Input: Operator $\mathcal{L} : \mathbb{R}^S \to \mathbb{R}^S$ and accuracy $\epsilon > 0$
Set $v_0 = 0$, $v_1 = \mathcal{L}v_0$, $n = 0$

1. **While** $sp(v_{n+1} - v_n) > \epsilon$ **do**
   (a) $n = n + 1$
   (b) $v_{n+1} = \mathcal{L}v_n$
2. **Return:** $g_n = \frac{1}{2}\big(\max_s\{v_{n+1}(s) - v_n(s)\} + \min_s\{v_{n+1}(s) - v_n(s)\}\big)$ and v_n

Figure 3: EVI.

A Policy Evaluation with Uncertainties

Consider a bounded parameter MDP $\mathcal{M}$ defined by a compact set $B_r(s,a) \subseteq [0, r_{\max}]$ and $B_p(s,a) \in \Delta_S$:

$$\mathcal{M} = \{M = (\mathcal{S}, \mathcal{A}, r, p), r(s,a) \in B_r(s,a), p(\cdot|s,a) \in B_p(s,a), \forall (s,a) \in \mathcal{S} \times \mathcal{A}\} \tag{16}$$

In this paper, we consider confidence sets B_r and B_p that are polytopes. We are interested in building a pessimistic (robust) estimate of the performance of a policy $\pi \in \Pi^{\text{SD}}$ in $\mathcal{M}$. This robust optimization problem can be written as:

$$\underline{g}^\pi := \inf_{M \in \mathcal{M}} \{g^\pi(M)\} \tag{17}$$

where $g^\pi(M)$ is the gain of policy π in the MDP M. Lemma 5 shows that there exists a solution to this problem that can be computed using EVI when the set $\mathcal{M}$ contains an ergodic MDP.

We recall that any bounded parameter MDP admits an equivalent representation as an extended MDP (Jaksch et al., 2010) with identical state space $\mathcal{S}$ but compact action space. For a deterministic policy $\pi \in \Pi^{\text{SD}}$, the extended (pessimistic) Bellman operator $\mathcal{L}_\pi$ is defined as:

$$\forall v \in \mathbb{R}^S, \forall s \in \mathcal{S}, \quad \mathcal{L}_\pi v(s) := \min_{r \in B_r(s,\pi(s))} r + \min_{p \in B_p(s,\pi(s))} \{p^\mathsf{T} v\} \tag{18}$$

Lemma 5. *Let $\mathcal{M}$ be a bounded-parameter MDP defined as in Eq. 16 such that exists an* ergodic *MDP $M \in \mathcal{M}$ w.h.p. Consider a policy $\pi \in \Pi^{SD}$, then:*

1. *There exists a tuple $(\widetilde{g}, \widetilde{h}) \in \mathbb{R} \times \mathbb{R}^S$ such that:*

$$\forall s \in \mathcal{S}, \quad \widetilde{g} + \widetilde{h}(s) = \mathcal{L}_\pi \widetilde{h}(s)$$

   where $\mathcal{L}_\pi$ is the Bellman operator of the extended MDP $\mathcal{M}^+$ associated to $\mathcal{M}$ (see Eq. 18).

2. *In addition, we have the following inequalities on the pair $(\tilde{g}, \tilde{h})$:*

$$\tilde{g} \le g^\pi(M) \qquad \text{and} \qquad sp(\tilde{h}) \le \max_{\pi \in \Pi^{SD}(M)} \max_{s \neq s'} \mathbb{E}_M^\pi\left(\tau(s')|s\right) := \Upsilon < +\infty$$

   where $\mathbb{E}_M^\pi$ is the expectation of using policy π in the MDP M and $\tau(s')$ is the minimal number of steps to reach state s'.

Proof. **Point 1.** We show that this policy evaluation problem is equivalent to a planning problem in an extended MDP $\mathcal{M}^-$ with negative reward. Consider the extended MDP $\mathcal{M}^- = (\mathcal{S}, \mathcal{A}^-, p^-, r^-)$ such that $\mathcal{A}_s^- = \{\pi(s)\} \times B_r(s,\pi(s)) \times B_p(s,\pi(s))$. For any state $s \in \mathcal{S}$ and action $a^- = (\pi(s), r(s,\pi(s)), p(\cdot|s,\pi(s))) \in \mathcal{A}_s$,

$$r^-(s,a^-) = -r(s,\pi(s))$$
$$p^-(\cdot|s,a^-) = p(\cdot|s,\pi(s))$$

Denote by $\mathcal{L}^-$ the optimal Bellman operator of $\mathcal{M}^-$. Since $B_r(s,\pi(s))$ and $B_p(s,\pi(s))$ are polytopes, $\mathcal{L}^-$ can be interpreted as an optimal Bellman operator with finite number of actions. A sufficient condition for the

existence of a solution of the optimality equations is that the MDP is weakly communicating (Puterman, 1994, Chap. 8-9). Note that $\mathcal{M}^-$ contains the model defined by P^π, i.e., the Markov chain induced by π in M.[7] Since P^π is ergodic, $\mathcal{M}^-$ is at least communicating and thus $\mathcal{L}^-$ converges to a solution of the optimality equations. Extended value iteration (Jaksch et al., 2010) on $\mathcal{L}^-$ converges toward a gain and bias (g^-, h^-) such that:

$$\begin{aligned} g^- + h^-(s) = L^- h^-(s) &= \max_{a \in \mathcal{A}_s^-} \{r^-(s,a) + p^-(\cdot|s,a)^\mathsf{T} h^-\} \\ &= \max_{r \in B_r(s,\pi(s))} \{-r\} + \max_{p \in B_p(s,\pi(s))} p^\mathsf{T} h^- \\ &= -\min\{B_r(s,\pi(s))\} + \max_{p \in B_p(s,\pi(s))} p^\mathsf{T} h^- \end{aligned}$$

By rearranging, we have that:

$$\begin{aligned} -g^- + (-h^-)(s) &= \min\{B_r(s,\pi(s))\} + \min_{p \in B_p(s,\pi(s))} p^\mathsf{T}(-h^-) \\ &= \mathcal{L}_\pi(-h^-)(s) \end{aligned}$$

Thus follows that $\widetilde{g} = -g^-$ and $\widetilde{h} = -h^-$. This shows the relationship between maximizing over policies in the extended MDP $\mathcal{M}^-$ and minimizing over the set of models induced by π.

Point 2. Let's begin by bounding the span of the bias $\tilde{h}$. Thanks to Theorem 4 of Bartlett and Tewari (2009), we have that the span of $\tilde{h}$ is upper-bounded by the diameter of the extended MDP $\mathcal{M}^-$, i.e:

$$sp(\tilde{h}) \leq \max_{s \neq s'} \inf_{\pi^- \in \Pi^{SD}(\mathcal{M}^-)} \mathbb{E}_{\pi^-}(\tau(s')|s)$$

where $\mathbb{E}_{\pi^-}$ is the expectation of using policy π^- in the extended MDP $\mathcal{M}^-$ and $\tau(s')$ is the hitting time of state s'. But let's define the policy $\pi^\star$ in the extended MDP $\mathcal{M}^-$ such that for a state s, it chooses the action:

$$\pi^\star(s) = (\pi(s), r^\star(s,\pi(s)), p^\star(.|s,\pi(s)))$$

with $r^\star$ and $p^\star$ the true parameter of the MDP M, this is possible because w.h.p the MDP $M^\pi \in \mathcal{M}^-$ with M^π the Markov chain induced by using policy π in the MDP M. Thus for any pair of states (s, s'):

$$\mathbb{E}_{\pi^\star}(\tau(s')|s) = \mathbb{E}_M^\pi(\tau(s')|s)$$

with $\mathbb{E}_M^\pi$ the expectation of using policy π in the MDP M. Therefore:

$$\begin{aligned} sp(\tilde{h}) &\leq \max_{s \neq s'} \inf_{\pi^- \in \Pi^{SD}(M^-)} \mathbb{E}_{\pi^-}(\tau(s')|s) \\ &\leq \mathbb{E}_M^\pi(\tau(s')|s) \leq \Upsilon := \max_{\pi \in \Pi^{SD}(M)} \max_{s \neq s'} \mathbb{E}_M^\pi(\tau(s')|s) \end{aligned}$$

And $\Upsilon < +\infty$ because M is assumed to be ergodic.

Let's show that the gain $\tilde{g}$ is a lower bound on the gain of the policy π in the MDP M. Indeed, because the operator $\mathcal{L}^-$ converges toward solution of the optimality equations for negative rewards, we have that, see (Puterman, 1994, Th. 8.4.1):

$$g^- \geq -g^\pi(M)$$

because reversing the sign of the rewards in the MDP M changes the sign of the gain of a policy. Thus, $\tilde{g} \leq g^\pi(M)$. □

As a consequence, we can use EVI on $\mathcal{L}_\pi$ to compute a solution for problem 17. EVI generates a sequence of vectors (v_i) such that $v_{i+1} = \mathcal{L}_\pi v_i$ and $v_0 = 0$. If the algorithm is stopped when $sp(v_{n+1} - v_n) \leq \epsilon$ we have (Puterman, 1994, Sec. 8.3.1) that:

$$|g_n - \widetilde{g}| \leq \epsilon/2 \quad \text{and} \quad \|\mathcal{L}_\pi v_n - v_n - g_n e\|_\infty \leq \epsilon \tag{19}$$

where $e = (1, \ldots, 1)$ and $g_n = \frac{1}{2}(\max_s\{v_{n+1}(s) - v_n(s)\} + \min_s\{v_{n+1}(s) - v_n(s)\})$. The following lemma shows how we can use the value produced by EVI to lower bound the expected sum of rewards under a policy π.

