Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, Ronald Ortner

▶ **To cite this version:**

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, Ronald Ortner. Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning. ICML 2018 - The 35th International Conference on Machine Learning, Jul 2018, Stockholm, Sweden. pp.1578-1586. hal-01941206

HAL Id: hal-01941206
https://inria.hal.science/hal-01941206

Submitted on 30 Nov 2018

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

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

Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning

Ronan Fruit [*1] **Matteo Pirotta** [*1] **Alessandro Lazaric** [2] **Ronald Ortner** [3]

Abstract

We introduce SCAL, an algorithm designed to perform efficient exploration-exploitation in any *unknown weakly-communicating* Markov decision process (MDP) for which an upper bound c on the span of the *optimal bias function* is known. For an MDP with S states, A actions and $\Gamma \leq S$ possible next states, we prove a regret bound of $\widetilde{\mathcal{O}}(c\sqrt{\Gamma SAT})$, which significantly improves over existing algorithms (e.g., UCRL and PSRL), whose regret scales linearly with the MDP *diameter* D. In fact, the optimal bias span is finite and often much smaller than D (e.g., $D = \infty$ in non-communicating MDPs). A similar result was originally derived by Bartlett and Tewari (2009) for REGAL.C, for which no tractable algorithm is available. In this paper, we *relax* the optimization problem at the core of REGAL.C, we carefully analyze its properties, and we provide the first *computationally efficient algorithm* to solve it. Finally, we report numerical simulations supporting our theoretical findings and showing how SCAL significantly outperforms UCRL in MDPs with *large* diameter and *small* span.

1. Introduction

While learning in an unknown environment, a reinforcement learning (RL) agent must trade off the *exploration* needed to collect information about the dynamics and reward, and the *exploitation* of the experience gathered so far to gain as much reward as possible. In this paper, we focus on the regret framework (Jaksch et al., 2010), which evaluates the exploration-exploitation performance by comparing the rewards accumulated by the agent and an optimal policy. A common approach to the exploration-exploitation dilemma is the *optimism in face of uncertainty* (OFU) principle: the agent maintains optimistic estimates of the value function and, at each step, it executes the policy with highest optimistic value (e.g., Brafman and Tennenholtz, 2002; Jaksch et al., 2010; Bartlett and Tewari, 2009). An alternative approach is posterior sampling (Thompson, 1933), which maintains a Bayesian distribution over MDPs (i.e., dynamics and expected reward) and, at each step, samples an MDP and executes the corresponding optimal policy (e.g., Osband et al., 2013; Abbasi-Yadkori and Szepesvári, 2015; Osband and Roy, 2017; Ouyang et al., 2017; Agrawal and Jia, 2017).

Given a finite MDP with S states, A actions, and diameter D (i.e., the time needed to connect any two states), Jaksch et al. (2010) proved that no algorithm can achieve regret smaller than $\Omega(\sqrt{DSAT})$. While recent work successfully closed the gap between upper and lower bounds w.r.t. the dependency on the number of states (e.g., Agrawal and Jia, 2017; Azar et al., 2017), relatively little attention has been devoted to the dependency on D. While the diameter quantifies the number of steps needed to "recover" from a bad state in the worst case, the actual regret incurred while "recovering" is related to the difference in potential reward between "bad" and "good" states, which is accurately measured by the span (i.e., the range) $sp\{h^*\}$ of the optimal bias function h^*. While the diameter is an upper bound on the bias span, it could be arbitrarily larger (e.g., weakly-communicating MDPs may have finite span and infinite diameter) thus suggesting that algorithms whose regret scales with the span may perform significantly better.[1] Building on the idea that the OFU principle should be *mitigated* by the bias span of the optimistic solution, Bartlett and Tewari (2009) proposed three different algorithms (referred to as REGAL) achieving regret scaling with $sp\{h^*\}$ instead of D. The first algorithm defines a span regularized problem, where the regularization constant needs to be carefully tuned depending on the state-action pairs visited in the future, which makes it unfeasible in practice. Alternatively, they propose a constrained variant, called REGAL.C, where the regularized problem is replaced by a constraint on the span. Assuming that an upper-bound c on the bias span of the optimal policy is

[*]Equal contribution [1]SequeL Team, INRIA Lille, France [2]Facebook AI Research, Paris, France [3]Montanuniversität Leoben, Austria. Correspondence to: Ronan Fruit <ronan.fruit@inria.fr>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

[1]The proof of the lower bound relies on the construction of an MDP whose diameter actually coincides with the bias span (up to a multiplicative numerical constant), thus leaving the open question whether the "actual" lower bound depends on D or the bias span. See (Osband and Roy, 2016) for a more thorough discussion.

known (i.e., $sp\{h^*\} \le c$), REGAL.C achieves regret upper-bounded by $\widetilde{\mathcal{O}}(\min\{D,c\}S\sqrt{AT})$. Unfortunately, they do not propose any computationally tractable algorithm solving the constrained optimization problem, which may even be ill-posed in some cases. Finally, REGAL.D avoids the need of knowing the *future* visits by using a doubling trick, but still requires solving a regularized problem, for which no computationally tractable algorithm is known.

In this paper, we build on REGAL.C and propose a constrained optimization problem for which we derive a computationally efficient algorithm, called SCOPT. We identify conditions under which SCOPT converges to the optimal solution and propose a suitable stopping criterion to achieve an ε-optimal policy. Finally, we show that using a slightly modified optimistic argument, the convergence conditions are always satisfied and the learning algorithm obtained by integrating SCOPT into a UCRL-like scheme (resulting into SCAL) achieves regret scaling as $\widetilde{\mathcal{O}}(\min\{D,c\}\sqrt{\Gamma SAT})$ when an upper-bound c on the optimal bias span is available, thus providing the first computationally tractable algorithm that can solve weakly-communicating MDPs.

