The appendix is organized as follows: We first provide some insights on extended value iterations useful in our construction of the regret.Then, the detailed proof of theorem 4.1 is given with bounds on the five terms in our decomposition of the regret. A final appendix provides technical lemmas about MDPs in $\mathcal{M}$.

A Proof of Theorem 4.1

A.1 Extended value iteration

For each episode k, we use the extended value iteration algorithm described in [11] to compute $\tilde{\pi}_k$ and $\tilde{M} \in \mathcal{M}_k$, an optimistic policy and MDP. The values we iteratively get are defined in the following way:

$$
\begin{aligned}
&u_0^{(k)}(s) = 0 \\
&u_{i+1}^{(k)}(s) = \max_{a \in \mathcal{A}} \left\{ \tilde{r}(s,a) + \max_{p(\cdot) \in \mathcal{P}(s,a)} \left\{ \sum_{s \in S} p(s') u_i^{(k)}(s') \right\} \right\},
\end{aligned}
\tag{16}
$$

where $\tilde{r}$ is the maximal reward from (4) and $\mathcal{P}(s,a)$ is the set of probabilities from (5).

Now, from [11, Theorem 7], we obtain the following lemma on the iterations of extended value iteration.

Lemma A.1. *For episode k, denote by i the last step of extended value iteration, stopped when:*

$$
\max_s \{u_{i+1}^{(k)}(s) - u_i^{(k)}(s)\} - \min_s \{u_{i+1}^{(k)}(s) - u_i^{(k)}(s)\} < \frac{r_{\max}}{\sqrt{t_k}}. \tag{17}
$$

The optimistic MDP $\tilde{M}_k$ and the optimistic policy $\tilde{\pi}_k$ that we choose are so that the gain is $\frac{1}{\sqrt{t_k}}-$close to the optimal gain:

$$
\tilde{\rho}_k := \min_s \rho(\tilde{M}_k, \tilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s'} \rho(M', \pi, s') - \frac{r_{\max}}{\sqrt{t_k}}. \tag{18}
$$

Moreover from [16, Theorem 8.5.6]:

$$
\left| u_{i+1}^{(k)}(s) - u_i^{(k)}(s) - \tilde{\rho}_k \right| \leq \frac{r_{\max}}{\sqrt{t_k}}, \tag{19}
$$

and when the optimal policy yields an irreducible and aperiodic Markov chain, we have that $\tilde{\rho}_k = \rho(\tilde{M}_k, \tilde{\pi}_k, s)$ for any s, so that we can define the bias:

$$
\tilde{h}_k(s_0) = \mathbb{E}_{s_0} \left[\sum_{t=0}^{\infty} (\tilde{r}(s_t, a_t) - \tilde{\rho}_k) \right]. \tag{20}
$$

By choosing iteration i large enough, from [16, Equation 8.2.5], we can also ensure that:

$$
\left| u_i^{(k)}(s) - (i-1)\tilde{\rho}_k - \tilde{h}_k(s) \right| < \frac{r_{\max}}{2\sqrt{t_k}}, \tag{21}
$$

so that we can define the following difference

$$
d_k(s) := \left| u_i^{(k)}(s) - \min_s u_i^{(k)}(s) - \left(\tilde{h}_k(s) - \min_s \tilde{h}_k(s) \right) \right| < \frac{r_{\max}}{\sqrt{t_k}}. \tag{22}
$$

A.2 Regret when M is out of the confidence bound

Let us compute $\mathbb{E}[\text{Reg}]$, the expected regret. We will mainly follow the approach in [11, Section 4], with a few tweaks. We start by splitting the regret into a sum over episodes and states.

We remind that $\overline{r}(s,a)$ is the overall mean reward and $N_T(s,a)$ the total count of visits. We also define $R_k(s) := \sum_a \nu_k(s,a)(\rho^* - \overline{r}(s,a))$ the regret at episode k induced by state s, with $\nu_k(s,a)$ the number of visit of (s,a) during episode k.

Let $R_{\text{in}} := \sum_s \sum_{k=1}^m R_k(s)\mathbb{1}_{M\in\mathcal{M}_k}$ and $R_{\text{out}} := \sum_s \sum_{k=1}^m R_k(s)\mathbb{1}_{M\notin\mathcal{M}_k}$. We therefore have the split:

$$\mathbb{E}\left[Reg\right] \leq \mathbb{E}\left[R_{\text{in}}\right] + \mathbb{E}\left[R_{\text{out}}\right]. \tag{23}$$

Now, let $\nu_k(s) = \sum_a \nu_k(s,a)$ and denote by $\mathcal{M}(t)$ the set of MDPs $\mathcal{M}_k$ such that $t_k \leq t < t_{k+1}$. For the terms out of the confidence sets, we have:

$$\begin{aligned}
R_{\text{out}} &\leq \sum_s \sum_{k=1}^m \nu_k(s)\mathbb{1}_{M\notin\mathcal{M}_k} \\
&\leq \sum_s \sum_{k=1}^m N_{t_k}(s)\mathbb{1}_{M\notin\mathcal{M}_k} \text{ using the stopping criterion} \\
&= \sum_{t=1}^T \sum_s \sum_{k=1}^m \mathbb{1}_{t_k=t} N_t(s)\mathbb{1}_{M\notin\mathcal{M}(t)} \leq \sum_{t=1}^T \sum_s N_t(s)\mathbb{1}_{M\notin\mathcal{M}(t)} \\
&= \sum_{t=1}^T \mathbb{1}_{M\notin\mathcal{M}(t)} \sum_s N_t(s) \leq \sum_{t=1}^T t\mathbb{1}_{M\notin\mathcal{M}(t)}.
\end{aligned}$$

Taking the expectations:

$$\begin{aligned}
\mathbb{E}\left[R_{\text{out}}\right] &\leq r_{\max} \sum_{t=1}^T t\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} \\
&\leq r_{\max} \sum_{t=1}^T \frac{tS}{2t^3} \leq r_{\max} \sum_{t=1}^T \frac{S}{2t^2} \text{ by Lemma B.1} \\
&\leq r_{\max} S.
\end{aligned} \tag{24}$$

Thus, we have dealt with the cases where the MDP M did not belong to any confidence set, for some episodes. We now need to deal with the rest.

A.3 Regret terms when M is in the confidence bound

We now assume that $M \in \mathcal{M}_k$ and deal with the terms in the confidence bound, so that we can omit the repetitions of the indicator functions. For each episode k, let $R_{\text{in},k} := \sum_s R_k$.