[7] We abuse of language since $\mathcal{M}^-$ is not formally a set. We should formally refer to the bounded parameter MDP associated to $\mathcal{M}^-$, i.e., built considering $B_p(s,\pi(s))$ and $B_r(s,\pi(s))$. Note that $p(\cdot|s,\pi(s)) \in B_p(s,\pi(s))$ w.h.p.

Lemma 6. *Let (g_n, v_n) the values computed by* EVI *using $\mathcal{L}_\pi$ and an accuracy ϵ. Then, the cumulative reward collected by policy π in M after t steps can be lower bounded by:*

$$\forall y \in \mathcal{S}, \quad \mathbb{E}_M \left[\sum_{i=1}^{t} r_i | s_1 = y, \pi \right] \geq t(g_n - \epsilon) - sp(v_n)$$

In addition,

$$sp(v_n) \leq \Upsilon$$

Proof. Using the inequalities in (19) we can write that:

$$\begin{aligned} v_n(s) + g_n \leq \mathcal{L}_\pi v_n(s) &= \min_{r \in B_r(s,\pi(s))} r + \min_{p \in B_p(s,\pi(s))} \{p^\intercal v\} + \epsilon \\ &\leq r(s,\pi(s)) + p(\cdot|s,\pi(s))^\intercal v_n + \epsilon \end{aligned}$$

since $r(s,\pi(s)) \in B_r(s,\pi(s))$ and $p(\cdot|s,\pi(s)) \in B_p(s,\pi(s))$ w.h.p. By iterating this inequality, we get that for all $t > 0$ and state s :

$$v_n(s) + t g_n \leq (t-1)\varepsilon + p^t(\cdot|s,\pi(s))^\intercal v_n + \mathbb{E}\left[\sum_{i=1}^{t} r_i \left(s_i, \pi(s_i)\right) | s_1 = s \right]$$

The statement follows by noticing that

$$sp(v_n) = \max_s v_n(s) - \min_s v_n(s) \geq \underbrace{p^t(\cdot|y,\pi(y))^\intercal v_n}_{\leq \max_s v_n(s)} - \underbrace{v_n(y)}_{\geq \min_s v_n(s)}, \quad \forall y \in \mathcal{S}$$

The last statement is a direct consequence of the argument developed in section 4.3.1 of Jaksch et al. (2010). This reasoning relies on the fact that the initial vector used in EVI is a zero span vector. □

B Regret Bound for CUCRL

Lemma 7. *The regret of* CUCRL2 *can be upper-bounded for some $\beta > 0$, with probability at least $1 - \frac{2\delta}{5}$, by:*

$$R(\text{CUCRL2}, T) \leq \beta \cdot \left(R(\text{UCRL2}, T | \Lambda_T) + (g^\star - g^{\pi_b}) \sum_{k \in \Lambda_T^c} T_k + \max\{r_{\max}, sp\left(h^{\pi_b}\right)\} \sqrt{SAT \ln(T/\delta)} \right)$$

Proof. Recall that $k_t = \sup\{k > 0 : t > t_k\}$ is the episode at time t and that the regret is defined as $R(\text{CUCRL2}, T) = \sum_{t=1}^{T} \left(g^\star - r_t(s_t, a_t) \right)$.

Since the baseline policy π_b may be stochastic, as a first step we replace the observed reward by its expectation. As done in (Fruit et al., 2018b) we use Azuma's inequality that gives, with probability at least $1 - \frac{\delta}{5}$:

$$\forall T \geq 1, \quad -\sum_{t=1}^{T} r_t \leq -\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \pi_{k_t}(s_t, a) r(s_t, a) + 2 r_{\max} \sqrt{T \ln\left(\frac{5T}{\delta}\right)} \tag{20}$$

We denote by $\Lambda_T = \Lambda_{k_T} \cup \{k_T\} \cdot \mathbb{1}_{(Eq.\ 15)}$ the set of episodes where the algorithm played an UCRL policy. Note that we cannot directly consider Λ_{k_T} since the set is updated at the end of the episode and the last episode may not have ended at T. Similarly we denote by $\Lambda_T^c = \Lambda_{k_T}^c \cup \{k_T\} \cdot \mathbb{1}_{(\neg Eq.\ 15)}$. Then, the regret of CUCRL2 can be

decomposed as follow:

$$\begin{aligned} R(\text{CUCRL2}, T) &= \sum_{t=1}^{T} \left(g^\star - \sum_{a \in \mathcal{A}} \pi_{k_t}(s_t, a) r(s_t, a) \right) + 2 r_{\max} \sqrt{T \ln\left(\frac{5T}{\delta}\right)} \\ &= 2 r_{\max} \sqrt{T \ln\left(\frac{5T}{\delta}\right)} + \underbrace{\sum_{k=1}^{k_T} \mathbb{1}_{(k \in \Lambda_T)} \sum_{t=t_k}^{t_{k+1}-1} (g^\star - r(s_t, a_t))}_{:= R(\text{UCRL2}, T | \Lambda_T)} \\ &\quad + \sum_{k=1}^{k_T} \mathbb{1}_{(k \in \Lambda_T^c)} \left((g^\star - g^{\pi_b})(t_{k+1} - t_k) + \underbrace{\sum_{t=t_k}^{t_{k+1}-1} \left(g^{\pi_b} - \sum_{a \in \mathcal{A}} \pi_b(s_t, a) r(s_t, a) \right)}_{:= \Delta_k^c} \right) \end{aligned} \tag{21}$$

Moreover, note that the UCRL2 policy is deterministic so we hate that $\sum_{a \in \mathcal{A}} \pi_{k_t}(s_t, a) r(s_t, a) = r(s_t, a_t)$ when $k_t \in \Lambda_T$. The second term, denoted $R(\text{UCRL2}, T | \Lambda_{k_T})$, is the regret suffered by UCRL2 over $\sum_{k \in \Lambda_{k_T}} T_k$ steps. The only difference with the orginal analysis (Jaksch et al., 2010) is that the confidence intervals used by UCRL are updated when using the baseline policy, however it does not affect the regret of UCRL because it only means the confidence intervals used shrinks faster for some state-action pairs. We will analyze this term in Lem. 9. To decompose Δ_k^c we can use the Bellman equations ($g^{\pi_b} e = L^{\pi_b} h^{\pi_b} - h^{\pi_b}$):

$$\begin{aligned} \sum_{k \in \Lambda_T^c} \Delta_k^c &= \sum_{k \in \Lambda_T^c} \sum_{t=t_k}^{t_{k+1}-1} \sum_{a} \pi_b(s_t, a) p(\cdot | s_t, a)^\intercal h^{\pi_b} - h_{\pi^b}(s_t) \\ &= \sum_{k \in \Lambda_T^c} \sum_{t=t_k}^{t_{k+1}-1} \underbrace{\sum_{a \in \mathcal{A}} \pi_b(s, a) \left(p(\cdot | s_t, a)^\intercal h^{\pi_b} \right) - h^{\pi_b}(s_{t+1})}_{:= \Delta_{k,t}^{c,p}} + \sum_{k \in \Lambda_T^c} \underbrace{\sum_{t=t_k}^{t_{k+1}-1} (h^{\pi_b}(s_{t+1}) - h^{\pi_b}(s_t))}_{:= \Delta_k^{c,2}} \end{aligned}$$

But, $\Delta_k^{c,2}$ can be bounded using a telescopic sum argument and the number of episodes:

$$\sum_{k \in \Lambda_T^c} \Delta_k^{c,2} = \sum_{k \in \Lambda_T^c} h^{\pi_b}(s_{t_{k+1}}) - h^{\pi_b}(s_{t_k}) \leq |\Lambda_T^c| sp\left(h^{\pi_b}\right)$$

Then it is easy to see that $(\Delta_{k,t}^{c,p})_{k,t}$ is a Martingale Difference Sequence with respect to the filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$ which is generated by all the randomness in the environment and in the algorithm up until time t: $|\Delta_{k,t}^{c,p}| \leq 2 \|h^{\pi_b}\|_\infty \leq 2 sp(h^{\pi_b})$ and $\mathbb{E}[\Delta_{k,t}^{c,p} | \mathcal{F}_t] = 0$. Thus with probability $1 - \frac{\delta}{5}$:

$$\sum_{k \in \Lambda_{k_T}^c} \sum_{t=t_k}^{t_{k+1}-1} \Delta_{k,t}^{c,p} \leq 4 sp\left(h^{\pi_b}\right) \sqrt{T \ln\left(\frac{5T}{\delta}\right)}$$

Therefore putting all the above together, we have that with probability at least $1 - \frac{2\delta}{5}$:

$$\begin{aligned} R(\text{CUCRL2}, T) \leq 2 r_{\max} \sqrt{T \ln\left(\frac{5T}{\delta}\right)} + R(\text{UCRL2}, T | \Lambda_T) + (g^\star - g^{\pi_b}) \sum_{k=1}^{k_T} \mathbb{1}_{(k \in \Lambda_T^c)} (t_{k+1} - t_k) \\ + sp\left(h^{\pi_b}\right) \left(|\Lambda_T^c| + 4 \sqrt{T \ln\left(\frac{5T}{\delta}\right)} \right) \end{aligned}$$

As shown in (Ouyang et al., 2017, Lem. 1), $k_T \leq \sqrt{2SAT \ln(T)}$ thus we can simply write that $|\Lambda_T^c| \leq \sqrt{2SAT \ln(T)}$. □

In the next lemma, we bound the total number of steps where CUCRL2 used the baseline policy.

