioned. For the rest of the proof, we will therefore slightly change our notations and denote by $\widehat{p}_n(s'|s,a)$, $\widehat{r}_n(s,a)$, $B_p^n(s'|s,a)$ and $B_r^n(s,a)$ the sample means and confidence intervals after the first n visits in (s,a). Thus, the r.v. that we denoted by $\widehat{p}_k$ in UCRL2 actually corresponds to $\widehat{p}_{N_k(s,a)}$ with our new notation (and similarly for $\widehat{r}_k$, B_p^k and B_r^k). This change of notation will make the proof easier.

$M \notin \mathcal{M}_k$ means that there exists $k \geq 1$ s.t. either $p(s'|s,a) \notin B_p^{N_k(s,a)}(s,a,s')$ or $r(s,a) \notin B_r^{N_k(s,a)}(s,a)$ for at least one $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. This means that there exists at least one value $n \geq 0$ s.t. either $p(s'|s,a) \notin B_p^n(s,a,s')$ or $r(s,a) \notin B_r^n(s,a)$. As a consequence we have the following inclusion

$$E \subseteq \bigcup_{s,a} \bigcup_{n=0}^{+\infty} \{r(s,a) \notin B_r^n(s,a)\} \cup \bigcup_{s'} \{p(s'|s,a) \notin B_p^n(s,a,s')\} \tag{19}$$

Using Boole's inequality we thus have:

$$\mathbb{P}(E) \leq \sum_{s,a} \sum_{n=0}^{+\infty} \left(\mathbb{P}\left(r(s,a) \notin B_r^n(s,a)\right) + \sum_{s'} \mathbb{P}\left(p(s'|s,a) \notin B_p^n(s,a,s')\right) \right) \tag{20}$$

Let's fix a 3-tuple $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and define for all $n \geq 0$

$$\epsilon_{p,n}^{sas'} := \widehat{\sigma}_{p,n}(s'|s,a) \sqrt{\frac{2\ln\left(30S^2A(n^+)^2/\delta\right)}{n^+}} + \frac{3\ln\left(30S^2A(n^+)^2/\delta\right)}{n^+} \tag{21}$$

$$\epsilon_{r,n}^{sa} := \widehat{\sigma}_{r,n}(s,a) \sqrt{\frac{2\ln\left(30SA(n^+)^2/\delta\right)}{n^+}} + \frac{3r_{\max}\ln\left(30SA(n^+)^2/\delta\right)}{n^+} \tag{22}$$

where $\widehat{\sigma}_{p,n}(s'|s,a)$ and $\widehat{\sigma}_{r,n}(s,a)$ denote the population variances obtained with the first n samples. It is immediate to verify that $\epsilon_{p,n}^{sas'} \leq \beta_{p,n}^{sas'}$ and $\epsilon_{r,n}^{sa} \leq \beta_{r,n}^{sa}$ a.s. (see Eq. 1 and 2 with $N_k(s,a)$ replaced by n). Using the empirical Bernstein inequality (Audibert et al., 2009, Thm. 1) we have that for all $n \geq 1$:

$$\mathbb{P}\left(|p(s'|s,a) - \widehat{p}_n(s'|s,a)| \geq \beta_{p,n}^{sas'}\right) \leq \mathbb{P}\left(|p(s'|s,a) - \widehat{p}_n(s'|s,a)| \geq \epsilon_{p,n}^{sas'}\right) \leq \frac{\delta}{10n^2S^2A} \tag{23}$$

$$\mathbb{P}\left(|r(s,a) - \widehat{r}_n(s,a)| \geq \beta_{r,n}^{sa}\right) \leq \mathbb{P}\left(|r(s,a) - \widehat{r}_n(s,a)| \geq \epsilon_{r,n}^{sa}\right) \leq \frac{\delta}{10n^2SA} \tag{24}$$

Note that when $n = 0$ (i.e., when there hasn't been any observation of (s,a)), $\epsilon_{p,0}^{sas'} \geq 1$ and $\epsilon_{r,0}^{sa} \geq r_{\max}$ so $\mathbb{P}\left(|p(s'|s,a) - \widehat{p}_0(s'|s,a)| \geq \epsilon_{p,0}^{sas'}\right) = \mathbb{P}\left(|r(s,a) - \widehat{r}_0(s,a)| \geq \epsilon_{r,0}^{sa}\right) = 0$ by definition. Since in addition (also by definition)

$$B_p^n(s,a,s') \subseteq \left[\widehat{p}_n(s'|s,a) - \beta_{p,n}^{sas'}, \widehat{p}_n(s'|s,a) + \beta_{p,n}^{sas'}\right] \text{ (see Eq. 3)}$$

and

$$B_r^n(s,a) \subseteq \left[\widehat{r}_n(s,a) - \beta_{r,n}^{sa}, \widehat{r}_k(s,a) + \beta_{r,n}^{sa}\right] \text{ (see Eq. 4)}$$

we conclude that for all $n \geq 1$

$$\mathbb{P}\left(p(s'|s,a) \notin B_p^n(s,a,s')\right) \leq \frac{\delta}{10n^2S^2A} \quad \text{and} \quad \mathbb{P}\left(r(s,a) \notin B_r^n(s,a)\right) \leq \frac{\delta}{10n^2SA}$$

and these probabilities are equal to 0 if $n = 0$. Plugging these inequalities into Eq. (20) we obtain:

$$\mathbb{P}\left(\exists T \geq 1, \exists k \geq 1 \text{ s.t.} M \notin \mathcal{M}_k\right) \leq \sum_{s,a} \left(0 + \sum_{n=1}^{+\infty} \left(\frac{\delta}{10n^2SA} + \sum_{s'} \frac{\delta}{10n^2S^2A}\right)\right) = \frac{2\pi^2\delta}{60} \leq \frac{\delta}{3}$$

which concludes the proof. $\square$