2. Preliminaries

We consider a finite *weakly-communicating* Markov decision process (Puterman, 1994, Sec. 8.3) $M = \langle \mathcal{S}, \mathcal{A}, r, p\rangle$ with a set of states $\mathcal{S}$ and a set of actions $\mathcal{A} = \bigcup_{s\in\mathcal{S}} \mathcal{A}_s$. Each state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}_s$ is characterized by a reward distribution with mean $r(s,a)$ and support in $[0, r_{\max}]$ as well as a transition probability distribution $p(\cdot|s,a)$ over next states. We denote by $S = |\mathcal{S}|$ and $A = \max_{s\in\mathcal{S}} |\mathcal{A}_s|$ the number of states and actions, and by Γ the maximum support of all transition probabilities. A Markov randomized *decision rule* $d : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ maps states to distributions over actions. The corresponding set is denoted by D^{MR}, while the subset of Markov deterministic decision rules is D^{MD}. A stationary *policy* $\pi = (d, d, \dots) =: d^\infty$ repeatedly applies the same decision rule d over time. The set of stationary policies defined by Markov randomized (resp. deterministic) decision rules is denoted by $\Pi^{\text{SR}}(M)$ (resp. $\Pi^{\text{SD}}(M)$). The *long-term average reward* (or *gain*) of a policy $\pi \in \Pi^{\text{SR}}(M)$ starting from $s \in \mathcal{S}$ is

$$g^\pi_M(s) := \lim_{T\to+\infty} \mathbb{E}_{\mathbb{Q}}\left[\frac{1}{T}\sum_{t=1}^{T} r(s_t,a_t)\right],$$

where $\mathbb{Q} := \mathbb{P}(\cdot|a_t \sim \pi(s_t); s_0 = s; M)$. Any stationary policy $\pi \in \Pi^{\text{SR}}$ has an associated bias function defined as

$$h^\pi_M(s) := \underset{T\to+\infty}{C\text{-}\lim}\, \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^{T} \big(r(s_t,a_t) - g^\pi_M(s_t)\big)\right],$$

that measures the expected total difference between the reward and the stationary reward in *Cesaro-limit*[2] (denoted $C\text{-}\lim$). Accordingly, the difference of bias values $h^\pi_M(s) - h^\pi_M(s')$ quantifies the (dis-)advantage of starting in state s rather than s'. In the following, we drop the dependency on M whenever clear from the context and denote by $sp\{h^\pi\} := \max_s h^\pi(s) - \min_s h^\pi(s)$ the *span* of the bias function. In weakly communicating MDPs, any optimal policy $\pi^* \in \arg\max_\pi g^\pi(s)$ has *constant* gain, i.e., $g^{\pi^*}(s) = g^*$ for all $s \in \mathcal{S}$. Let $P_d \in \mathbb{R}^{S\times S}$ and $r_d \in \mathbb{R}^S$ be the transition matrix and reward vector associated with decision rule $d \in D^{\text{MR}}$. We denote by L_d and L the Bellman operator associated with d and *optimal* Bellman operator

$$\forall v \in \mathbb{R}^S,\ \ L_d v := r_d + P_d v; \quad Lv := \max_{d\in D^{\text{MR}}} \{r_d + P_d v\}.$$

For any policy $\pi = d^\infty \in \Pi^{SR}$, the gain g^π and bias h^π satisfy the following system of *evaluation equations*

$$g^\pi = P_d g^\pi; \quad h^\pi = L_d h^\pi - g^\pi. \tag{1}$$

Moreover, there exists a policy $\pi^* \in \arg\max_\pi g^\pi(s)$ for which $(g^*, h^*) = (g^{\pi^*}, h^{\pi^*})$ satisfy the *optimality equation*

$$h^* = Lh^* - g^* e, \quad \text{where} \quad e = (1,\dots,1)^\intercal. \tag{2}$$

Finally, we denote by $D := \max_{(s,s')\in\mathcal{S}\times\mathcal{S}, s\neq s'}\{\tau_M(s \to s')\}$ the diameter of M, where $\tau_M(s\to s')$ is the minimal expected number of steps needed to reach s' from s in M.

Learning problem. Let M^* be the true *unknown* MDP. We consider the learning problem where $\mathcal{S}$, $\mathcal{A}$ and $r_{\max}$ are *known*, while rewards r and transition probabilities p are *unknown* and need to be estimated on-line. We evaluate the performance of a learning algorithm $\mathfrak{A}$ after T time steps by its cumulative *regret* $\Delta(\mathfrak{A}, T) = Tg^* - \sum_{t=1}^T r_t(s_t,a_t)$.

3. Optimistic Exploration-Exploitation

Since our proposed algorithm SCAL (Sec. 6) is a tractable variant of REGAL.C and thus a modification of UCRL, we first recall their common structure summarized in Fig. 1.

3.1. Upper-Confidence Reinforcement Learning

UCRL proceeds through episodes $k = 1, 2\dots$ At the beginning of each episode k, UCRL computes a set of plausible MDPs defined as $\mathcal{M}_k = \big\{M = \langle \mathcal{S}, \mathcal{A}, \widetilde{r}, \widetilde{p}\rangle : \widetilde{r}(s,a) \in B^k_r(s,a),\ \widetilde{p}(s'|s,a) \in B^k_p(s,a,s'), \sum_{s'}\widetilde{p}(s'|s,a) = 1\big\}$, where B^k_r and B^k_p are high-probability confidence intervals on the rewards and transition probabilities of the true MDP M^*, which guarantees that $M^* \in \mathcal{M}_k$ w.h.p. We use confidence intervals constructed using empirical Bernstein's inequality (Audibert et al., 2007; Maurer and Pontil, 2009)

$$\beta^{sa}_{r,k} := \sqrt{\frac{14\widehat{\sigma}^2_{r,k}(s,a) b_{k,\delta}}{\max\{1, N_k(s,a)\}}} + \frac{\frac{49}{3} r_{\max} b_{k,\delta}}{\max\{1, N_k(s,a)-1\}},$$

$$\beta^{sas'}_{p,k} := \sqrt{\frac{14\widehat{\sigma}^2_{p,k}(s'|s,a) b_{k,\delta}}{\max\{1, N_k(s,a)\}}} + \frac{\frac{49}{3} b_{k,\delta}}{\max\{1, N_k(s,a)-1\}},$$

[2] For policies with an aperiodic chain, the standard limit exists.

where $N_k(s,a)$ is the number of visits in (s,a) before episode k, $\widehat{\sigma}^2_{r,k}(s,a)$ and $\widehat{\sigma}^2_{p,k}(s'|s,a)$ are the empirical variances of $r(s,a)$ and $p(s'|s,a)$ and $b_{k,\delta} = \ln(2SAt_k/\delta)$. Given the empirical averages $\widehat{r}_k(s,a)$ and $\widehat{p}_k(s'|s,a)$ of rewards and transitions, we define $\mathcal{M}_k$ by $B^k_r(s,a) := [\widehat{r}_k(s,a) - \beta^{sa}_{r,k}, \widehat{r}_k(s,a) + \beta^{sa}_{r,k}] \cap [0, r_{\max}]$ and $B^k_p(s,a,s') := [\widehat{p}_k(s'|s,a) - \beta^{sas'}_{p,k}, \widehat{p}_k(s'|s,a) + \beta^{sas'}_{p,k}] \cap [0,1]$.