Lemma 8. *For any, $\delta > 0$, the total length of episodes where the baseline policy is played by CUCRL after T steps is upper-bounded with probability $1 - 2\delta/5$ by:*

$$\sum_{l \in \Lambda^c_{k_T}} T_l \leq 2\sqrt{SAT\ln(T)} + \frac{16\sqrt{TL^\delta_T}}{\alpha g^{\pi_b}}\Big[(D+\Upsilon)\sqrt{SA} + r_{\max} + \sqrt{SA}sp(h^{\pi_b})\Big] + \frac{112SAL^\delta_T}{(\alpha g^{\pi_b})^2}(1 + S(D+\Upsilon)^2)$$

with $L^\delta_T := \ln\left(\frac{5SAT}{\delta}\right)$ a logarithmic term in T.

Proof. Let τ be the last episode played conservatively: $\tau = \sup\{k > 0 : k \in \Lambda^c_k\}$. At the beginning of episode τ the conservative condition is not verified that is to say:

$$\underbrace{\sum_{l \in \Lambda_{\tau-1}} T_l\left(g^{\pi_b} - g_l^- + \varepsilon_l\right)}_{:=\Delta^1_\tau} + sp(h^{\pi_b})\left(|\Lambda^c_{\tau-1}| + (1-\alpha)\right) + \sum_{l \in \Lambda_{\tau-1}\cup\{\tau\}} sp\left(h_l^-\right) +$$

$$+ (T_{\tau-1}+1)\left((1-\alpha)g^{\pi_b} - g_\tau^- + \epsilon_\tau\right)\mathbb{1}_{\{(1-\alpha)g^{\pi_b} \geq g_\tau^- + \epsilon_\tau\}} \geq \alpha\sum_{l=1}^{\tau-1} T_l g^{\pi_b} \tag{22}$$

Let's proceeding by analysing each term on the RHS of Eq. 22. First, we have that $|\Lambda^c_{\tau-1}| \leq k_T \leq \sqrt{2SAT\ln(T)}$, thus:

$$sp(h^{\pi_b})\left(|\Lambda^c_{\tau-1}| + (1-\alpha)\right) \leq \left(\sqrt{2SAT\ln(T)} + 1\right)sp(h^{\pi_b}) \tag{23}$$

On the other hand, thanks to Lem. 6, we have:

$$\sum_{l \in \Lambda_{\tau-1}\cup\{\tau\}} sp\left(h_l^-\right) \leq (|\Lambda_{\tau-1}|+1)\Upsilon \leq 2\sqrt{2SAT\ln(T)}\Upsilon \tag{24}$$

Before analysing Δ^1_τ, let's bound the contribution of episode τ:

$$(T_{\tau-1}+1)\left((1-\alpha)g^{\pi_b} - g_\tau^- - \epsilon_\tau\right)\mathbb{1}_{\{(1-\alpha)g^{\pi_b} \geq g_\tau^- + \epsilon_\tau\}} \leq (1-\alpha)g^{\pi_b}k_T \leq (1-\alpha)r_{\max}\sqrt{2SAT\ln(T)} \tag{25}$$

where we used the fact that for all episode k, we have $T_k \leq k$. Indeed the dynamic episode condition is such that for an episode k, $T_k \leq T_{k-1} + 1$ thus by iterating this inequality, $T_k \leq T_0 + k = k$. At this point using equations 22, 24 and 25 we have:

$$\Delta^1_\tau + \left(\sqrt{2SAT\ln(T)} + 1\right)sp(h^{\pi_b}) + 2\sqrt{2SAT\ln(T)}\Upsilon + (1-\alpha)r_{\max}\sqrt{2SAT\ln(T)} \geq \alpha\sum_{l=1}^{\tau-1} T_l g^{\pi_b}$$

Let's finish by analysing Δ^1_τ. Let's define the event, $\Gamma = \left\{\exists T > 0, \exists k \geq 1, \text{ s.t } \ M \notin \mathcal{M}_k\right\}$, by definition of B^k_r and B^k_p, $\mathbb{P}(\Gamma) \leq \delta/5$, see (Lazaric et al., 2019, App. B.2) for a complete proof. We have that on the event Γ^c, for any $l \in \Lambda_{\tau-1}$, $(g_l^-, h_l^-) = \text{EVI}(\mathcal{L}_l^{\pi_l}, \varepsilon_l)$ is such that $|\underline{g}^{\pi_l} - g_l^-| \leq \epsilon_l$ (see App. A) where $\underline{g}^{\pi_l}$ is the true gain: $\underline{g}^{\pi_l} + \underline{h}^{\pi_l} = \mathcal{L}_l^{\pi_l}\underline{h}^{\pi_l}$. Thus, since $\varepsilon_l \leq r_{\max}/\sqrt{t_l}$:

$$\Delta^1_\tau = \sum_{l \in \Lambda_{\tau-1}} T_l\left(g^{\pi_b} - g_l^- + \epsilon_l\right) \leq 2\sum_{l \in \Lambda_{\tau-1}} T_l\varepsilon_l + \sum_{l \in \Lambda_{\tau-1}} T_l(g^{\pi_b} - \underline{g}^{\pi_l})$$

$$\leq 4r_{\max}\sqrt{T} + \sum_{l \in \Lambda_{\tau-1}} T_l(\widetilde{g}_l - \underline{g}^{\pi_l})$$

where $\widetilde{g}_l$ is the optimistic gain at episode l (see Lazaric et al. (2019)) thus the last inequality comes from $g^{\pi_b} \leq g^\star \leq \widetilde{g}_l$ for every episode l. We can also define the optimistic bias at episode l, $\widetilde{h}_l$, the pair $(\widetilde{g}_l, \widetilde{h}_l)$ is such that:

$$\forall s \in \mathcal{S}, \qquad \widetilde{g}_l + \widetilde{h}_l(s) := \mathcal{L}_l^+\widetilde{h}_l := \max_a\left\{\max_{r \in B^l_r(s,a)} r + \max_{p \in B^l_p(s,a))} p^\intercal\widetilde{h}_l\right\}$$

Recall that $\widetilde{\pi}_l \in \Pi^{\mathrm{SD}}$ is the optimistic policy at episode l and when $l \in \Lambda_{\tau-1}$, $\pi_l = \widetilde{\pi}_l$. Then, by using Bellman equations:

$$\begin{aligned}
\sum_{l\in\Lambda_{\tau-1}} T_l(\widetilde{g}_l - \underline{g}^{\pi_l}) &= \sum_{l\in\Lambda_{\tau-1}} \sum_{t=t_l}^{t_{l+1}-1} (\widetilde{g}_l - \underline{g}^{\pi_l}) = \sum_{l\in\Lambda_{\tau-1}} \sum_{t=t_l}^{t_{l+1}-1} \big(\underbrace{\mathcal{L}_l^{+}\widetilde{h}_l(s_t)}_{:=\mathcal{L}_l^{+,\pi_l}\widetilde{h}_l(s_t)}, -\widetilde{h}_l(s_t) - \mathcal{L}_l^{\pi_l}\underline{h}^{\pi_l}(s_t) + \underline{h}^{\pi_l}(s_t)\big) \\
&= \sum_{l\in\Lambda_{\tau-1}} \sum_{t=t_l}^{t_{l+1}-1} \max_{r\in B_r^l(s_t,a_t)} r - \min_{r\in B_r^l(s_t,a_t)} r + \max_{p\in B_p^l(s_t,a_t)} p^{\intercal}\widetilde{h}_l - \min_{p\in B_p^l(s_t,a_t)} p^{\intercal}\underline{h}^{\pi_l} - \widetilde{h}_l(s_t) + \underline{h}^{\pi_l}(s_t) \\
&\le \sum_{l\in\Lambda_{\tau-1}} \sum_{t=t_l}^{t_{l+1}-1} 2 \max_{r\in B_r^l(s_t,a_t)} r + \max_{q\in B_p^l(s_t,a_t)} (q-p^\star)^{\intercal}\widetilde{h}_l \\
&\quad - \min_{q\in B_p^l(s_t,a_t)} (q-p^\star)^{\intercal}\underline{h}^{\pi_l} + p^\star(\cdot|s_t,a_t)^{\intercal}\left(\widetilde{h}_l - \underline{h}^{\pi_l}\right) - \big(\widetilde{h}_l(s_{t+1}) - \underline{h}^{\pi_l}(s_{t+1})\big) \\
&\quad + \big(\widetilde{h}_l(s_{t+1}) - \widetilde{h}_l(s_t) + \underline{h}^{\pi_l}(s_{t+1}) - \underline{h}^{\pi_l}(s_t)\big)
\end{aligned}$$