We defined $\tilde{\pi}_k$ the optimistic policy computed at episode k, now define $\tilde{P}_k := (\tilde{p}_k(s'|s,\tilde{\pi}_k(s)))$ the transition matrix of that policy on the optimistic MDP $\tilde{M}_k$. Define also $\mathbf{v}_k := (\nu_k(s,\tilde{\pi}_k)$ the row vector of visit counts during episode k. Following the same steps as in [11], we get the inequality on the regret of episode k, assuming $M \in \mathcal{M}_k$, using Lemma A.1:

$$\begin{aligned}
R_{\text{in},k} &= \sum_{s,a} \nu_k(s,a)(\rho^* - \overline{r}(s,a)) \\
&\leq \sum_{s,a} \nu_k(s,a)(\tilde{\rho}_k - \overline{r}(s,a)) + r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}} \\
&= \sum_{s,a} \nu_k(s,a)(\tilde{\rho}_k - \tilde{r}_k(s,a)) + \sum_{s,a} \nu_k(s,a)(\tilde{r}_k - \overline{r}(s,a)) + r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}}.
\end{aligned}$$

Then with (19) and using the definition of the iterated values from EVI, we have for a given state s and $a_s := \tilde{\pi}_k(s)$:

$$\left|(\tilde{\rho}_k - \tilde{r}_k(s,a_s)) - \left(\sum_{s'} \tilde{p}_k(s'|s,a_s)u_i^{(k)}(s') - u_i^{(k)}(s)\right)\right| \leq \frac{r_{\max}}{\sqrt{t_k}},$$

so that:

$$R_{\text{in},k} \leq \mathbf{v}_k\left(\tilde{\mathbf{P}}_k - \mathbf{I}\right)\mathbf{u}_i + \sum_{s,a} \nu_k(s,a)(\tilde{r}_k - \overline{r}(s,a)) + 2r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}}.$$

Remember that for any state s: $|d_k(s)| \leq \frac{r_{\max}}{\sqrt{t_k}}$, where $\tilde{\mathbf{h}}_k$ is the bias of the average optimal policy for the optimist MDP, and:

$$d_k(s) := \left(u_i^{(k)}(s) - \min_x u_i^{(k)}(x)\right) - \left(\tilde{\mathbf{h}}_k(s) - \min_x \tilde{\mathbf{h}}_k(x)\right).$$

Notice that the unit vector is in the kernel of $\left(\tilde{\mathbf{P}}_k - \mathbf{I}\right)$. Therefore, in the first term, we can replace $\mathbf{u}_i$ by any translation of it. We get:

$$\mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{u}_i = \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k + \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{d}_k.$$

so that:

$$\begin{aligned} R_{\text{in}} \leq & \underbrace{\sum_k \sum_{s,a} \nu_k(s,a)(\tilde{r}_k - \overline{r}(s,a))}_{R_{\text{rewards}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k}_{R_{\text{bias}}} \\ & + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{d}_k + 2r_{\max} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{t_k}}}_{R_{\text{EVI}}}. \end{aligned}$$

Then, using the assumption on empirical rewards (4), as $M \in \mathcal{M}_k$, and noticing that $N_{t_k} \leq t_k$:

$$R_{\text{rewards}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{25}$$

For the term $\mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{d}_k$, which does not appear in the analysis of [11], we obtain

$$\begin{aligned} \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{I}\right) \mathbf{d}_k & \leq \sum_s \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left\|\tilde{p}_k\left(\cdot|s, \tilde{\pi}_k(s)\right) - \mathbb{1}_s\right\|_1 \cdot \sup_{s'} |d_k(s')| \\ & \leq 2r_{\max} \sum_s \frac{\nu_k\left(s, \tilde{\pi}_k(s)\right)}{\sqrt{t_k}} \leq 2r_{\max} \sum_{s,a} \frac{\nu_k\left(s,a\right)}{\sqrt{t_k}} \\ & \leq 2r_{\max} \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}, \end{aligned}$$

where in the last inequality we used that $\max\{1, N_{t_k}(s,a)\} \leq t_k \leq T$. Thus, for $T \geq \frac{e^2}{2AT}$ the regret term coming from the consequences and approximations of EVI satisfies

$$R_{\text{EVI}} \leq r_{\max} 2\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{26}$$

Now, by defining P_k the transition matrix of the optimistic policy $\tilde{\pi}_k$ in the true MDP M, we have the following decomposition of the middle term:

$$\underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \mathbf{h}^*}_{R_{\text{trans}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \left(\tilde{\mathbf{h}}_k - \mathbf{h}^*\right)}_{R_{\text{diff}}} + \underbrace{\sum_k \mathbf{v}_k \left(\mathbf{P}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k}_{R_{\text{ep}}}$$

Overall:

$$\begin{aligned} R_{\text{in}} \leq & \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \mathbf{h}^*}_{R_{\text{trans}}} + \underbrace{\sum_k \mathbf{v}_k \left(\tilde{\mathbf{P}}_k - \mathbf{P}_k\right) \left(\tilde{\mathbf{h}}_k - \mathbf{h}^*\right)}_{R_{\text{diff}}} + \underbrace{\sum_k \mathbf{v}_k \left(\mathbf{P}_k - \mathbf{I}\right) \tilde{\mathbf{h}}_k}_{R_{\text{ep}}} \\ & \underbrace{+ r_{\max} 4\sqrt{2\log(2AT)} \sum_k \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}}_{R_{\text{EVI}} + R_{\text{rewards}}}. \end{aligned}$$

A.3.1 Bound on R_{trans}

Let us deal with the first term R_{trans}. To bound this term, we will use our knowledge of the optimal bias $\mathbf{h}^*$ and the control of the difference of the transition matrices, and for the second term we will control the difference of the biases.

Notice that for a fixed state $1 \leq s \leq S-1$:

$$\sum_{s'} p\left(s'|s, \tilde{\pi}_k(s)\right) h^*(s') = \sum_{s'} p\left(s'|s, \tilde{\pi}_k(s)\right) \left(h^*(s') - h^*(s)\right) + h^*(s).$$

The same is true for $\tilde{p}_k$, and knowing the MDP is a birth and death process:

$$\begin{aligned}
R_{\text{trans}} &= \sum_k \sum_s \sum_{s'} \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left(\tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right) - p\left(s'|s, \tilde{\pi}_k(s)\right)\right) \cdot h^*(s') \\
&= \sum_k \sum_s \sum_{s'} \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left(\tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right) - p\left(s'|s, \tilde{\pi}_k(s)\right)\right) \cdot \left(h^*(s') - h^*(s)\right) \\
&\leq \sum_k \sum_s \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left\|\tilde{p}_k\left(\cdot|s, \tilde{\pi}_k(s)\right) - p\left(\cdot|s, \tilde{\pi}_k(s)\right)\right\|_1 \sup_{s'} \partial h^*(s) \\
&\leq 4\sqrt{2\log\left(2AT\right)} \sum_k \sum_{s,a} \frac{\Delta(s)\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}},
\end{aligned}$$

where in the last inequality, we used the knowledge on the bounded variations of the optimal bias from Lemma 3.2, and that the optimistic MDP has transitions close to the true transitions.