Once $\mathcal{M}_k$ has been computed, UCRL finds an approximate solution $(\widetilde{M}^*_k, \widetilde{\pi}^*_k)$ to the optimization problem

$$(\widetilde{M}^*_k, \widetilde{\pi}^*_k) \in \underset{M \in \mathcal{M}_k, \pi \in \Pi^{\text{SD}}(M)}{\arg\max}\, g^\pi_M. \tag{3}$$

Since $M^* \in \mathcal{M}_k$ w.h.p., it holds that $g^*_{\widetilde{M}_k} \geq g^*_{M^*}$. As noticed by Jaksch et al. (2010), problem (3) is equivalent to finding $\widetilde{\mu}^* \in \arg\max_{\mu \in \Pi^{\text{SD}}(\widetilde{\mathcal{M}}_k)} \{g^\mu_{\widetilde{\mathcal{M}}_k}\}$ where $\widetilde{\mathcal{M}}_k$ is the *extended* MDP (sometimes called *bounded-parameter* MDP) implicitly defined by $\mathcal{M}_k$. More precisely, in $\widetilde{\mathcal{M}}_k$ the (finite) action space $\mathcal{A}$ is "extended" to a compact action space $\widetilde{\mathcal{A}}_k$ by considering every possible value of the confidence intervals $B^k_r(s,a)$ and $B^k_p(s,a,s')$ as fictitious actions. The equivalence between the two problems comes from the fact that for each $\widetilde{\mu} \in \Pi^{\text{SD}}(\widetilde{\mathcal{M}}_k)$ there exists a pair $(\widetilde{M}, \widetilde{\pi})$ such that the policies $\widetilde{\pi}$ and $\widetilde{\mu}$ induce the same Markov reward process on respectively $\widetilde{M}$ and $\widetilde{\mathcal{M}}_k$, and conversely. Consequently, (3) can be solved by running so-called *extended* value iteration (EVI): starting from an initial vector $u_0 = 0$, EVI recursively computes

$$u_{n+1}(s) = \max_{a, \widetilde{r}, \widetilde{p}} \left[\widetilde{r}(s,a) + \widetilde{p}(\cdot|s,a)^\mathsf{T} u_n\right] = \widetilde{L} u_n(s), \tag{4}$$

where $\widetilde{L}$ is the *optimistic* optimal Bellman operator associated to $\widetilde{\mathcal{M}}_k$. If EVI is stopped when $sp\{u_{n+1} - u_n\} \leq \varepsilon_k$, then the greedy policy $\widetilde{\mu}_k$ w.r.t. u_n is guaranteed to be ε_k-optimal, i.e., $g^{\widetilde{\mu}_k}_{\widetilde{\mathcal{M}}_k} \geq g^*_{\widetilde{\mathcal{M}}_k} - \varepsilon_k \geq g^*_{M^*} - \varepsilon_k$. Therefore, the policy $\widetilde{\pi}_k$ associated to $\widetilde{\mu}_k$ is an *optimistic* ε_k-*optimal* policy, and UCRL executes $\widetilde{\pi}_k$ until the end of episode k.

3.2. A first relaxation of REGAL.C

REGAL.C follows the same steps as UCRL but instead of solving problem (3), it tries to find the best *optimistic* model $\widetilde{M}^*_{\text{RC}} \in \mathcal{M}_{\text{RC}}$ having constrained *optimal* bias span i.e.,

$$(\widetilde{M}^*_{\text{RC}}, \widetilde{\pi}^*_{\text{RC}}) = \underset{M \in \mathcal{M}_{\text{RC}}, \pi \in \Pi^{\text{SD}}(M)}{\arg\max}\, g^\pi_M, \tag{5}$$

where $\mathcal{M}_{\text{RC}} := \{M \in \mathcal{M}_k : sp\{h^*_M\} \leq c\}$ is the set of plausible MDPs with bias span of the *optimal* policy bounded by c. Under the assumption that $sp\{h^*_{M^*}\} \leq c$, REGAL.C discards any MDP $M \in \mathcal{M}_k$ whose *optimal* policy has a span larger than c (i.e., $sp\{h^*_M\} > c$) and otherwise looks for the MDP with highest *optimal* gain $g^*(M)$. Unfortunately, there is no guarantee that all MDPs in $\mathcal{M}_{\text{RC}}$ are weakly communicating and thus have constant gain. As a result, we suspect this problem to be ill-posed (i.e., the maximum is most likely not well-defined). Moreover, even if it is well-posed, searching the space $\mathcal{M}_{\text{RC}}$ seems to be computationally intractable. Finally, for any $M \in \mathcal{M}_k$, there may be several optimal policies with different bias spans and some of them may not satisfy the optimality equation (2) and are thus difficult to compute.

Input: Confidence $\delta \in]0,1[$, $r_{\max}$, $\mathcal{S}$, $\mathcal{A}$, a constant $c \geq 0$
For episodes $k = 1, 2, \ldots$ **do**

1. Set $t_k = t$ and episode counters $\nu_k(s,a) = 0$.
2. Compute estimates $\widehat{p}_k(s'|s,a)$, $\widehat{r}_k(s,a)$ and a confidence set $\mathcal{M}_k$ (UCRL, REGAL.C), resp. $\mathcal{M}^\ddagger_k$ (SCAL).
3. Compute an $r_{\max}/\sqrt{t_k}$-approximation $\widetilde{\pi}_k$ of the solution of Eq. 3 (UCRL), resp. Eq. 5 (REGAL.C), resp. Eq. 15 (SCAL).
4. Sample action $a_t \sim \widetilde{\pi}_k(\cdot|s_t)$.
5. **While** $\nu_k(s_t, a_t) \leq \max\{1, N_k(s_t, a_t)\}$ **do**
   (a) Execute a_t, obtain reward r_t, and observe next state s_{t+1}.
   (b) Set $\nu_k(s_t, a_t)$ += 1.
   (c) Sample action $a_{t+1} \sim \widetilde{\pi}_k(\cdot|s_{t+1})$ and set t += 1.
6. Set $N_{k+1}(s,a) = N_k(s,a) + \nu_k(s,a)$.

Figure 1. The general structure of optimistic algorithms for RL.