where $p^\star$ is the transition probability of the true MDP, $M^\star$. By a simple telescopic sum argument, we have:

$$\sum_{l\in\Lambda_{\tau-1}} \sum_{t=t_l}^{t_{l+1}-1} \widetilde{h}_l(s_{t+1}) - \widetilde{h}_l(s_t) + \underline{h}^{\pi_l}(s_{t+1}) - \underline{h}^{\pi_l}(s_t) = |\Lambda_{\tau-1}|\big(sp(\widetilde{h}_l) + sp(\underline{h}^{\pi_l})\big)$$

At this point we need to explicitly define the concentration inequality used to construct the confidence sets B_r^l and B_p^l. For every $(s,a) \in \mathcal{S}\times\mathcal{A}$, we define $\beta_l^k(s,a)$ such that:

$$\forall l \ge 1, \qquad B_r^l(s,a) \subset [\widetilde{r}_l(s,a) - \beta_r^l(s,a), \widetilde{r}_l(s,a) + \beta_r^l(s,a)]$$

where $\widehat{r}_l(s,a)$ is the empirical average of the reward received when visiting the state-action pairs (s,a) at the beginning of episode l. For every $(s,a) \in \mathcal{S}\times\mathcal{A}$, we define $\beta_p^l(s,a)$ as:

$$B_p^l(s,a) = \big\{p \in \Delta_S : \|p(\cdot|s,a) - \widehat{p}_l(\cdot|s,a)\|_1 \le \beta_p^l(s,a)\big\}$$

with $\widehat{p}_l$ is the empirical average of the observed transitions. Choosing those β_r^l and β_p^l is done thanks to concentration inequalities such that event Γ^c holds with high enough probability. In the following, we use:

$$\forall s,a \qquad \beta_r^l(s,a) = \sqrt{\frac{7SAL_T^\delta}{2\max\{1,N_l(s,a)\}}} \text{ and } \beta_p^l(s,a) = S\sqrt{\frac{14AL_T^\delta}{\max\{1,N_l(s,a)\}}}$$

where $L_T^\delta = \ln\left(\frac{5SAT}{\delta}\right)$. For other choices of β_r^l and β_p^l refer to (Lazaric et al., 2019). Similarly to what done in (Jaksch et al., 2010, Sec. 4.3.1 and 4.3.2), by using Holder's inequality and recentering the bias functions, we write:

$$\begin{aligned}
\sum_{l\in\Lambda_{\tau-1}} T_l(\widetilde{g}_l - \underline{g}^{\pi_l}) \le |\Lambda_{\tau-1}|\big(sp(\widetilde{h}_l) + sp(\underline{h}^{\pi_l})\big) &+ \underbrace{\sum_{l\in\Lambda_{\tau-1}} \sum_{t=t_l}^{t_{l+1}-1} 2\beta_r^l(s_t,a_t) + \beta_p^l(s_t,a_t)\big(sp(\widetilde{h}_l) + sp(\underline{h}^{\pi_l})\big)}_{:=(a)} \\
&+ \underbrace{\sum_{l\in\Lambda_{\tau-1}} \sum_{t=t_l}^{t_{l+1}-1} p^\star(\cdot|s_t,a_t)^{\intercal}\left(\widetilde{h}_l + \underline{h}^{\pi_l}\right) - \big(\widetilde{h}_l(s_{t+1}) - \underline{h}^{\pi_l}(s_{t+1})\big)}_{:=(b)}
\end{aligned}$$

To finish, the proof of this lemma, we need to bound the term (a) and (b). In the following, we use the fact that $sp(\widetilde{h}_l) + sp(\underline{h}^{\pi_l}) \le D + \Upsilon$ (see Lem. 6) and again that $|\Lambda_{\tau-1}| \le k_T \le \sqrt{2SAT\ln(T)}$. Let's begin with (a), by definition of the radius of the confidence sets, we have:

$$\sum_{l\in\Lambda_{\tau-1}} \sum_{t=t_l}^{t_{l+1}-1} \beta_r^l(s_t,a_t) = \sqrt{\frac{7SAL_T^\delta}{2}} \sum_{l\in\Lambda_{\tau-1}} \sum_{t=t_l}^{t_{l+1}-1} \sqrt{\frac{1}{\max\{1,N_l(s_t,a_t)\}}} \le \sqrt{\frac{7SAL_T^\delta}{2}} \sqrt{\sum_{l=1}^{\tau-1} T_l}$$

and,

$$\sum_{l\in\Lambda_{\tau-1}}\sum_{t=t_l}^{t_{l+1}-1}\beta_p^l(s_t,a_t)\le S\sqrt{14L_T^\delta A}\sqrt{\sum_{l=1}^{\tau-1}T_l}$$

The second term (b) is easy to bound because it is a Martingale Difference Sequence with respect to the filtration generated by all the randomness in the algorithm and the environment before the current step. For any time t, the σ-algebra generated by the history up to time t included is $\mathcal{F}_t=\sigma(s_1,a_1,r_1,\dots,s_t,a_t,r_t,s_{t+1})$. Define $X_t=\mathbb{1}_{(k_t\in\Lambda_T)}(p(\cdot|s_t,\pi_{k_t}(s_t))^\mathsf{T}u_{k_t}-u_{k_t}(s_{t+1}))$ with $u_{k_t}=\widetilde{h}_{k_t}-\underline{h}^{\pi_{k_t}}$. Since π_{k_t} is $\mathcal{F}_t$ measurable, $E[X_t|\mathcal{F}_{t-1}]=0$ and $|X_t|\le 2(D+\Upsilon)$. Then $(X_t,\mathcal{F}_t)_t$ is an MDS and nothing change compared to the analysis of UCRL2. Therefore using Azuma-Hoeffding inequality, we have, with probability $1-\frac{\delta}{5}$ that:

$$(b)\le 2(D+\Upsilon)\sqrt{2TL_T^\delta}$$

⚠ Algorithmically, it is possible to evaluate the gain of the policies played in the past episodes at the beginning of the current episode. While this will provide a better estimate for the conservative condition, it will break the MDS structure in *(b)* since $\underline{h}^{\pi_l}$ will be not measurable w.r.t. $\mathcal{F}_l$ since it is computed with samples collected after episode l. Thus putting the bound for (a) and (b) together, we have:

$$\begin{aligned}\sum_{l\in\Lambda_{\tau-1}}T_l(\widetilde{g}_l-\underline{g}^{\pi_l})\le{}&(D+\Upsilon)\sqrt{2SAT\ln(T)}+\sqrt{14SAL_T^\delta}\sqrt{\sum_{l=1}^{\tau-1}T_l}+\big(D+\Upsilon\big)S\sqrt{14L_T^\delta A}\sqrt{\sum_{l=1}^{\tau-1}T_l}\\&+2(D+\Upsilon)\sqrt{2TL_T^\delta}\end{aligned}$$

That is to say,

$$\begin{aligned}&\boxed{\begin{aligned}&4r_{\max}\sqrt{T}+(D+\Upsilon)\sqrt{2SAT\ln(T)}+2(D+\Upsilon)\sqrt{2TL_T^\delta}\\&+\Big(\sqrt{2SAT\ln(T)}+1\Big)sp(h^{\pi_b})+\sqrt{2SAT\ln(T)}\Upsilon+(1-\alpha)r_{\max}\sqrt{2SAT\ln(T)}\end{aligned}}:=b_T\\&+\sqrt{14SAL_T^\delta}\sqrt{\sum_{l=1}^{\tau-1}T_l}+\big(D+\Upsilon\big)S\sqrt{14L_T^\delta A}\sqrt{\sum_{l=1}^{\tau-1}T_l}\ge\alpha\sum_{l=1}^{\tau-1}T_lg^{\pi_b}\end{aligned}$$

Rearranging the terms and calling $X=\sum_{l=1}^{\tau-1}T_l$, we have:

$$\alpha g^{\pi_b}X\le b_T+\left(\sqrt{14SAL_T^\delta}+\big(D+\Upsilon\big)S\sqrt{14L_T^\delta A}\right)\sqrt{X}$$

We have a quadratic equation and thus:

$$\sum_{l=1}^{\tau-1}T_l\le\frac{2b_T}{\alpha g^{\pi_b}}+\frac{56SAL_T^\delta}{(\alpha g^{\pi_b})^2}(2+2S(D+\Upsilon)^2)$$

Therefore, as τ is the last episode where CUCRL2 played the policy π_b, we have $\sum_{l\in\Lambda_T^c}T_l=\sum_{l\in\Lambda_\tau^c}T_l$. Also, because of the condition on the length of an episode $T_k\le k$ for every k, therefore:

$$\sum_{l\in\Lambda_T^c}T_l=\sum_{l\in\Lambda_\tau^c}T_l\le\sum_{l=1}^{\tau-1}T_l+T_\tau\le k_T+\frac{2b_T}{\alpha g^{\pi_b}}+\frac{56SAL_T^\delta}{(\alpha g^{\pi_b})^2}(2+2S(D+\Upsilon)^2)$$

□

The following lemma states the regret of the UCRL2 algorithm conditioned on running only the episodes in the set Λ_T.