A.3.2 Bound on R_{diff}

We now deal with the term involving the difference of bias, R_{diff}. For each episode k with policy π_k, denote by x_k the state such that the confidence bounds are at their worst and denote by $a_k := \pi_k(x_k)$ the corresponding action used at this state, so that $N_{t_k}(x_k, a_k)$ is minimal. We therefore have that $\sqrt{\frac{\log(2At_k)}{\max\{1, N_{t_k}(x_k,a_k)\}}}$ is maximal for episode k. The true MDP being within the confidence bounds, with a triangle inequality:

$$\|P_k - P^*\|_\infty \leq 4\sqrt{\frac{2\log\left(2At_k\right)}{\max\{1, N_{t_k}(x_k, a_k)\}}},$$

and

$$\|r_k - r^*\|_\infty \leq 2r_{\max}\sqrt{\frac{2\log\left(2At_k\right)}{\max\{1, N_{t_k}(x_k, a_k)\}}}.$$

Then using Lemma C.4, and noticing that to bound the biases $\tilde{\mathbf{h}}_k, \mathbf{h}^*$ and the quantity $\|\sum_{t=1}^T \tilde{P}_k^t \tilde{r}_k\|$ is bounded by the same diameter D, using the same argument as in [11] (Equation (11)), and noticing that $D \geq 1$:

$$\|\tilde{\mathbf{h}}_k - \mathbf{h}^*\|_\infty \leq 12 T_{hit} r_{\max} D \sqrt{\frac{2\log\left(2At_k\right)}{\max\{1, N_{t_k}(x_k, a_k)\}}}. \tag{27}$$

Hence,

$$\begin{aligned}
R_{\text{diff}} &\leq \sum_s \sum_{s'} \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left(\tilde{p}_k\left(s'|s, \tilde{\pi}_k(s)\right) - p\left(s'|s, \tilde{\pi}_k(s)\right)\right) \cdot \left(\tilde{h}_k(s') - h^*(s')\right) \\
&\leq \sum_s \nu_k\left(s, \tilde{\pi}_k(s)\right) \cdot \left\|\tilde{p}_k\left(\cdot|s, \tilde{\pi}_k(s)\right) - p\left(\cdot|s, \tilde{\pi}_k(s)\right)\right\|_1 \|\tilde{\mathbf{h}}_k - \mathbf{h}^*\|_\infty \\
&\leq 48 D^2 r_{\max} \log\left(2AT\right) \Sigma,
\end{aligned}$$

where in the last inequality we have used (27) and that by definition of D

$$T_{hit} := \inf_{s' \in \mathcal{S}} \sup_{s \in \mathcal{S}} \mathbb{E}_s\, \tau_{s'}^{\pi^*} \leq \mathbb{E}_{S-1}\, \tau_0^{\pi^0} \leq D,$$

and we called

$$\Sigma := \sum_{s,a} \sum_{k} \sum_{t=t_k}^{t_{k+1}-1} \frac{\mathbb{1}_{\{s_t, a_t = s,a\}}}{\sqrt{\max\{1, N_{t_k}(s,a)\}}\sqrt{\max\{1, N_{t_k}(x_k, a_k)\}}}.$$

By the choice of x_k, $N_{t_k}(x_k, a_k) \leq N_{t_k}(s,a)$ for any state-action pair (s,a), so that we can rewrite, with $I_k := t_{k+1} - t_k$ the length of episode k:

$$\Sigma \leq \sum_{s,a} \sum_{k} \sum_{t=t_k}^{t_{k+1}-1} \frac{\mathbb{1}_{\{s_t, a_t = s,a\}}}{\max\{1, N_{t_k}(x_k, a_k)\}} \leq \sum_{k} \frac{I_k}{\max\{1, N_{t_k}(x_k, a_k)\}}.$$

Now define $Q_{\max} := \left(\frac{10D}{m^{\max}(S-1)}\right)^2 \log\left(\left(\frac{10D}{m^{\max}(S-1)}\right)^4\right)$, and $I(T) := \max\left\{Q_{\max}, T^{1/4}\right\}$. We split the sum depending on whether the episodes are shorter than $I(T)$ or not, and call $K_{\leq I}$ the number of such episodes. This yields:

$$\Sigma \leq K_{\leq I} I(T) + \sum_{k, I_k > I(T)} \frac{I_k}{\max\{1, N_{t_k}(x_k, a_k)\}}.$$

Using the stopping criterion for episodes:

$$\Sigma \leq K_{\leq I} I(T) + \sum_{k, I_k > I(T)} \frac{I_k}{\max\{1, \nu_k(x_k, a_k)\}}.$$

Now denote by $\mathcal{E}$ the event:

$$\mathcal{E} = \left\{\forall k \text{ s.t } I_k > I(T),\ \frac{1}{\max\{1, \nu(x_k, a_k)\}} \leq \frac{2}{m^{\max}(S-1) I_k}\right\}.$$

By splitting the sum, using the above event, we get:

$$\begin{aligned}\Sigma &\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} \sum_{k, I_k > I(T)} \frac{2}{m^{\max}(S-1)} + \mathbb{1}_{\bar{\mathcal{E}}} \sum_{k, I_k > I(T)} I_k \\ &\leq K_{\leq I} I(T) + \mathbb{1}_{\mathcal{E}} \left(K_T - K_{\leq I}\right) \frac{2}{m^{\max}(S-1)} + \mathbb{1}_{\bar{\mathcal{E}}} T.\end{aligned}$$

We use Corollary C.6 to get $\mathbb{P}\left(\bar{\mathcal{E}}\right) \leq \frac{1}{4T}$, so that when taking the expectation:

$$\mathbb{E}\left[\Sigma\right] \leq \mathbb{E}\left[K_{\leq I}\right] I(T) + \mathbb{E}\left[(K_T - K_{\leq I})\right] \frac{2}{m^{\max}(S-1)} + \frac{1}{4}$$

Now using Lemma B.3, $SA \geq 4$, $I(T) \geq \frac{2}{m^{\max}(S-1)}$ and that $\frac{1}{\log 2} + \frac{1}{4} \leq 2$:

$$\mathbb{E}\left[\Sigma\right] \leq \mathbb{E}\left[K_T\right] I(T) + \frac{1}{4} \leq 2SA \log(2AT) I(T).$$

We therefore have that:

$$\mathbb{E}\left[R_{\text{diff}}\right] \leq 96 r_{\max} SAD^2 I(T) \log^2\left(2AT\right). \tag{28}$$