In this paper, we slightly modify problem (5) as follows:

$$(\widetilde{M}^*_c, \widetilde{\pi}^*_c) \in \underset{M \in \mathcal{M}_k, \pi \in \Pi_c(M)}{\arg\max}\, g^\pi_M, \tag{6}$$

where the search space of policies is defined as

$$\Pi_c(M) := \left\{\pi \in \Pi^{\text{SR}} : sp\{h^\pi_M\} \leq c \ \wedge \ sp\{g^\pi_M\} = 0\right\},$$

and $\max_{\pi \in \Pi_c(M)}\{g^\pi_M\} = -\infty$ if $\Pi_c(M) = \emptyset$. Similarly to (3), problem (6) is equivalent to solving $\widetilde{\mu}^*_c \in \arg\max_{\mu \in \Pi_c(\widetilde{\mathcal{M}}_k)} \{g^\mu_{\widetilde{\mathcal{M}}_k}\}$. Unlike (5), for *every* MDP in $\mathcal{M}_k$ (not just those in $\mathcal{M}_{\text{RC}}$), (6) considers *all* (stationary) policies with *constant gain* satisfying the span constraint (not just the deterministic optimal policies).

Since g^π_M and $sp\{h^\pi_M\}$ are in general non-continuous functions of (M, π), the argmax in (5) and (6) may not exist. Nevertheless, by reasoning in terms of supremum value, we can show that (6) is always a *relaxation* of (5) (where we enforce the additional constraint of constant gain).

Proposition 1. *Define the following restricted set of MDPs* $\mathcal{E}_k = \mathcal{M}_{\text{RC}} \cap \{M \in \mathcal{M}_k : sp\{g^*_M\} = 0\}$. *Then*

$$\sup_{M \in \mathcal{E}_k, \pi \in \Pi^{SD}} g^\pi_M \leq \sup_{M \in \mathcal{M}_k, \pi \in \Pi_c(M)} g^\pi_M.$$

Proof. The result follows from the fact that $\mathcal{E}_k \subseteq \mathcal{M}_k$ and $\forall M \in \mathcal{E}_k$, $\arg\max_{\pi \in \Pi^{\text{SD}}}\{g^\pi_M\} \subseteq \Pi_c(M)$. □

As a result, the *optimism* principle is preserved when moving from (5) to (6) and since the set of admissible MDPs $\mathcal{M}_k$ is the same, any algorithm solving (6) would enjoy the same regret guarantees as REGAL.C. In the following we further characterise problem (6), introduce a *truncated* value

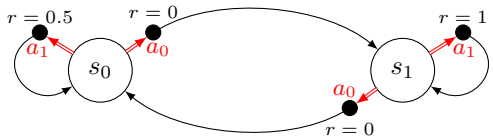

Figure 2. Toy example with deterministic transitions and reward for all actions.

iteration algorithm to solve it, and finally integrate it into a UCRL-like scheme to recover REGAL.C regret guarantees.

4. The Optimization Problem

In this section we analyze some properties of the following optimization problem, of which (6) is an instance,

$$\sup_{\pi \in \Pi_c(M)} \{g_M^\pi\}, \tag{7}$$

where M is any MDP (with discrete or compact action space) s.t. $\Pi_c(M) \neq \emptyset$. Problem (7) aims at finding a policy that maximizes the gain g_M^π within the set of randomized policies with constant gain (i.e., $sp\{g_M^\pi\} = 0$) and bias span smaller than c (i.e., $sp\{h_M^\pi\} \leq c$). Since $g_M^\pi \in [0, r_{\max}]$ the supremum always exists and we denote it by $g_c^*(M)$. The set of maximizers is denoted by $\Pi_c^*(M) \subseteq \Pi_c(M)$, with elements $\pi_c^*(M)$ (if $\Pi_c^*(M)$ is non-empty).

In order to give some intuition about the solutions of problem (7), we introduce the following illustrative MDP.

Example 1. *Consider the two-states MDP depicted in Fig. 2. For a generic stationary policy $\pi \in \Pi^{SR}$ with decision rule $d \in D^{MR}$ we have that*

$$d = \begin{bmatrix} x & 1-x \\ y & 1-y \end{bmatrix}; \quad P_d = \begin{bmatrix} 1-x & x \\ y & 1-y \end{bmatrix}, \quad r_d = \begin{bmatrix} \frac{1-x}{2} \\ 1-y \end{bmatrix}.$$

We can compute the gain $g = [g_1, g_2]$ and the bias $h = [h_1, h_2]$ by solving the linear system (1). For any $x > 0$ or $y > 0$, we obtain

$$g_1 = g_2 = \frac{1}{2} + x\frac{1-3y}{2(x+y)}; \quad h_2 - h_1 = \frac{1}{2} + \frac{1-3y}{2(x+y)},$$

while for $x = 0$, $y = 0$, we have $g_1 = 1/2$ and $g_2 = 1$, with $h_2 = h_1 = 0$. Note that $0 \leq sp\{h^\pi\} \leq 1$ for any $\pi \in \Pi^{SR}$. In the following, we will use this example choosing particular values for x, y, and c to illustrate some important properties of optimization problem (7).

Randomized policies. The following lemma shows that, unlike in unconstrained gain maximization where there always exists an optimal deterministic policy, the solution of (7) may indeed be a randomized policy.

Lemma 2. *There exists an MDP M and a scalar $c \geq 0$, such that $\Pi_c^*(M) \neq \emptyset$ and $\Pi_c^*(M) \cap \Pi^{SD}(M) = \emptyset$.*

Proof. Consider Ex. 1 with constraint $1/2 < c < 1$. The only deterministic policy π_D with constant gain and bias span smaller than c is defined by the decision rule with $x = 0$ and $y = 1$, which leads to $g^{\pi_D} = 1/2$ and $sp\{h^{\pi_D}\} = 1/2$. On the other hand, a randomized policy π_R can satisfy the constraint and maximize the gain by taking $x = 1$ and $y = (1-c)/(1+c)$, which gives $sp\{h^{\pi_R}\} = c$ and $g^{\pi_R} = c > g^{\pi_D}$, thus proving the statement. □

Constant gain. The following lemma shows that if we consider non-constant gain policies, the supremum in (7) may not be well defined, as no *dominating* policy exists. A policy $\pi \in \Pi^{\text{SR}}$ is *dominating* if for any policy $\pi' \in \Pi^{\text{SR}}$, $g^\pi(s) \geq g^{\pi'}(s)$ in all states $s \in \mathcal{S}$.