Lemma 9. *For any $\delta > 0$, we have that after T, the regret of* UCRL2 *is upper bounded with probability at least $1 - \delta/5$ by:*

$$R(\text{UCRL2}, T|\Lambda_T) \leq \beta DS\sqrt{AT\ln\left(\frac{5T}{\delta}\right)} + \beta DS^2 A \ln\left(\frac{5T}{\delta}\right)$$

with β a numerical constant.

Proof. The same type of bound has been shown in numerous work before Jaksch et al. (2010); Lazaric et al. (2019), however the proof presented in those works can not be readily applied to our setting. Indeed, when the algorithm chooses to play the baseline policy for an episode, then the confidence sets used in CUCRL2 are updated for the state-action pairs encountered during this episode. However, in the classic proof for the UCRL2 algorithm the confidence sets are the same between the end of one episode and the beginning of the next one are the same. This may not be the case for CUCRL2.

Fortunately, when using the baseline policy during an episode, the confidence sets for every state-action pairs are either the same as the previous episode or are becoming tighter around the true parameters of the MDP $M^\star$. Thus, proving Lemma 9 is similar to the proof presented in Lazaric et al. (2019), the only difference resides in bounding the sum, $\sum_{k\in\Lambda_{k_T}}\sum_{t=t_k}^{t_{k+1}-1} 1/\sqrt{N_k^+(s_t,a_t)}$, which is bounded by the square root of the total number of samples in the proof of Lazaric et al. (2019) whereas in the case CUCRL2 it is bounded by the square root of the total number of samples gathered while exploring the set of policies plus the number of samples collected while playing the baseline policies. Therefore, at the end of the day both quantities are bounded by a constant times the square root of T.

A doubt someone could have is on controlling the term

$$\begin{aligned}
&\sum_{k=1}^{k_T} \mathbb{1}_{(k\in\Lambda_T)} \sum_{t=t_k}^{t_{k+1}-1} \left(p(\cdot|s_t,\pi_k(s_t))^\mathsf{T} u_k - u_k(s_t)\right) \\
&= \sum_{k=1}^{k_T} \mathbb{1}_{(k\in\Lambda_T)} \sum_{t=t_k}^{t_{k+1}-1} \underbrace{\left(p(\cdot|s_t,\pi_k(s_t))^\mathsf{T} u_k - u_k(s_{t+1})\right)}_{\Delta_k^p} \\
&\quad + \sum_{k=1}^{k_T} \mathbb{1}_{(k\in\Lambda_T)} \sum_{t=t_k}^{t_{k+1}-1} u_k(s_{t+1}) - u_k(s_t) \\
&= \sum_{k=1}^{k_T} \mathbb{1}_{(k\in\Lambda_T)} \Delta_k^p + \underbrace{\left(u_k(s_{t_{k+1}}) - u_k(s_{t_k})\right)}_{\leq sp(w_k)\leq D}
\end{aligned}$$

For any time t, the σ-algebra generated by the history up to time t included is $\mathcal{F}_t = \sigma(s_1, a_1, r_1, \ldots, s_t, a_t, r_t, s_{t+1})$. Define $X_t = \mathbb{1}_{(k_t\in\Lambda_T)}(p(\cdot|s_t,\pi_{k_t}(s_t))^\mathsf{T} u_k - u_k(s_{t+1}))$. Since π_{k_t} is $\mathcal{F}_t$ measurable, $E[X_t|\mathcal{F}_{t-1}] = 0$ and $|X_t| \leq 2D$. Then $(X_t, \mathcal{F}_t)_t$ is an MDS and nothing change compared to the analysis of UCRL2.

□

Finally, plugging Lemmas 8 and 9 into Lem. 7, we have that there exists a numerical constant C_1 such that with probability $1-\delta$:

$$\begin{aligned}
R(\text{CUCRL2}, T) \leq C_1 \Bigg(DS\sqrt{ATL_T^\delta} + (g^\star - g^{\pi_b}) \Bigg(\sqrt{SAT\ln(T)} + \frac{\sqrt{TSAL_T^\delta}}{\alpha g^{\pi_b}} \max\{sp(h^{\pi_b}), D + \Upsilon\} \\
+ \frac{S^2 AL_T^\delta}{(\alpha g^{\pi_b})^2}(D+\Upsilon)^2 \Bigg) + \max\{r_{\max}, sp\,(h^{\pi_b})\}\sqrt{SAT\ln(T/\delta)} \Bigg)
\end{aligned}$$

C Conservative Exploration in Finite Horizon Markov Decision Processes

In this section, we show how the conservative setting can be applied to finite horizon MDPs. Let's consider a finite-horizon MDP (Puterman, 1994, Chp. 4) $M = (\mathcal{S}, \mathcal{A}, p, r, H)$ with state space $\mathcal{S}$ and action space $\mathcal{A}$. Every state-action pair is characterized by a reward distribution with mean $r(s,a)$ and support in $[0,1]$ and a transition distribution $p(\cdot|s,a)$ over next state. We denote by $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$ the number of states and actions, and by H the horizon of an episode. A Markov randomized decision rule $d : \mathcal{S} \to P(\mathcal{A})$ maps states to distributions over actions. A policy π is a sequence of decision rules, i.e., $\pi = (d_1, d_2, \dots, d_H)$. We denote by Π^{MR} (resp. Π^{MD}) the set of Markov randomized (resp. deterministic) policies. The value of a policy $\pi \in \Pi^{\text{MR}}$ is measured trough the value function

$$\forall t \in [H], \forall s \in \mathcal{S} \qquad V_t^\pi(s) = \mathbb{E}^\pi \left[\sum_{l=t}^{H} r_l(s_l, a_l) \mid s_t = s \right]$$

where the expectation is defined w.r.t. the model and policy (i.e., $a_l \sim d_l(s_l)$). This function gives the expected total reward that one could get by following policy π starting in state s, at time t. There exists an optimal policy $\pi^\star \in \Pi^{\text{MD}}$ (Puterman, 1994, Sec. 4.4) for which $V_t^\star = V_t^{\pi^\star}$ satisfies the *optimality equations*:

$$\forall t \in [H], \forall s \in \mathcal{S}, \qquad V_t^\star(s) = \max_{a \in \mathcal{A}} \left\{ r_t(s,a) + p(\cdot|s,a)^\intercal V_{t+1}^\star \right\} := L_t^\star V_t^\star \tag{26}$$

where $V_{H+1}^\star(s) = 0$ for any state $s \in \mathcal{S}$. The value function can be computed using backward induction (e.g., Puterman, 1994; Bertsekas, 1995) when the reward and transitions are known. Given a policy $\pi \in \Pi^{\text{MD}}$, the associated value function satisfies the *evaluation equations* $V_t^\pi(s) := L_t^\pi V_{t+1}^\pi(s) = r(s, d_t(s)) + p(\cdot|s, d_t(s))^\intercal V_{t+1}^\pi$. The optimal policy is thus defined as $\pi^\star = \arg\max_{\pi \in \Pi^{\text{MD}}} \{L_t^\pi V_t^\star\}$, $\forall t \in [H]$.

In the following we assume that the learning agent known $\mathcal{S}$, $\mathcal{A}$ and r_{max}, while the reward and dynamics are *unknown* and need to be estimated online. Given a finite number of episode K, we evaluate the performance of a learning algorithm $\mathfrak{A}$ by its cumulative regret

$$R(\mathfrak{A}, K) = \sum_{k=1}^{K} V_1^\star(s_{k,1}) - V_1^{\pi_k}(s_{k,1})$$

where π_k is the policy executed by the algorithm at episode k.