A.3.3 Bound on the main terms: Exploiting the stochastic ordering

In Section 4.3 we have shown that:

$$R_{\text{trans}} \leq 4\sqrt{2\log\left(2AT\right)} \sum_{s,a} \frac{\Delta(s) \nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}. \tag{29}$$

To control this term as well as R_{EVI} (26) and R_{rewards} (25), we need to control the terms in the sum in a way that does not make the parameters D or S appear, as this will be one of the main contributing terms. To do so, we need to sum over the episodes and take the expectation, so that with Lemma B.4, we get:

$$\begin{aligned}\mathbb{E}\left[\sum_{s,a}\sum_{k}\frac{\nu_k(s,a)}{\sqrt{\max\{1,N_{t_k}(s,a)\}}}\right] &\leq 3\mathbb{E}\left[\sum_{s,a}\sqrt{N_T(s,a)}\right]\\ &\leq 3\sum_{s}\sqrt{\mathbb{E}\left[N_T(s)\right]A}\text{ by Jensen's inequality.}\end{aligned}$$

We will use the following lemma to carry on the computations:

Lemma A.2. *Let m^{π^0} be the stationary measure of the Markov chain under policy π^0, such that for every state s: $\pi^0(s)=0$. Let $f:\mathcal{S}\to\mathbb{R}^+$ be a non-negative non-decreasing function on the state space. Then for any state $s\in\mathcal{S}$,*

$$\mathbb{E}\left[\sum_{s'\geq s} f(s')N_t(s')\right] \leq t\sum_{s'\geq s} f(s')m^{\pi^0}(s') \tag{30}$$

Proof. Let $s\in\mathcal{S}$. For any state s', define $N_t^{m^{\pi^0},\pi^0}(s')$ the number of visits when the starting state follows the initial distribution m^{π^0}, and the MDP always executes the policy π^0 at every timestep instead of the policy determined by the algorithm UCRL2. Notice already that for any state s':

$$\mathbb{E}\left[N_t^{m^{\pi^0},\pi^0}(s')\right] = tm^{\pi^0}(s')$$

On the other hand, for $x\in\mathcal{S}$, we have the stochastic ordering:

$$\sum_{s'\geq x} N_t(s') \leq_{st} \sum_{s'\geq x} N_t^{m^{\pi^0},\pi^0}(s'),$$

so that for any non-negative non-decreasing function f, with the convention $f(-1)=0$:

$$\begin{cases}(f(x)-f(x-1))\sum_{s'\geq x} N_t(s') \leq_{st} (f(x)-f(x-1))\sum_{s'\geq x} N_t^{m^{\pi^0},\pi^0}(s')\\ f(s-1)\sum_{s'\geq s} N_t(s') \leq_{st} f(s-1)\sum_{s'\geq s} N_t^{m^{\pi^0},\pi^0}(s'),\end{cases} \tag{31}$$

and then summing the equation above for $s\leq x\leq S-1$ and switching the sums yields:

$$\sum_{s'\geq s} N_t(s')\sum_{x=s}^{s'}[f(x)-f(x-1)] \leq_{st} \sum_{s'\geq s} N_t^{m^{\pi^0},\pi^0}(s')\sum_{x=s}^{s'}[f(x)-f(x-1)],$$

which simplifies to:

$$\sum_{s'\geq s} N_t(s')[f(s')-f(s-1)] \leq_{st} \sum_{s'\geq s} N_t^{m^{\pi^0},\pi^0}(s')[f(s')-f(s-1)].$$

Now summing with the second equation in (31), we get the following equation:

$$\sum_{s'\geq s} N_t(s')f(s') \leq_{st} \sum_{s'\geq s} N_t^{m^{\pi^0},\pi^0}(s')f(s').$$

Taking the expectation in this last inequality finishes the proof. □

Now, we can conclude our bound on R_{trans}. Since

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}(\Delta(s)+r_{\max})\frac{\nu_k(s,a)}{\sqrt{\max\{1,N_{t_k}(s,a)\}}}\right] \leq 3\sqrt{A}\sum_{s\geq 0}(\Delta(s)+r_{\max})\sqrt{\mathbb{E}\left[N_T(s)\right]}, \tag{32}$$

let f be a non-negative non-decreasing function over the state space, such that $F := \sum_{s\geq 0} f(s)^{-1}$ exists. Then by concavity:

$$\begin{aligned}
\sum_{s\geq 0}(\Delta(s)+r_{\max})\sqrt{\mathbb{E}\left[N_T(s)\right]} &= F\sum_{s\geq 0}\frac{1}{Ff(s)}\sqrt{f(s)^2(\Delta(s)+r_{\max})^2\mathbb{E}\left[N_T(s)\right]} \\
&\leq F\sqrt{\sum_{s\geq 0}\frac{f(s)^2(\Delta(s)+r_{\max})^2\mathbb{E}\left[N_T(s)\right]}{Ff(s)}} \text{ by concavity} \\
&= \sqrt{F\sum_{s\geq 0}f(s)(\Delta(s)+r_{\max})^2\mathbb{E}\left[N_T(s)\right]} \\
&\leq \sqrt{TF\sum_{s\geq 0}f(s)(\Delta(s)+r_{\max})^2 m^{\pi^0}(s)} \text{ using Lemma A.2,}
\end{aligned}$$

so that overall, (32) becomes:

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}\frac{(\Delta(s)+r_{\max})\nu_k(s,a)}{\sqrt{\max\{1,N_{t_k}(s,a)\}}}\right] \leq 3\sqrt{ATF}\sqrt{\sum_{s\geq 0}f(s)(\Delta(s)+r_{\max})^2 m^{\pi^0}(s)}. \tag{33}$$

This is the term mainly contributing to the regret.

A.3.4 Bound on the main terms: Introducing E_2

Now, using Lemma B.5 which gives the stationary distribution of m^0, we can compute the expectation under m^0 of a well-chosen function f:

Lemma A.3. *Let us choose the increasing function* $f : s \mapsto \frac{\max\{1, s(s-1)\}}{(\Delta(s)+r_{\max})^2}$. *Then* $F \leq 3(C+r_{\max})^2$ *and* $\sum_{s\geq 0}(\Delta(s)+r_{\max})^2 f(s) m^{\pi^0}(s) = \mathbb{E}_{m^{\pi^0}}\left[(\Delta+r_{\max})^2\cdot f\right] \leq \left(1+\frac{\lambda^2}{\mu^2}\right)$, *so that:*

$$E_2 := F\mathbb{E}_{m^{\pi^0}}\left[(\Delta+r_{\max})^2\cdot f\right] \leq 3(C+r_{\max})^2\left(1+\frac{\lambda^2}{\mu^2}\right).$$

Proof. For F, we obtain:

$$F \leq (C+r_{\max})^2\left(2+\sum_{s=2}^{S-1}\frac{1}{s(s-1)}\right) = (C+r_{\max})^2\left(2+\sum_{s=2}^{S-1}\left(\frac{1}{s-1}-\frac{1}{s}\right)\right) \leq 3(C+r_{\max})^2$$