Lemma 3. *There exists an MDP M and a scalar $c \geq 0$, such that there exists no dominating policy π in Π^{SR} with constrained bias span (i.e., $sp\{h^\pi\} \leq c$).*

Proof. Consider Ex. 1 with constraint $1/2 < c < 1$. As shown in the proof of Lem. 2, the optimal stationary policy π_R with constant gain has $g_c^* = [c, c]$. On the other hand, the only policy π with non-constant gain is $x = 0$, $y = 0$, which has $sp\{h^\pi\} = 0 < c$ and $g^\pi(s_0) = 1/2 < c = g_c^*$ and $g^\pi(s_1) = 1 > c = g_c^*$, thus proving the statement. □

On the other hand, when the search space is restricted to policies with constant gain, the optimization problem is well posed. Whether problem (7) always admits a maximizer is left as an open question. The main difficulty comes from the fact that, in general, $\pi \mapsto g^\pi$ is not a continuous map and Π_c is not a closed set. For instance in Ex. 1, although the maximum is attained, the point $x = 0$, $y = 0$ does not belong to Π_c (i.e., Π_c is not closed) and g^π is not continuous at this point. Notice that when the MDP is *unichain* (Puterman, 1994, Sec. 8.3), Π_c is compact, g^π is continuous, and we can prove the following lemma (see App. A):

Lemma 4. *If M is unichain then $\Pi_c^*(M) \neq \emptyset$.*

We will later show that for the specific instances of (7) that are encountered by our algorithm SCAL, Lem. 4 holds.

5. Planning with SCOPT

In this section, we introduce SCOPT and derive sufficient conditions for its convergence to the solution of (7). In the next section, we will show that these assumptions always hold when SCOPT is carefully integrated into UCRL (while in App. B we show that they may not hold in general).

5.1. Span-constrained value and policy operators

SCOPT is a version of (relative) value iteration (Puterman, 1994; Bertsekas, 1995), where the optimal Bellman operator is modified to return value functions with span bounded by c, and the stopping condition is tailored to return a *constrained greedy* policy with near-optimal gain. We first introduce a *constrained* version of the optimal Bellman operator L.

Input: Initial vector $v_0 \in \mathbb{R}^S$, reference state $\overline{s} \in \mathcal{S}$, contractive factor $\gamma \in (0,1)$, accuracy $\varepsilon \in (0, +\infty)$
Output: Vector $v_n \in \mathbb{R}^S$, policy $\pi_n = (G_c v_n)^\infty$

1. Initialize $n = 0$ and $v_1 = T_c v_0 - (T_c v_0)(\overline{s})e$,
2. **While** $sp\{v_{n+1} - v_n\} + \frac{2\gamma^n}{1-\gamma} sp\{v_1 - v_0\} > \varepsilon$ **do**
   (a) $n \mathrel{+}= 1$.
   (b) $v_{n+1} = T_c v_n - (T_c v_n)(\overline{s})e$.

Figure 3. Algorithm SCOPT.

Definition 1. *Given $v \in \mathbb{R}^S$ and $c \geq 0$, we define the value operator $T_c : \mathbb{R}^S \to \mathbb{R}^S$ as*

$$T_c v = \begin{cases} Lv(s) & \forall s \in \overline{\mathcal{S}}(c,v), \\ c + \min_s\{Lv(s)\} & \forall s \in \mathcal{S} \setminus \overline{\mathcal{S}}(c,v), \end{cases} \tag{8}$$

where $\overline{\mathcal{S}}(c,v) = \{s \in \mathcal{S} | Lv(s) \leq \min_s\{Lv(s)\} + c\}$.

In other words, operator T_c applies a *span truncation* to the one-step application of L, that is, for any state $s \in \mathcal{S}$, $T_c v(s) = \min\{Lv(s), \min_x Lv(x) + c\}$, which guarantees that $sp\{T_c v\} \leq c$. Unlike L, operator T_c is not always associated with a decision rule d s.t. $T_c v = L_d v$ (see App. B). We say that T_c is *feasible* at $v \in \mathbb{R}^S$ and $s \in \mathcal{S}$ if there exists a distribution $\delta_v^+(s) \in \mathcal{P}(A)$ such that

$$T_c v(s) = \sum_{a \in \mathcal{A}_s} \delta_v^+(s,a)\big[r(s,a) + p(\cdot|s,a)^\mathsf{T} v\big]. \tag{9}$$

When a distribution $\delta_v^+(s)$ exists in all states, we say that T_c is *globally feasible* at v, and δ_v^+ is its associated decision rule, i.e., $T_c v = L_{\delta_v^+} v$. In the following lemma, we identify sufficient and necessary conditions for (global) *feasibility*.

Lemma 5. *Operator T_c is* feasible *at $v \in \mathbb{R}^S$ and $s \in \mathcal{S}$ if and only if*

$$\min_{a \in \mathcal{A}_s}\{r(s,a) + p(\cdot|s,a)^\mathsf{T} v\} \leq \min_{s'}\{Lv(s')\} + c. \tag{10}$$

Furthermore, let

$$D(c,v) := \big\{d \in D^{MR} \mid sp\{L_d v\} \leq c\big\} \tag{11}$$

be the set of randomized decision rules d whose associated operator L_d returns a span-constrained value function when applied to v. Then, $T_c v$ is globally feasible *if and only if $D(c,v) \neq \emptyset$, in which case we have*

$$T_c v = \max_{\delta \in D(c,v)} L_\delta v, \quad \textit{and} \quad \delta_v^+ \in \arg\max_{\delta \in D(c,v)} L_\delta v. \tag{12}$$

The last part of this lemma shows that when T_c is globally feasible at v (i.e., $D(c,v) \neq \emptyset$), $T_c v = L_{\delta_v^+} v$ is the *componentwise maximal* value function of the form $L_\delta v$ with decision rule $\delta \in D^{\text{MR}}$ satisfying $sp\{L_\delta v\} \leq c$. Surprisingly, even in the presence of a constraint on the one-step value span, such a *componentwise* maximum still exists (which is not as straightforward as in the case of the greedy operator L). Therefore, whenever $D(c,v) \neq \emptyset$, optimization problem (12) can be seen as an LP-problem (see App. A.2).