Conservative Condition Designing a conservative condition, in this setting is much easier than in the average reward case as evaluating a policy can be done through the value function which gives an estimation of the expected reward over an episode. Thus, we can use this evaluation of a policy to use in place of rewards in the bandits condition. Formally, denote by $\pi_b \in \Pi^{\text{MR}}$ the baseline policy and assume that $V_t^{\pi_b}$ is known. In general, this assumption is not restrictive since the baseline performance can be estimated from historical data. Given a conservative level $\alpha \in (0,1)$, we define the conservative condition as:

$$\forall k \in [K], \quad \sum_{l=1}^{k} V_1^{\pi_l}(s_{l,1}) \geq (1-\alpha) \sum_{l=1}^{k} V_1^{\pi_b}(s_{l,1}) \qquad \text{w.h.p} \tag{27}$$

where π_l is the policy executed by the algorithm at episode l and $s_{l,1}$ is the starting state of episode l before policy π_l is chosen. The initial state can be chosen arbitrarily but should be revealed at the beginning of each episode. Note that this condition is random due the choice of the policies $(\pi_l)_l$ and also because of the starting states thus the condition is required to hold with high probability.

Note that Eq. 27 requires to evaluated the performance of policy π_l on the true (unknown) MDP. In order derive a practical condition, we need to construct an estimate of $V_1^{\pi_l}$. In order to be conservative, we are interesting in deriving a lower bound on the value function of a generic policy π which can be used in Eq. 27.

Pessimistic value function estimate. We recall that OFU algorithms (e.g., UCB-VI and EULER) builds uncertainties around the rewards and dynamics that are used to perform an optimistic planning. Formally, denote by $\widehat{p}_k(\cdot|s,a)$ and $\widehat{r}_k(s,a)$ the empirical transitions and rewards at episode k. Then, with high probability

$$|(p(\cdot|s,a) - \widehat{p}_k(\cdot|s,a))^\intercal v| \leq \beta_k^p(s,a) \quad \text{and} \quad |r(s,a) - \widehat{r}_k(s,a)| \leq \beta_k^r(s,a)$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $v \in [0,H]^S$. This uncertainties are used to compute an exploration bonus $b_k(s,a) = \beta_k^v(s,a) + \beta_k^r(s,a)$ that can be used to compute an optimistic estimate of the optimal value function. Formally, at episode k, optimistic backward induction (e.g., Azar et al., 2017, Alg. 2) computes an estimate value function $\bar{v}_{k,h}$ such that $\bar{v}_{k,h} \geq V_t^\star$ for any state s. The same approach can be used to compute a pessimistic estimate of the optimal value function by subtracting the exploration bonus to the reward (e.g., Zanette and Brunskill, 2019).

The only difference in the conservative setting is that we are interesting to compute a pessimistic estimate for a policy different from the optimal one. We thus define the *pessimistic evaluation equations* for any episode k, step h, state s and policy $\pi \in MR$ as:

$$\underline{v}_{k,h}^\pi(s) := \underline{L}_{k,h}^\pi \underline{v}_{k,h+1}^\pi = \sum_a \pi_{k,h}(s,a) \left(\widehat{r}_k(s,a) - b_k(s,a) + \widehat{p}_k(\cdot|s,a)^\mathsf{T} \underline{v}_{k,h+1}^\pi \right) \tag{28}$$

with $\underline{v}_{k,H+1}^\pi(s) = 0$ for all states $s \in \mathcal{S}$. This value function is pessimistic (see Lem. 10) and can be computed using backward induction with $\underline{L}_k^\pi$.

Lemma 10. *Let $\pi = (d_1, \dots, d_H) \in MR$ and $(\underline{v}_{k,h}^\pi)_{h\in[H]}$ be the value function given by backward induction using Eq. 28 then with high probability:*

$$\forall (h,s) \in [H] \times \mathcal{S}, \qquad V_h^\pi(s) \geq \underline{v}_{k,h}^\pi(s)$$

Proof. On the event that the concentration inequalities holds, let $\widehat{r}_k(s,a)$ be the empirical reward at episode k and $\widehat{p}_k(.|s,a)$ the empirical distribution over the next state from (s,a) at episode k. We proceed with a backward induction. At time H the statement is true. For $h < H$:

$$\begin{aligned}
\underline{v}_{k,h}^\pi(s) - V_h^\pi(s) &= \sum_a d_h(s,a) \left(\widehat{r}_k(s,a) - b_k(s,a) + \widehat{p}_k(\cdot|s,a)^\mathsf{T} \underline{v}_{k,h+1}^\pi \right) - L_h^\pi V_{h+1}^\pi(s) \\
&= \sum_a d_h(s,a) \left(\underbrace{\widehat{r}_k(s,a) - r(s,a) - \beta_k^r(s,a)}_{\leq 0} \right) \\
&\quad + \sum_a d_h(s,a) \left(\widehat{p}_k(\cdot|s,a)^\mathsf{T} \underline{v}_{k,h+1}^\pi - p(\cdot|s,a)^\mathsf{T} V_{h+1}^\pi - \beta_k^p(s,a) \right) \\
&\leq \sum_a d_h(s,a) \left(\widehat{p}_k(\cdot|s,a)^\mathsf{T} \underline{v}_{k,h+1}^\pi - p(\cdot|s,a)^\mathsf{T} V_{h+1}^\pi - \beta_k^p(s,a) \right) \\
&\leq \sum_a d_h(s,a) \left((\widehat{p}_k(\cdot|s,a) - p(\cdot|s,a))^\mathsf{T} V_{h+1}^\pi - \beta_k^p(s,a) \right) \leq 0
\end{aligned}$$

where the first inequality is true because of the confidence intervals on the reward function and the penultimate inequality is true because of the backward induction hypothesis. □

Thanks to this result, we can formulate a condition that the algorithm can check, at the beginning of episode k to decide if a policy is safe to play or not :

$$\sum_{l \in \mathcal{S}_{k-1} \cup \{k\}} \underline{v}_{l,1}^{\pi_l}(s_{l,1}) + \sum_{l \in \mathcal{S}_{k-1}^c} V_{l,1}^{\pi_b}(s_{l,1}) \geq (1-\alpha) \sum_{l=1}^k V_{l,1}^{\pi_b}(s_{l,1}) \tag{29}$$

where $\mathcal{S}_{k-1}$ is the set of episodes where the algorithm previously played non-conservatively, $\mathcal{S}_{k-1}^c = [k-1] \setminus \mathcal{S}_{k-1}$ is the set of episodes played conservatively and $(\pi_l)_l$ is the policies that the OFU algorithm (e.g., UCB-VI) would execute without the conservative constraint.

Alg. 4 shows the generic structure of any conservative exploration algorithm for MDPs. First, it computes an optimistic policy by leveraging on an OFU algorithm and the collected history. Then it checks the conservative condition. When Eq. 29 is verified it plays the optimistic policy otherwise it plays conservatively by executing policy π_b. This allows to build some budget for playing exploratory actions in the future.

Input: Policy π_b, $\delta \in (0,1)$, $r_{\max}$, $\mathcal{S}$, $\mathcal{A}$, $\alpha' \in (0,1)$, H
Initialization: Set $\mathcal{H} = \emptyset$, $\mathcal{S}_0 = \emptyset$ and $\mathcal{S}_0^c = \emptyset$
For episodes $k = 1, 2, \ldots$ **do**

1. Compute optimistic policy π_k using any OFU algorithm on history $\mathcal{H}$.
2. Compute pessimistic estimate $\underline{v}_k^{\pi_k}$ as in Eq. 28.
3. **if** Equation (29) not verified: **then**
   (a) $\pi_k = \pi_b$, $\mathcal{S}_{k+1}^c = \mathcal{S}_k^c \cup \{k\}$ and $\mathcal{S}_{k+1} = \mathcal{S}_k$
   else:
   (a) $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{k\}$ and $\mathcal{S}_{k+1}^c = \mathcal{S}_k^c$
4. **for** $h = 1, \ldots, H$ **do**
   (a) Execute $a_{k,h} = \pi_k(s_{k,h})$, obtain reward $r_{k,h}$, and observe $s_{k,h}$.
   (b) **if** $\pi_k \neq \pi_b$ **then:** add $(s_{k,h}, a_{k,h}, r_{k,h}, s_{k,h+1})$ to $\mathcal{H}$

Figure 4: CUCB-VI algorithm.

Regret Guarantees We analyse Alg. 4 with UCB-VI. Before to introduce the upper-bound to the regret of CUCB-VI we introduce the following assumption on the baseline policy.

Assumption 3. *The baseline policy $\pi_b \in \Pi^{MR}$ is such that $r_b := \min_s\{V_1^{\pi_b}(s)\} > 0$.*

We can now state the main results:

Proposition 1. *For $\delta > 0$, the regret of conservative* UCB-VI *(*CUCB-VI*) is upper-bounded with probability at least $1-\delta$ by:*

$$R(\textit{CUCB-VI}, K) \leq 2H\sqrt{SAHKL_K} + 8HSAL_K^2 + H\sqrt{2KL_K} + 2HS\sqrt{AHKL_K} + \frac{1}{4\alpha r_b(\underline{\Delta}_b + \alpha r_b)}\left(16H^3L_K + \left(200H^5S^2A + 128H^5SA\right)L_K^2\right), \tag{30}$$

where $L_K = \max\{\ln(3KHSA/\delta), 1\}$ and $\underline{\Delta}_b = \min_{s\in\mathcal{S}}\{V_1^\star(s) - V_1^{\pi_b}(s)\}$.