Using the following computations:

$$\begin{aligned}
\sum_{s=2}^{S-1}s(s-1)\binom{S-1}{s}\left(\frac{\lambda}{(S-1)\mu}\right)^s &= (S-2)(S-1)\sum_{s=2}^{S}\binom{S-3}{s-2}\left(\frac{\lambda}{(S-1)\mu}\right)^s \\
&= (S-2)(S-1)\left(\frac{\lambda}{(S-1)\mu}\right)^2\sum_{s=0}^{S-3}\binom{S-3}{s}\left(\frac{\lambda}{(S-1)\mu}\right)^s \\
&= (S-2)(S-1)\left(\frac{\lambda}{(S-1)\mu}\right)^2\left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-3} \\
&\leq \left(\frac{\lambda}{\mu}\right)^2\left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-3},
\end{aligned}$$

and using that $1+\frac{\lambda}{\mu} \leq \left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-1}$, we get:

$$\left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-1}\mathbb{E}_{m^{\pi^0}}\left[(\Delta+r_{\max})^2\cdot f\right] \leq \left(1+\frac{\lambda^2}{\mu^2}\right)\left(1+\frac{\lambda}{(S-1)\mu}\right)^{S-1},$$

so that finally

$$\mathbb{E}_{m^{\pi^0}}\left[(\Delta + r_{\max})^2 \cdot f\right] \leq \left(1 + \frac{\lambda^2}{\mu^2}\right),$$

which concludes the proof. □

Finally (33) becomes:

$$\mathbb{E}\left[\sum_{s,a}\sum_{k}\frac{(\Delta(s) + r_{\max})\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}}\right] \leq 3\sqrt{E_2 AT}, \tag{34}$$

and thus:

$$\mathbb{E}\left[R_{\text{trans}} + R_{\text{rewards}} + R_{\text{EVI}}\right] \leq 12\sqrt{2E_2 AT\log(2AT)}. \tag{35}$$

In particular:

$$\mathbb{E}\left[R_{\text{trans}} + R_{\text{rewards}} + R_{\text{EVI}}\right] \leq 30(C + r_{\max})\sqrt{\left(1 + \frac{\lambda^2}{\mu^2}\right) AT\log(2AT)}. \tag{36}$$

A.3.5 Bound on R_{ep}

It remains to deal with the following regret term:

$$R_{\text{ep}} = \sum_k \mathbf{v}_k\left(\mathbf{P}_k - \mathbf{I}\right)\tilde{\mathbf{h}}_k.$$

We will follow the core of the proof from [11]. Define $X_t := \left(p(\cdot|s_t, a_t) - \mathbf{e}_{s_t}\right)\tilde{\mathbf{h}}_{k(t)}\mathbb{1}_{M\in\mathcal{M}_{k(t)}}$, where $k(t)$ is the episode containing step t and $\mathbf{e}_i$ the vector with i-th coordinate 1 and 0 for the other coordinates.

$$\begin{aligned}\mathbf{v}_k\left(\mathbf{P}_k - \mathbf{I}\right)\tilde{\mathbf{h}}_{\mathbf{k}} &= \sum_{t=t_k}^{t_{k+1}-1} X_t + \tilde{\mathbf{h}}_k(s_{t_{k+1}}) - \tilde{\mathbf{h}}_k(s_{t_k})\\ &\leq \sum_{t=t_k}^{t_{k+1}-1} X_t + Dr_{\max}.\end{aligned}$$

By summing over the episodes we get:

$$R_{\text{ep}} \leq \sum_{t=1}^{T} X_t + K_T D r_{\max}.$$

Notice that $\mathbb{E}\left[X_t|s_1, a_1, \ldots, s_t, a_t\right] = 0$, so that when taking the expectations, only the term in the number of episodes remains.

On the other side, using Lemma B.3, we get when taking the expectation:

$$\mathbb{E}\left[R_{\text{ep}}\right] \leq SA\log_2\left(\frac{8T}{SA}\right) \cdot Dr_{\max}.$$

Assuming $SA \geq 4$, and using $\log(2) \geq \frac{1}{2}$:

$$\mathbb{E}\left[R_{\text{ep}}\right] \leq 2r_{\max}SAD\log(2AT). \tag{37}$$

We can now gather the expected regret terms when the true MDP is within the confidence bounds. Using (28), (35) and (37):

$$\mathbb{E}\left[R_{\text{in}}\right] \leq 96r_{\max}SAD^2 I(T)\log^2(2AT) + 12\sqrt{2E_2 AT\log(2AT)} + 2r_{\max}SAD\log(2AT),$$

which gives with (23) and (24), assuming that $T \geq S^2$:

$$\mathbb{E}\left[Reg\right] \leq 97 r_{\max} SAD^2 I(T) \log^2 \left(2AT\right) + 12\sqrt{2E_2 AT \log\left(2AT\right)}.$$

which finally gives:

$$\mathbb{E}\left[Reg\right] \leq 97 r_{\max} SAD^2 I(T) \log^2 \left(2AT\right) + 19\sqrt{E_2 AT \log\left(2AT\right)}.$$

B Technical Lemmas

B.1 Probability of the confidence bounds

This first lemma is from [11, Lemma 17] and adapted to our confidence bounds.

Lemma B.1. *For $t > 1$, the probability that the MDP M is not within the set of plausible MDPs $\mathcal{M}_t$ is bounded by:*

$$\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} < \frac{S}{2t^3}.$$

Proof. Fix a state action pair (s, a), and n the number of visits of this pair before time t. Recall that $\hat{p}$ and $\hat{r}$ are the empirical transition probabilities and rewards from the n observations. Knowing that from each pair, there are at most 3 transitions, a Weissman's inequality gives for any $\varepsilon_p > 0$:

$$\mathbb{P}\left\{\|\hat{p}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geq \varepsilon_p\right\} \leq 6 \exp\left(-\frac{n\varepsilon_p^2}{2}\right).$$

So that for the choice of $\varepsilon_p = \sqrt{\frac{2}{n}\log\left(16At^4\right)} \leq \sqrt{\frac{8}{n}\log\left(2At\right)}$, we get:

$$\mathbb{P}\left\{\|\hat{p}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geq \sqrt{\frac{8}{n}\log\left(2At\right)}\right\} \leq \frac{3}{8At^4}.$$