Definition 2. *Given $v \in \mathbb{R}^S$ and $c \geq 0$, let $\widetilde{\mathcal{S}}(c,v)$ be the set of states where $T_c v$ is* feasible *(condition (10)) with $\delta_v^+(s)$ be the associated decision rule (Eq. 9). We define the operator $G_c : \mathbb{R}^S \to D^{MR}$ as*[3]

$$G_c v = \begin{cases} \delta_v^+(s) & s \in \widetilde{\mathcal{S}}(C,v), \\ \arg\min_{a \in \mathcal{A}_s}\big\{r(s,a) + p(\cdot|s,a)^\mathsf{T} v\big\} & s \in \mathcal{S} \setminus \widetilde{\mathcal{S}}(C,v). \end{cases}$$

As a result, if T_c is globally feasible at v, by definition $G_c v = \delta_v^+$. Note that computing δ_v^+ is *not* significantly more difficult than computing a greedy policy (see App. C for an *efficient implementation*).

We are now ready to introduce SCOPT (Fig. 3). Given a vector $v_0 \in \mathbb{R}^S$ and a reference state $\overline{s}$, SCOPT implements relative value iteration where L is replaced by T_c, i.e.,

$$v_{n+1} = T_c v_n - T_c v_n(\overline{s})e. \tag{13}$$

Notice that the term $(T_c v_n)(\overline{s})e$ subtracted at any iteration n prevents v_n from increasing linearly with n and thus avoids numerical instability. However, the subtraction can be dropped without affecting the convergence properties of SCOPT. If the stopping condition is met at iteration n, SCOPT returns policy $\pi_n = d_n^\infty$ where $d_n = G_c v_n$.

5.2. Convergence and Optimality Guarantees

In order to derive convergence and optimality guarantees for SCOPT we need to analyze the properties of operator T_c. We start by proving that T_c preserves the one-step *span contraction* properties of L.

Assumption 6. *The optimal Bellman operator L is a 1-step γ-span-contraction, i.e., there exists a $\gamma < 1$ such that for any vectors $u, v \in \mathbb{R}^S$, $sp\{Lu - Lv\} \leq \gamma sp\{u - v\}$.*[4]

Lemma 7. *Under Asm. 6, T_c is a γ-span contraction.*

The proof of Lemma 7 relies on the fact that the truncation of L in the definition of T_c is non-expansive in span seminorm. Details are given in App. D, where it is also shown that T_c preserves other properties of L such as *monotonicity* and *linearity*. It then follows that T_c admits a fixed point solution to an optimality equation (similar to L) and thus SCOPT converges to the corresponding bias and gain, the latter being an upper-bound on the optimal solution of (7). We formally state these results in Lem. 8.

Lemma 8. *Under Asm. 6, the following properties hold:*

1. Optimality equation and uniqueness: *There exists a solution $(g^+, h^+) \in \mathbb{R} \times \mathbb{R}^S$ to the optimality equation*

$$T_c h^+ = h^+ + g^+ e. \tag{14}$$

[3] When there are several policies δ_v^+ achieving $T_c v(s) = L_{\delta_v^+} v(s)$ in state $s \in \mathcal{S}$, G_c chooses an arbitrary decision rule.

[4] In the undiscounted setting, if the MDP is unichain, L is a J-stage contraction with $S \geq J \geq 1$.

If $(g, h) \in \mathbb{R} \times \mathbb{R}^S$ *is another solution of* (14)*, then* $g = g^+$ *and there exists* $\lambda \in \mathbb{R}$ *s.t.* $h = h^+ + \lambda e$.

2. Convergence: *For any initial vector* $v_0 \in \mathbb{R}^S$*, the sequence* (v_n) *generated by* SCOPT *converges to a solution vector* h^+ *of the optimality equation* (14)*, and*
$$\lim_{n \to +\infty} T_c^{n+1} v_0 - T_c^n v_0 = g^+ e.$$

3. Dominance: *The gain* g^+ *is an upper-bound on the supremum of* (7)*, i.e.,* $g^+ \geq g_c^*$.

A direct consequence of point 2 of Lem. 8 (convergence) is that SCOPT always stops after a finite number of iterations. Nonetheless, T_c may not always be globally feasible at h^+ (see App. B) and thus there may be no policy associated to optimality equation (14). Furthermore, even when there is one, Lem. 8 provides no guarantee on the performance of the policy returned by SCOPT after a finite number of iterations. To overcome these limitations, we introduce an additional assumption, which leads to stronger performance guarantees for SCOPT.

Assumption 9. *Operator* T_c *is globally feasible at any vector* $v \in \mathbb{R}^S$ *such that* $sp\{v\} \leq c$.

Theorem 10. *Assume Asm. 6 and 9 hold and let* γ *denote the contractive factor of* T_c *(Asm. 6). For any* $v_0 \in \mathbb{R}^S$ *such that* $sp\{v_0\} \leq c$*, any* $\overline{s} \in \mathcal{S}$ *and any* $\varepsilon > 0$*, the policy* π_n *output by* SCOPT$(v_0, \overline{s}, \gamma, \varepsilon)$ *is such that* $\|g^+ e - g^{\pi_n}\|_\infty \leq \varepsilon$*. Furthermore, if in addition the policy* $\pi^+ = (G_c h^+)^\infty$ *is* unichain*,* g^+ *is the solution to optimization problem* (7) *i.e.,* $g^+ = g_c^*$ *and* $\pi^+ \in \Pi_c^*$.

The first part of the theorem shows that the stopping condition used in Fig. 3 ensures that SCOPT returns an ε-optimal policy π_n. Notice that while $sp\{h^+\} = sp\{T_c h^+\} \leq c$ by definition of T_c, in general when the policy $\pi^+ = (G_c h^+)^\infty$ associated to h^+ is *not unichain*, we might have $sp\{h^+\} < sp\{h^{\pi^+}\}$. On the other hand, Corollary 8.2.7. of Puterman (1994) ensures that if π^+ is unichain then $sp\{h^+\} = sp\{h^{\pi^+}\}$, hence the second part of the theorem. Notice also that even if π^+ is unichain, we cannot guarantee that π_n satisfies the span constraint, i.e., $sp\{h^{\pi_n}\}$ may be arbitrary larger than c. Nonetheless, in the next section, we show that the definition of T_c and Thm. 10 are sufficient to derive regret bounds when SCOPT is integrated into UCRL.

6. Learning with SCAL

In this section we introduce SCAL, an optimistic online RL algorithm that employs SCOPT to compute policies that efficiently balance exploration and exploitation. We prove that the assumptions stated in Sec. 5.2 hold when SCOPT is integrated into the optimistic framework. Finally, we show that SCAL enjoys the same regret guarantees as REGAL.C, while being the first implementable and efficient algorithm to solve bias-span constrained exploration-exploitation.