Proof. Let's define the high probability event, $\mathcal{E}$, that is such that in this event, all the concentration inequalities holds and the Martingale Difference Sequence concentration inequalities also holds :

$$\mathcal{E}_{1,\delta} := \bigcap_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bigcap_{k\in[K]} \left\{ ||p(.|s,a) - \hat{p}_k(.|s,a)||_1 \leq \sqrt{\frac{2S\ln(3KSA/\delta)}{\max\{1, N_k(s,a)\}}} \right\} \bigcap \left\{ |\hat{r}_k(s,a) - r(s,a)| \leq 2r_{\max}\sqrt{\frac{\ln(3KSA/\delta)}{\max\{1, N_k(s,a)\}}} \right\}$$

$$\mathcal{E}_{2,\delta} := \bigcap_{k\in[K]} \left\{ \sum_{l\in S_k}\sum_{h=1}^{H} \varepsilon_{k,h} \leq H^{3/2}\sqrt{2\#S_k \ln(3KH/\delta)} \right\}$$

and finally, $\mathcal{E} := \mathcal{E}_{1,\delta} \cap \mathcal{E}_{2,\delta}$, then $\mathcal{E}$ holds with probability at least $1-\delta$. Indeed,

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{t=1}^{HK} \frac{\delta}{3HK} + \sum_{s,a}\sum_{k} \frac{2\delta}{3KSA} \leq \delta$$

Under this event, we have that for all episode $k \in S_K$:

$$\overline{v}_{k,1}(s_{k,1}) - \underline{v}_{k,1}^{\pi_k}(s_{k,1}) \leq \sum_{h=1}^{H} \varepsilon_{k,h} + 5\beta_k^p(s_{k,h}, d_h^k(s_{k,h})) + 2\beta_k^r(s_{k,h}, d_h^k(s_{k,h})),$$

where $(\varepsilon_{k,h})_{k\in\mathcal{S}_K,h\in[H]}$ is a martingale difference sequence with respect to the filtration $(\mathcal{F}_{k,h})_{k\in\mathcal{S}_K,h\in[H]}$ that is generated by all the randomness before step h of episode k. Indeed, for an episode k, let $\pi_k = (d_1^k, \dots, d_H^k)$, decomposing π_k into successive decision rules.

$$\overline{v}_{k,1}(s_{k,1}) - \underline{v}^{\pi_k}_{k,2} \leq 2\beta_k^r(s_{k,1}, d_1^k(s_{k,1})) + \hat{p}_k(. \mid s_{k,1}, d_1^k(s_{k,1}))^\intercal(\overline{v}_{k,2} - \underline{v}^{\pi_k}_{k,2}) + 2\beta_k^p(s_{k,1}, d_1^k(s_{k,1}))$$

Thus by defining, $B_{k,h} := 3\beta_k^p(s_{k,h}, d_h^k(s_{k,h})) + 2\beta_k^r(s_{k,h}, d_h^k(s_{k,h}))$, we have :

$$\begin{aligned}
\overline{v}_{k,1}(s_{k,1}) - \underline{v}^{\pi_k}_{k,1}(s_{k,1}) \leq & B_{k,1} + (\hat{p}_k(. \mid s_{k,1}, d_1^k(s_{k,1}) - p(.|s_{k,1}, d_1^k(s_{k,1}))^\intercal(\overline{v}_{k,2} - \underline{v}^{\pi_k}_{k,2}) + (\overline{v}_{k,2}(s_{k,2}) - \underline{v}^{\pi_k}_{k,2}(s_{k,2})) \\
& - (\overline{v}_{k,2}(s_{k,2}) - \underline{v}^{\pi_k}_{k,2}(s_{k,2})) + p(.|s_{k,1}, d_1^k(s_{k,1})^\intercal(\overline{v}_{k,2} - \underline{v}^{\pi_k}_{k,2}) \\
& \leq p(.|s_{k,1}, d_1^k(s_{k,1})^\intercal(\overline{v}_{k,2} - \underline{v}^{\pi_k}_{k,h}) - (\overline{v}_{k,2}(s_{k,2}) - \underline{v}^{\pi_k}_{k,2}(s_{k,2})) + 2\beta_k^p(s_{k,1}, d_1^k(s_{k,1})) + B_{k,1} \\
& + (\overline{v}_{k,2}(s_{k,2}) - \underline{v}^{\pi_k}_{k,2}(s_{k,2}))
\end{aligned}$$

But let's define $\varepsilon_{k,h} := p(.|s_{k,h}, d_h^k(s_{k,h}))^\intercal(\overline{v}_{k,h} - \underline{v}^{\pi_k}_{k,h}) - (\overline{v}_{k,h}(s_{k,h+1}) - \underline{v}^{\pi_k}_{k,h}(s_{k,h+1}))$ then $(\varepsilon_{k,h})_{k\in[K],h\in[H]}$ is a Martingale Difference Sequence with respect to the filtration $\mathcal{F}_{k,h}$ which is generated by all the randomness in the environment and the algorithm before step h of episode k . Then, by recursion, we have :

$$\overline{v}_{k,1}(s_{k,1}) - \underline{v}^{\pi_k}_{k,1}(s_{k,1}) \leq \sum_{h=1}^{H} B_{k,h} + \varepsilon_{k,h} + 2\beta_k^p(s_{k,h}, d_h^k(s_{k,h}))$$

The regret of algorithm CUCB-VI can be decomposed as :

$$\begin{aligned}
R(\text{CUCB-VI}, K) &= \sum_{k\in\mathcal{S}_K^c} V_1^\star(s_{k,1}) - V_1^{\pi_b}(s_{k,1}) + \sum_{k\in\mathcal{S}_K} V_1^\star(s_{k,1}) - V_1^{\pi_k}(s_{k,1}) \\
&\leq |\mathcal{S}_K^c|\overline{\Delta}_b + R(\text{UCB-VI}, |\mathcal{S}_K|)
\end{aligned}$$

where $\overline{\Delta}_b = \max_{s\in\mathcal{S}} V^\star(s) - V^{\pi_b}(s)$. Therefore bounding the regret amounts to bound the number of episode played conservatively. To do so, let's consider, τ the last episode played conservatively, then before the beginning of episode τ, the condition 29 is not verified and thus :

$$\alpha \sum_{k=1}^{\tau} V_1^{\pi_b}(s_{k,1}) \leq \sum_{k\in\mathcal{S}_{\tau-1}\cup\{\tau\}} \underbrace{V_1^{\pi_b}(s_{k,1}) - \underline{v}^{\pi_k}_{k,1}(s_{k,1})}_{=\Delta_{k,1}}$$

Thus, let's finish this analysis by bounding $\Delta_{k,1} = V_1^{\pi_b}(s_{k,1}) - \underline{v}^{\pi_k}_{k,1}(s_{k,1})$ for all $k \in \mathcal{S}_K$. But:

$$\Delta_{k,1} = V_1^{\pi_b}(s_{k,1}) - V_1^\star(s_{k,1}) + V_1^\star(s_{k,1}) - \underline{v}^{\pi_k}_{k,1}(s_{k,1}) \leq -\underline{\Delta}_b + \overline{v}_{k,1}(s_{k,1}) - \underline{v}^{\pi_k}_{k,1}(s_{k,1}),$$

where $\underline{\Delta}_b := \min_s V_1^\star(s) - V_1^{\pi_b}(s)$. Now, we need to bound the sum over all the non-conservative episodes of the difference between the optimistic and pessimistic value function. That is to say :

$$\begin{aligned}
\sum_{l\in S_{\tau-1}} \sum_{h=1}^{H} \beta_k^r(s_{k,h}, d_h^k(s_{k,h})) &= \sum_{l\in\mathcal{S}_{\tau-1}} \sum_{h=1}^{H} 2Hr_{\max}\sqrt{\frac{2\ln(3KSA/\delta)}{\max\{1, N_k(s_{k,h}, d_h^k(s_{k,h}))}} \\
&\leq 2r_{\max}H^2\sqrt{2SAH|\mathcal{S}_{\tau-1}|(1+\ln(|\mathcal{S}_{\tau-1}|H))\ln(3KSA/\delta)}
\end{aligned}$$