We can do similar computations for the confidence on rewards, with a Hoeffding inequality:

$$\mathbb{P}\left\{|\hat{r}(s,a) - r(s,a)| \geq \varepsilon_r\right\} \leq 2 \exp\left(-\frac{2n\varepsilon_r^2}{r_{\max}^2}\right),$$

and choosing $\varepsilon_r = r_{\max}\sqrt{\frac{1}{2n}\log\left(16At^4\right)} \leq r_{\max}\sqrt{\frac{2}{n}\log\left(2At\right)}$, so that:

$$\mathbb{P}\left\{|\hat{r}(s,a) - r(s,a)| \geq r_{\max}\sqrt{\frac{2}{n}\log\left(2At\right)}\right\} \leq \frac{1}{8At^4}.$$

Now with a union bound for all values of $n \in \{0, 1, \cdots, t-1\}$, we get:

$$\mathbb{P}\left\{\|\hat{p}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geq \sqrt{\frac{8\log\left(2At\right)}{\max\{1, N_t(s,a)\}}}\right\} \leq \frac{3}{8At^3},$$

and

$$\mathbb{P}\left\{|\hat{r}(s,a) - r(s,a)| \geq r_{\max}\sqrt{\frac{2\log\left(2At\right)}{\max\{1, N_t(s,a)\}}}\right\} \leq \frac{1}{8At^3},$$

and finally, when summing over all state-action pairs, $\mathbb{P}\left\{M \notin \mathcal{M}(t)\right\} < \frac{S}{2t^3}$. □

B.2 Number of visits for an MDP in $\mathcal{M}$

This lemma is needed in the proof of Lemma C.5.

Lemma B.2 (Azuma-Hoeffding inequality)**.** *Let $X_1, X_2, \dots$ be a martingale difference sequence with $|X_i| \le RD$ for all i and some $R > 0$. Then for all $\varepsilon > 0$ and $n \in \mathbb{N}$:*

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \ge \varepsilon\right\} \le \exp\left(-\frac{\varepsilon^2}{2nDR}\right).$$

The two following lemmas are proved in [11, Appendix C.2 and Appendix C.3] respectively. Bounding the number of episodes is notably useful to obtain equation (28).

Lemma B.3. *Denote by K_t the number of episodes up to time t, and let $t > SA$. It is bounded by:*

$$K_t \le SA \log_2\left(\frac{8t}{SA}\right).$$

The following lemma is used to simplify regret terms, notably (29).

Lemma B.4. *For any fixed state action pair (s, a) and time T, we have:*

$$\sum_{k=1} \frac{\nu_k(s,a)}{\sqrt{\max\{1, N_{t_k}(s,a)\}}} \le 3\sqrt{N_{T+1}(s,a)},$$

B.3 Diameter and Span of MDPs in $\mathcal{M}$

For completeness, and to support the discussion in Section 4.2, the section details the behavior of the diameter and the span of MDPs in $\mathcal{M}$, as functions of S.

Under policy π^0, it is possible to get en explicit expression for the stationary distribution of the states.

Lemma B.5. *Under the stationary policy π^0, the stationary measure $m^{\pi^0}(s)$ of the MDP is given by:*

$$m^{\pi^0}(s) = \frac{\binom{S-1}{s}\left(\frac{\lambda}{(S-1)\mu}\right)^{s}}{\left(1 + \frac{\lambda}{(S-1)\mu}\right)^{(S-1)}}.$$

This lemma is shown in the proof of [1, Lemma 3.3].

First, it should be clear that under any policy π, the diameter of the MDP under π is extremely large because the probability to move from state s to state $s+1$ is smaller and smaller as s grows. Actually, this is also true for the local diameter, more precisely the expected time to go from s to $s+1$ grows extremely fast with s.

This discussion is formalized in the following result.

Lemma B.6. *For any $M \in \mathcal{M}$ and any policy π, the diameter D^π as well as the local diameter $D^\pi(s-1, s)$ grow as S^{S-2}.*

Proof. Under policy π, the following sequence of inequalities follows from the stochastic comparison with π^0 and monotonicity under π^0.

$$D^\pi \ge \tau^\pi(0, S-1) \ge \tau^{\pi^0}(0, S-1) \ge \tau^{\pi^0}(S-2, S-1),$$

where $\tau^\pi(x, y)$ is the expected time to go from x to y under policy π.

Now, starting from $S-2$, the Markov chain moves to $S-1$ with probability $p := \lambda/(U(S-1))$ and the time to reach $S-1$ is equal to 1 or moves to $S-2$ or $S-3$ with probability $1-p$. Therefore, $\tau^{\pi^0}(S-2, S-1)$ is bounded by $1-p$ times the return time to $S-2$ in the chain truncated at $S-2$, bounded in turn by the inverse of the stationary measure of state $S-2$ in this chain . Using Lemma B.5,

$$\tau^{\pi^0}(S-2, S-1) \geq (1-p)\left(\frac{(S-2)\mu}{\lambda}\right)^{(S-2)}\left(1+\frac{\lambda}{(S-2)\mu}\right)^{(S-2)} \tag{38}$$

$$= \exp\left(\frac{\lambda}{\mu}-2\right)\left(\frac{\mu}{\lambda}\right)^{S-2} S^{S-2}(1+o(1/S)). \tag{39}$$

As for the maximal local diameter, $\max_s D^{\pi}(s-1,s) \geq \max_s \tau^{\pi^0}(s-1,s) \geq \tau^{\pi^0}(S-2,S-1)$ and the same argument as before applies.

□

Let us now consider the bias of the optimal policy in M. Using Lemma 3.2, the bias $h^*(s)$ is decreasing and concave in s, with bounded increments. Therefore, its span, defined as span $(h^*) := \max_s h^*(s) - \min_s h^*(s)$, satisfies

$$(h^*(0)-h^*(1))S \leq \text{span } (h^*) \leq (h^*(S-2)-h^*(S-1))S \leq C(S-1).$$

This implies that the span of the bias behaves as a linear function of S for all M.

C Generic lemmas on ergodic MDPs

C.1 From bias variations to probability transition variations

The three first lemmas of this subsection are used in the proof of Lemma C.4. This lemma is needed to obtain equation (27).