Based on Def. 1, we define $\widetilde{T}_c$ as the *span truncation* of the optimal Bellman operator $\widetilde{L}$ of the bounded-parameter MDP $\widetilde{\mathcal{M}}_k$ (see Sec. 3). Given the structure of problem (6), one might consider applying SCOPT (using $\widetilde{T}_c$) to the extended MDP $\widetilde{\mathcal{M}}_k$. Unfortunately, in general $\widetilde{L}$ does not satisfy Asm. 6 and 9 and thus $\widetilde{T}_c$ may not enjoy the properties of Lem. 8 and Thm. 10. To overcome this problem, we slightly modify $\widetilde{\mathcal{M}}_k$ as described in Def. 3.

Definition 3. *Let* $\widetilde{\mathcal{M}}$ *be a bounded-parameter (extended) MDP. Let* $1 \geq \eta > 0$ *and* $\overline{s} \in \mathcal{S}$ *an arbitrary state. We define the "modified" MDP* $\widetilde{\mathcal{M}}^\ddagger$ *associated to* $\widetilde{\mathcal{M}}$ *by*[5]
$$B_r^\ddagger(s,a) = [0, \max\{B_r(s,a)\}],$$
$$B_p^\ddagger(s,a,s') = \begin{cases} B_p(s,a,s') & \text{if } s' \neq \overline{s}, \\ B_p(s,a,\overline{s}) \cap [\eta, 1] & \text{otherwise}, \end{cases}$$
where we assume that η *is small enough so that:* $B_p(s,a,\overline{s}) \cap [\eta,1] \neq \emptyset$*,* $\sum_{s' \in \mathcal{S}} \min\{B_p^\ddagger(s,a,s')\} \leq 1$*, and* $\sum_{s' \in \mathcal{S}} \max\{B_p^\ddagger(s,a,s')\} \geq 1$*. We denote by* $\widetilde{L}^\ddagger$ *the optimal Bellman operator of* $\widetilde{\mathcal{M}}^\ddagger$ *(cf. Eq. 4) and by* $\widetilde{T}_c^\ddagger$ *the span truncation of* $\widetilde{L}^\ddagger$ *(cf. Def. 1).*

By slightly *perturbing* the confidence intervals B_p of the transition probabilities, we enforce that the "*attractive*" state $\overline{s}$ is reached with non-zero probability from any state-action pair (s,a) implying that the *ergodic coefficient* of $\widetilde{\mathcal{M}}^\ddagger$
$$\gamma = 1 - \min_{\substack{s,u \in \mathcal{S},\ a,b \in \mathcal{A} \\ \widetilde{p},\ \widetilde{q} \in B_p^\ddagger}} \left\{ \sum_{j \in \mathcal{S}} \underbrace{\min\{\widetilde{p}(j|s,a), \widetilde{q}(j|u,b)\}}_{\geq \eta \text{ if } j = \overline{s}} \right\}$$
is smaller than $1 - \eta < 1$, so that $\widetilde{L}^\ddagger$ is γ-contractive (Puterman, 1994, Thm. 6.6.6), i.e., Asm. 6 holds. Moreover, for any policy $\pi \in \Pi^{\mathrm{SR}}(\widetilde{\mathcal{M}}^\ddagger)$, state $\overline{s}$ necessarily belongs to all *recurrent classes* of π implying that π is unichain and so $\widetilde{\mathcal{M}}^\ddagger$ is unichain. As is shown in Thm. 11, the η-perturbation of B_p introduces a small bias ηc in the final gain.

By *augmenting* (without perturbing) the confidence intervals B_r of the rewards, we ensure two nice properties. First of all, for any vector $v \in \mathbb{R}^S$, $\widetilde{L} v = \widetilde{L}^\ddagger v$ and thus by definition $\widetilde{T}_c v = \widetilde{T}_c^\ddagger v$. Secondly, there exists a decision rule $\delta \in D^{\mathrm{MR}}(\widetilde{\mathcal{M}}^\ddagger)$ such that $\forall s \in \mathcal{S}$, $\widetilde{r}_\delta^\ddagger(s) = 0$ meaning that $sp\{\widetilde{L}_\delta^\ddagger v\} = sp\{\widetilde{P}_\delta^\ddagger v\} \leq sp\{v\}$ (Puterman, 1994, Proposition 6.6.1). Thus if $sp\{v\} \leq c$ then $sp\{\widetilde{L}_\delta^\ddagger v\} \leq c$ and so $\delta \in \widetilde{D}^\ddagger(c, v) \neq \emptyset$ which by Lem. 5 implies that $\widetilde{T}_c^\ddagger$ is globally feasible at v. Therefore, Asm. 9 holds in $\widetilde{\mathcal{M}}^\ddagger$.

When combining both the perturbation of B_p and the augmentation of B_r we obtain Thm. 11 (proof in App. E).

[5] For any closed interval $[a,b] \subset \mathbb{R}$, $\max\{[a,b]\} := b$ and $\min\{[a,b]\} := a$

Theorem 11. *Let $\widetilde{\mathcal{M}}$ be a bounded-parameter (extended) MDP and $\widetilde{\mathcal{M}}^{\ddagger}$ its "modified" counterpart (see Def. 3). Then*

1. *$\widetilde{L}^{\ddagger}$ is a γ-span contraction with $\gamma \leq 1-\eta < 1$ (i.e., Asm. 6 holds) and thus Lem. 8 applies to $\widetilde{T}_c^{\ddagger}$. Denote by (g^+, h^+) a solution to equation (14) for $\widetilde{T}_c^{\ddagger}$.*
2. *$\widetilde{T}_c^{\ddagger}$ is globally feasible at any $v \in \mathbb{R}^S$ s.t. $sp\{v\} \leq c$ (i.e., Asm. 9 holds) and $\widetilde{\mathcal{M}}^{\ddagger}$ is unichain implying that $\pi^+ = G_c h^+$ is unichain. Thus Thm. 10 applies to $\widetilde{T}_c^{\ddagger}$.*
3. *$\forall \mu \in \Pi_c(\widetilde{\mathcal{M}}), \quad g^+ = g_c^*(\widetilde{\mathcal{M}}^{\ddagger}) \geq g^{\mu}(\widetilde{\mathcal{M}}) - \eta c.$*