Also :

$$\begin{aligned}
\sum_{l\in S_{\tau-1}} \sum_{h=1}^{H} \beta_k^p(s_{k,h}, d_h^k(s_{k,h})) &= \sum_{l\in\mathcal{S}_{\tau-1}} \sum_{h=1}^{H} H\sqrt{\frac{2S\ln(3KSA/\delta)}{\max\{1, N_k(s_{k,h}, d_h^k(s_{k,h}))}} \\
&\leq H^2S\sqrt{2AH\#S_{\tau-1}(1+\ln(\#S_{\tau-1}H))\ln(3KSA/\delta)}
\end{aligned}$$

and, under the event $\mathcal{E}$, $\sum_{l\in\mathcal{S}_{\tau-1}}\sum_{h=1}^{H}\varepsilon_{k,h} \leq 2H^{3/2}\sqrt{2|\mathcal{S}_{\tau-1}|\ln(3KH/\delta)}$. On the other hand, for the episode τ, we can only bound the difference in value function by H. Finally, we have that $\tau = 1 + |\mathcal{S}_{\tau-1}^c| + |\mathcal{S}_{\tau-1}|$ and

thus if we assume that $r_b := \min_s V^{\pi_b}(s) > 0$:

$$\alpha r_b(|\mathcal{S}^c_{\tau-1}|+1) \leq \alpha \sum_{k=1}^{\tau} V_1^{\pi_b}(s_{k,1}) \leq -(\underline{\Delta}_b + \alpha r_b)|\mathcal{S}_{\tau-1}| + 2H^{3/2}\sqrt{2|\mathcal{S}_{\tau-1}|\ln(3KH/\delta)}$$
$$+5H^2 S\sqrt{2AH|\mathcal{S}_{\tau-1}|(1+\mathcal{S}_{\tau-1}|H))\ln(3KSA/\delta)}$$
$$+4r_{\max}H^2\sqrt{2SAH|\mathcal{S}_{\tau-1}|(1+\ln(|\mathcal{S}_{\tau-1}|H))\ln(3KSA/\delta)}$$

Thus, the function on the RHS in bounded and using lemma 8 of Kazerouni et al. (2017), we have :

$$\alpha r_b(|\mathcal{S}^c_{\tau-1}|+1) \leq \frac{1}{4(\underline{\Delta}_b+\alpha r_b)}\left(16H^3\ln\left(\frac{3KH}{\delta}\right) + \left(200H^5S^2A + 128r_{\max}^2H^5SA\right)\times\right.$$
$$\left.\times(1+\ln(HK))\ln\left(\frac{3KSA}{\delta}\right)\right)$$

But by definition, $|\mathcal{S}^c_{\tau-1}|+1 = |\mathcal{S}^c_K|$. Hence the result. □

Experiments Finally, we end this presentation of conservativeness in finite horizon MDPs with some experiments. We consider a classic 3×4 gridworld problem with one goal state, a starting state and one trap state, we set $H = 10$, and the reward of any action in all the state to -2, the reward in the goal state to 10 and the reward of falling in the trapping state to -20. We normalize the rewards to be in $[0, 1]$. The baseline policy is describing a path around the pit, see Fig 5. On the two position adjacent to the goal the baseline policy is

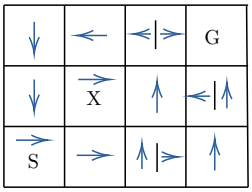

Figure 5: Illustration of the baseline policy. S is the starting state, X is the pit a,d G is the goal state.

stochastic with a probability of reaching the goal of 1/2 for the position on the right of the goal and below the goal, respectively. On the last line the probability of going up or right is also uniform. Figure 6 shows the impact of the conservative constraint on the regret of UCB-VI for a conservative coefficient $\alpha = 0.05$. Fig 6 also shows the constraint as a function of the time for UCB-VI and CUCB-VI that is to say: $\sum_{l=1}^{t} V^{\pi_l}(s_0) - (1-\alpha)V^{\pi_b}(s_0)$ as a function of episode t with s_0 the starting state of the gridworld. In the first 10% episodes (i.e until episode 300) the condition was violated by UCB-VI 83% of the time.

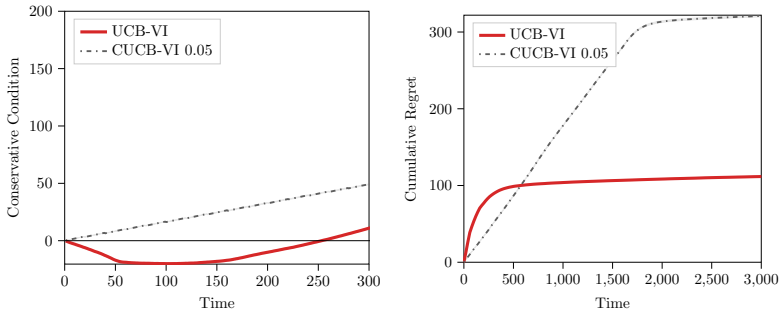

Figure 6: Regret and Conservative Condition for the gridworld problem

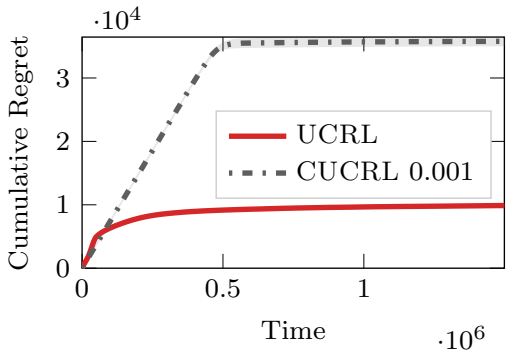

Figure 7: Regret of UCRL2 and CUCRL2 on the Cost-Based Maintenance problem described in D

D Experiments

For average reward problems we consider "simplified" Bernstein confidence intervals given by:

$$\beta_r^k(s,a) = \sigma_r(s,a)\sqrt{\frac{\ln(SA/\delta)}{N_k^+(s,a)}} + r_{\max}\frac{\ln(SA/\delta)}{N_k^+(s,a)} \quad \text{and} \quad \beta_p^k(s,a,s') = \sigma_p(s,a,s')\sqrt{\frac{\ln(SA/\delta)}{N_k^+(s,a)}} + \frac{\ln(SA/\delta)}{N_k^+(s,a)}$$

where $N_k^+(s,a) = \max\{1, N_k(s,a)\}$, $\sigma_r(s,a)$ is the empirical standard deviation and $\sigma_p(s,a,s') = \sqrt{\widehat{p}(s'|s,a)(1-\widehat{p}(s'|s,a))}$.

D.1 Single-Product Stochastic Inventory Control

Maintaining inventories is necessary for any company dealing with physical products. We consider the case of single product without backlogging. The state space is the amount of products in the inventory, $\mathcal{S} = \{0, \dots, M\}$ where M is the maximum capacity. Given the state s_t at the beginning of the month, the manager (agent) has to decide the amount of units a_t to order. We define D_t to be the random demand of month t and we assume a time-homogeneous probability distribution for the demand. The inventory at time $t+1$ is given by

$$s_{t+1} = \max\{0, s_t + a_t - D_t\}$$

The action space is $\mathcal{A}_s = \{0, \dots, M - s\}$. As in (Puterman, 1994), we assume a fixed cost $K > 0$ for placing orders and a varible cost $c(a)$ that increases with the quantity ordered: $O(a) = \begin{cases} K + c(a) & a > 0 \\ 0 & \text{otherwise} \end{cases}$. The cost of maintaining an inventory of s items is defined by the nondecreasing function $h(s)$. If the inventory is available to meet a demand j, the agent receives a revenue of $f(j)$. The reward is thus defined as $r(s_t, a_t, s_{t+1}) = -O(a_t) - h(s_t + a_t) + f(s_t + a_t - s_{t+1})$. In the experiments, we use $K = 4$, $c(x) = 2x$, $h(x) = x$ and $f(x) = 8x$.

In all the experiments, we normalize rewards such that the support is in $[0, 1]$ and we use noise proportional to the reward mean: $r_t(s,a) = (1 + c\eta_t)r(s,a)$ where $\eta_t \sim \mathcal{N}(0,1)$ (we set $c = 0.1$).

D.2 Cost-Based Maintenance

The system is composed by N components in an active redundant, parallel setting, which are subject to economic and stochastic dependence through load sharing. Each component $j \in [N]$ is described by its operational level $x_j = \{0, \dots, L\}$. The level L denotes that the component has failed. The deterioration process is modelled using a Poisson process. If all components have failed, the system is shut down and a penalty cost p is paid. The replacement of a failed component cost c_c, while the same operation on an active component cost c_p (usually $c_c \geq c_p$). There is also a fixed cost for maintenance c_s. At each time step, it is possible to replace simultaneously multiple components. Please refer to (Olde Keizer, 2016) for a complete description of dynamics and rewards.

We terminate the analysis of CUCRL2 with a more challenging test. We consider the condition-based maintenance problem (CBM, Olde Keizer, 2016) a multi-component system subject to structural, economic and

stochastic dependences. We report a complete description of the problem in App. D. The resulting MDP has $S = 121$ states and $A = 4$ actions. The maintenance policy is often implemented as a threshold policy based on the deterioration level. Such a threshold policy is not necessarily optimal for a system with economic dependence and redundancy. We simulate this scenario by considering a strong (almost optimal) threshold policy for CBM without economic dependence as baseline. We make it stochastic by selecting with probability 0.3 a random action. As a result we have that the optimal gain $g^\star = 0.89$ while the baseline gain is $g^{\pi_b} = 0.82$. Fig. 7 shows the cumulative regret for UCRL2 and CUCRL2 with $\alpha = 0.001$. UCRL2 explores faster than CUCRL2 but violates the conservative condition 53% of times in the initial phase (up to $t = 140000$), incurring in multiple complete system failures. On the other hand, CUCRL2 never violates the conservative condition.