Lemma C.1. *For a MDP with rewards $r \in [0, r_{\max}]$ and transition matrix P, denote by $J_s(\pi, T) := \mathbb{E}\left[\sum_{t=0}^{T} r(s_t, \pi(s_t))\right]$ the expected cumulative rewards until time T starting from state s, under policy π. Let D_π be the diameter under policy π. The following inequality holds: span $(J(\pi, T)) \leq r_{\max} D_\pi$.*

Proof. Let $s, s' \in \mathcal{S}$. Call $\tau_{s\to s'}$ the random time needed to reach state s' from state s under policy π. Then:

$$\begin{aligned} J_s(\pi,T) &= \mathbb{E}\left[\sum_{t=0}^{T} r(s_t)\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{\tau_{s\to s'}-1} r(s_t)\right] + \mathbb{E}\left[\sum_{t=\tau_{s\to s'}}^{T} r(s_t)\right] \\ &\leq r_{\max}\mathbb{E}\left[\tau_{s\to s'}\right] + J_{s'}(\pi,T) \\ &\leq r_{\max} D_\pi + J_{s'}(\pi,T), \end{aligned}$$

which proves the lemma. □

Lemma C.2. *Consider two ergodic MDPs M and M'. For $i \in 1, 2$, let $r_i \in [0, r_{\max}]$ and P_i be the rewards and transition matrix of MDP M_i under policy π_i, where both MDPs have the same state and action spaces. Denote by g_i the average reward obtained under policy π_i in the MDP M_i. Then the difference of the gains is upper bounded.*

$$|g-g'| \leq \|r-r'\|_\infty + r_{\max} D_\pi \|P-P'\|_\infty.$$

Proof. Define for any state s the following correction term $b(s) := r_{\max} D_\pi \|p(\cdot|s) - p'(\cdot|s)\|$. Let us show by induction that for $T \geq 0$,

$$\sum_{t=0}^{T-1} P^t r \leq \sum_{t=0}^{T-1} P'^t (r+b).$$

This is true for $T = 0$. Assume that the inequality is true for some $T \geq 0$, then

$$\begin{aligned}
\sum_{t=0}^{T} P^t r - \sum_{t=0}^{T} P'^t (r+b) &= -b + P \sum_{t=0}^{T-1} P^t r - P' \sum_{t=0}^{T-1} P'^t (r+b) \\
&= -b + P' \left(\sum_{t=0}^{T-1} P^t r - \sum_{t=0}^{T-1} P'^t (r+b) \right) + (P - P') \sum_{t=0}^{T} P^t r \\
&\leq -b + (P - P') \sum_{t=0}^{T} P^t r \text{ by induction hypothesis}
\end{aligned}$$

Notice that, for any state s:

$$\begin{aligned}
\left((P - P') \sum_{t=0}^{T} P^t r \right)(s) &\leq \|p(\cdot|s) - p'(\cdot|s)\| \cdot \text{span}\,(J(T)) \\
&\leq r_{\max} D_\pi \|p(\cdot|s) - p'(\cdot|s)\| \text{ by Lemma C.1} \\
&= b(s)
\end{aligned}$$

In the same manner we show that:

$$\sum_{t=0}^{T} P^t r \geq \sum_{t=0}^{T} P'^t (r - b).$$

Hence, as P' has non-negative coefficients, denoting by e the unit vector:

$$\left\| \sum_{t=0}^{T} P^t r - \sum_{t=0}^{T} P'^t r \right\|_\infty \leq \|b\|_\infty \left\| \sum_{t=0}^{T} P'^t \cdot e \right\|_\infty = \|b\|_\infty (T+1).$$

We can also show that:

$$\left\| \sum_{t=0}^{T} P'^t r - \sum_{t=0}^{T} P'^t r' \right\|_\infty = \left\| \sum_{t=0}^{T} P'^t (r - r') \right\|_\infty \leq \|r - r'\|_\infty (T+1)$$

And therefore with a multiplication by $\frac{1}{T+1}$ and by taking the Cesáro limit in $\left\| \sum_{t=0}^{T} P^t r - \sum_{t=0}^{T} P'^t r' \right\|_\infty$, and with a triangle inequality:

$$|g - g'| \leq \|r - r'\|_\infty + \|b\|_\infty,$$

where $\|b\|_\infty = r_{\max} D_\pi \|P - P'\|_\infty$. □

Lemma C.3. *Let P be the stochastic matrix of an ergodic Markov chain with state space $1, \ldots, S$. The matrix $A := I - P$ has a block decomposition*

$$A = \begin{pmatrix} A_S & b \\ c & d \end{pmatrix};$$

then A_S, of size $(S-1) \times (S-1)$ is invertible and $\|A_S^{-1}\|_\infty = \sup_{i \in \mathcal{S}} \mathbb{E}_i\, \tau_S$, where $\mathbb{E}_i\, \tau_S$ is the expected time to reach state S from state i.

Remark that this lemma is true for any state in $\mathcal{S}$.

Proof. $(\mathbb{E}_i\, \tau_S)_i$ is the unique vector solution to the system:

$$\begin{cases} v(S) = 0 \\ \forall i \neq S,\, v(i) = 1 + \sum_{j \in \mathcal{S}} P(i,j) v(j) \end{cases}$$

We can rewrite this system of equations as: $\tilde{A} v = e - e_S$, where $\tilde{A}$ is the matrix

$$\tilde{A} := \begin{pmatrix} A_S & b \\ 0 & 1 \end{pmatrix},$$

e the unit vector and e_S the vector with value 1 for the last state and 0 otherwise. Then $\widetilde{A}$ and A_S are invertible and we write:

$$\tilde{A}^{-1} = \begin{pmatrix} A_S^{-1} & -A_S^{-1}b \\ 0 & 1 \end{pmatrix}.$$

Thus, by computing $\tilde{A}^{-1}(e - e_S)$, for $i \neq S$, $(\mathbb{E}_i \, \tau_S)_i = A_S^{-1}e$. By definition of the infinite norm and using that A_S is an M-matrix and that its inverse has non-negative components, $\|A_S^{-1}\|_\infty = \sup_{i \in \mathcal{S}} \mathbb{E}_i \, \tau_S$. ☐

In the following lemma, we use the same notations as in Lemma C.2 with a common state space $\{1, \dots S\}$.

Lemma C.4. *Let the biases h, h' be the biases of the two MDPs verify their respective Bellman equations with the renormalization choice $h(S) = h'(S) = 0$. Let $\sup_{s \in \mathcal{S}} \mathbb{E}_s \, \tau^\pi_{s'}$ be the worst expected hitting time to reach the state s' with policy π, and call $T_{hit} := \inf_{s' \in \mathcal{S}} \sup_{s \in \mathcal{S}} \mathbb{E}_s \, \tau_{s'}$. We have the following control of the difference:*

$$\|h - h'\|_\infty \leq 2T_{hit}(D' r_{\max} \|P - P'\|_\infty + \|r - r'\|_\infty)$$

Notice that although the biases are unique up to a constant additive term, the renormalization choice does not matter as the unit vector is in the kernel of $(P - P')$.

Proof. The computations in this proof follow the same idea as in the proof of [10, Theorem 4.2]. The biases verify the following Bellman equations $r - ge = (I - P)h$, and also the arbitrary renormalization equations, thanks to the previous remark: $h(S) = 0$. Using the same notations as in the proof of Lemma C.3, we can write the system of equations $\tilde{A}h = \tilde{r} - \tilde{g}$, with $\tilde{r}$ and $\tilde{g}$ respectively equal to r and g everywhere but on the last state, where their value is replaced by 0.

We therefore have that $h = \tilde{A}^{-1}(\tilde{r} - \tilde{g})$, and with identical computations, $h' = \tilde{A'}^{-1}(\tilde{r'} - \tilde{g'})$. By denoting $dX := X - X'$ for any vector or matrix X, we get:

$$dh = -\tilde{A}^{-1}(d\tilde{r} - d\tilde{g} + d\tilde{A}h').$$

The previously defined block decompositions are:

$$\tilde{A}^{-1} = \begin{pmatrix} A_S^{-1} & -A_S^{-1}b \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad d\tilde{A} = \begin{pmatrix} A_S - A'_S & b - b' \\ 0 & 0 \end{pmatrix}.$$

For $s < S$, $dh(s) = -e_s^T A_S^{-1}(dA_S h' + d\tilde{r} - d\tilde{g})$ and $dh(S) = 0$. Now by taking the norm and using C.1:

$$\|dh\|_\infty \leq \|A_S^{-1}\|_\infty (r_{\max} D' \|dA_S\|_\infty + \|d\tilde{r}\| + |d\tilde{g}|).$$

Notice that $\|dA_S\|_\infty \leq \|dP\|_\infty$, $\|d\tilde{r}\| \leq \|dr\|$ and $\|d\tilde{g}\| = |dg|$. Using Lemma C.2 and Lemma C.3, and taking the infimum for the choice of the state of renormalization implies the claimed inequality for the biases. ☐

C.2 A McDiarmid's inequality

Lemma C.5. *Recall that $m^{\max}$ is the stationary measure of the Markov chain under policy $\pi^{\max}$, such that for every state s: $\pi^{\max}(s) = A_{\max}$.*

Let k be an episode, and assume that the length of this episode I_k is at least $I(T) = 1 + \max\left\{Q_{\max}, T^{1/4}\right\}$, with $Q_{\max} := \left(\frac{10D}{m^{\max}(S-1)}\right)^2 \log\left(\left(\frac{10D}{m^{\max}(S-1)}\right)^4\right)$. Then, with probability at least $1 - \frac{1}{4T}$:

$$\nu_k(x_k, a_k) \geq m^{\max}(S-1)I_k - 5D\sqrt{I_k \log I_k}.$$

We will now prove Lemma C.5:

Proof. Let k be an episode such that $I_k \geq I(T)$, and first consider it is of fixed length I. Denote by $\mathring{r}$ the vector of reward equal to 1 if the current state is x_k and 0 otherwise. Denote by $\mathring{g}_{\pi_k}$ the gain associated to the policy π_k for the transitions p and rewards $\mathring{r}$, and similarly define $\mathring{h}_{\pi_k}$ the bias, translated so that $\mathring{h}_{\pi_k}(S-1) = 0$. Notice in that case, that if we denote by m_k the stationary distribution under policy π_k, that $m^{\max}(S-1) \leq m_k(s)$ for any state s, assuming that $S \geq \frac{\lambda}{\mu} + 1$. Then:

$$
\begin{aligned}
\nu_k(x_k, a_k) &= \sum_{u=t_k}^{t_{k+1}-1} \mathring{r}(s_u) \\
&= \sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\pi_k} + \mathring{h}_{\pi_k}(s_u) - \left\langle p\left(\cdot|s_u, \pi_k(s_u)\right), \mathring{h}_{\pi_k} \right\rangle \text{ using a Bellman's equation} \\
&= \sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\pi_k} + \mathring{h}_{\pi_k}(s_u) - \mathring{h}_{\pi_k}(s_{u+1}) + \mathring{h}_{\pi_k}(s_{u+1}) - \left\langle p\left(\cdot|s_u, \pi_k(s_u)\right), \mathring{h}_{\pi_k} \right\rangle.
\end{aligned}
$$

By Azuma-Hoeffding inequality B.2, following the same proof as in section 4.3.2 of [11], notice that $X_u = \mathring{h}_{\pi_k}(s_{u+1}) - \left\langle p\left(\cdot|s_u, \pi_k(s_u)\right), \mathring{h}_{\pi_k} \right\rangle$ form a martingale difference sequence with $|X_u| \leq D$:

$$
\mathbb{P}\left\{ \sum_{u=t_k}^{t_{k+1}-1} X_u \geq D\sqrt{10 I \log I} \right\} \leq \frac{1}{I^5}.
$$

Using that $\left| \mathring{h}_{\pi_k}(s_{t_k}) - \mathring{h}_{\pi_k}(s_{t_{k+1}}) \right| \leq D$, with probability at least $1 - \frac{1}{I^2}$:

$$
\nu_k(x_k, a_k) \geq \sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\pi_k} - 5D\sqrt{I \log I}.
$$

On the other hand:

$$
\sum_{u=t_k}^{t_{k+1}-1} \mathring{g}_{\pi_k} = \nu_k(s_k, a_k) m_k(x_k),
$$

so that, using that $m_k(x_k) \geq m^{\max}(S-1)$, with probability at least $1 - \frac{1}{I^5}$:

$$
\nu_k(x_k, a_k) \geq m^{\max}(S-1) I - 5D\sqrt{I \log I}.
$$

We now use a union bound over the possible values of the episode lengths I_k, between $I(T)+1$ and T:

$$
\begin{aligned}
\mathbb{P}\left\{ \nu_k(x_k, a_k) < m^{\max}(S-1) I_k - 5D\sqrt{I_k \log I_k} \right\} &\leq \sum_{I=I(T)+1}^{T} \frac{1}{I^5} \leq \sum_{I=T^{1/4}+1}^{T} \frac{1}{I^5} \\
&\leq \frac{1}{4T},
\end{aligned}
$$

so that we now have that with probability at least $1 - \frac{1}{4T}$:

$$
\nu_k(x_k, a_k) \geq m^{\max}(S-1) I_k - 5D\sqrt{I_k \log I_k}.
$$

□

We can show a corollary of Lemma C.5 that we will use for the regret computations:

Corollary C.6. *For an episode k such that its length I_k is greater than $I(T)$,with probability at least $1 - \frac{1}{4T}$:*

$$
\nu_k(x_k, a_k) \geq \frac{m^{\max}(S-1)}{2} I_k.
$$

Proof. With Lemma C.5, it is enough to show that $5D\sqrt{I_k \log I_k} \leq \frac{m^{\max}(S-1)}{2} I_k$, *i.e.* that $\sqrt{\frac{I_k}{\log I_k}} \geq \frac{10D}{m^{\max}(S-1)} =: B$. By monotonicity, as $I_k \geq Q_{\max} = B^2 \log B^4$ we can show instead that $B^2 \log B^4 \geq B^2 \log \left(B^2 \log B^4\right)$.

This last inequality is true, using that $\log x \geq \log(2 \log x)$ for $x > 1$. This proves the corollary. $\square$