SCAL (cf. Fig. 1) is a variant of UCRL that applies SCOPT (instead of EVI, see Eq. 4) on the bounded parameter MDP $\widetilde{\mathcal{M}}_k^{\ddagger}$ (instead of $\widetilde{\mathcal{M}}_k$, cf. step 2 in Fig. 1) in each episode k to solve the optimization problem

$$\max_{M \in \widetilde{\mathcal{M}}_k^{\ddagger}, \pi \in \Pi_c(M)} g_M^{\pi}, \tag{15}$$

whose maximum is denoted by $g_c^*(\widetilde{\mathcal{M}}_k^{\ddagger})$. The intervals $B_p^{\ddagger}$ of $\widetilde{\mathcal{M}}_k^{\ddagger}$ are constructed using parameter[6] $\eta_k = r_{\max}/(c \cdot t_k)$ and an arbitrary attractive state $\overline{s} \in \mathcal{S}$. SCOPT is run at step 3 in Fig. 1 with an initial value function $v_0 = 0$, the same reference state $\overline{s}$ used for the construction of $B_p^{\ddagger}$, contraction factor $\gamma_k = 1 - \eta_k$, and accuracy $\varepsilon_k = r_{\max}/\sqrt{t_k}$. SCOPT finally returns an optimistic (nearly) optimal policy satisfying the span constraint. This policy is executed until the end of the episode.

Thm. 11 ensures that the specific instance of problem (6) for SCAL (i.e., problem (15)) is well defined and admits a maximizer $\pi_c^*(\widetilde{\mathcal{M}}_k^{\ddagger})$ that can be efficiently computed using SCOPT. Moreover, up to an accuracy $\eta_k \cdot c = r_{\max}/t_k$, policy $\pi_c^*(\widetilde{\mathcal{M}}_k^{\ddagger})$ is still optimistic w.r.t. all policies in the set of constrained policies $\Pi_c(\widetilde{\mathcal{M}}_k)$ for the *initial* extended MDP. Since the true (unknown) MDP M^* belongs to $\mathcal{M}_k$ with high probability, under the assumption that $sp\{h_{M^*}^*\} \leq c$, $g_c^*(\widetilde{\mathcal{M}}_k^{\ddagger}) \geq g_{M^*}^* - r_{\max}/t_k$. As briefly mentioned in Sec. 5, in practice SCOPT can only output an approximation $\widetilde{\mu}_k$ of $\pi_c^*(\widetilde{\mathcal{M}}_k^{\ddagger})$ and we have no guarantees on $sp\{h^{\widetilde{\mu}_k}\}$. However, the regret proof of SCAL only uses the fact that $sp\{v_n\} \leq c$ and this is always satisfied by definition of $\widetilde{T}_c^{\ddagger}$. We are now ready to prove the following regret bound (see App. F).

Theorem 12. *For any* weakly communicating *MDP M such that $sp\{h_M^*\} \leq c$, with probability at least $1-\delta$ it holds that for any $T \geq 1$, the regret of* SCAL *is bounded as*

$$\Delta(\text{SCAL}, T) = \mathcal{O}\left(\max\{r_{\max}, c\}\sqrt{\Gamma S A T \ln\left(\frac{T}{\delta}\right)}\right),$$

where $\Gamma = \max_{s \in \mathcal{S}, a \in \mathcal{A}} \|p(\cdot|s,a)\|_0 \leq S$ is the maximal number of states that can be reached from any state.

[6] Notice that given that $\beta_{p,k}^{sa} \geq \eta_k$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ (see definition in Sec. 3), the assumptions of Def. 3 hold trivially.

The previous bound shows that when $c \leq r_{\max} D$, SCAL scales linearly with c, while UCRL scales linearly with $r_{\max} D$ (all other terms being equal). Notice that the gap between $sp\{h^*\}$ and D can be arbitrarily large, and thus the improvement can be significant in many MDPs. As an extreme case, in weakly communicating MDPs the diameter can be infinite, leading UCRL to suffer linear regret, while SCAL is still able to achieve sub-linear regret. However when $c > r_{\max} D$, given that the true MDP M^* may not belong to $\mathcal{M}_k^{\ddagger}$, we cannot guarantee that the span of the value function v_n returned by SCOPT is bounded by $r_{\max} D$. Nevertheless, we can slightly modify SCAL to address this case: at the beginning of any episode k, we run both SCOPT (with the same inputs) and EVI (as in UCRL) in parallel and pick the policy associated to the value with smallest span. With this modification, SCAL enjoys the best of both worlds, i.e., the regret scales with $\min\{\max\{r_{\max}, c\}, r_{\max} D\}$ instead of c. When c is wrongly chosen ($c < sp\{h_{M^*}^*\}$), SCAL converges to a policy in $\Pi_c^*(M^*)$ which can be arbitrarily worse than the true optimal policy in M^*. For this reason we cannot prove a regret bound in this scenario. Finally, notice that the benefit of SCAL over UCRL comes at a negligible additional computational cost.

7. Numerical Experiments

In this section, we numerically validate our theoretical findings. The code is available on GitHub. In particular, we show that the regret of UCRL indeed scales linearly with the diameter, while SCAL achieves much smaller regret that only depends on the span. This result is even more extreme in the case of non-communicating MDPs, where $D = \infty$. Consider the simple but descriptive three-state domain shown in Fig. 4(a) (results in a more complex domain are reported in App. G). In this example, the learning agent only has to choose which action to play in state s_2 (in all other states there is only one action to play). The rewards are distributed as Bernoulli with parameters shown in Fig. 4(a) and $r_{\max} = 1$. The optimal policy π^* is such that $\pi^*(s_2) = a_1$ with gain $g^* = \frac{2}{3}$ and bias $h^* = \left[\frac{-2-\delta}{3(1-\delta)}, \frac{-1}{1-\delta}, 0\right]$. If δ is small, $sp\{h^*\} = \frac{1}{1-\delta} \approx 1$, while $D \approx \frac{1}{\delta}$. Fig. 4(b) shows that, as predicted by theory, the regret of UCRL (for a fixed horizon T) grows linearly with $\frac{1}{\delta} \approx D$. The optimal bias span however is roughly equal to 1. Therefore, we expect SCAL to clearly outperform UCRL on this example. In all the experiments, we noticed that perturbing the extended MDP was not necessary to ensure convergence of SCOPT and so we set $\eta_k = 0$. We also set $\gamma_k = 0$ to speed-up the execution of SCOPT (see stopping condition in Fig. 3).

Communicating MDPs. We first set $\delta = 0.005 > 0$, giv-

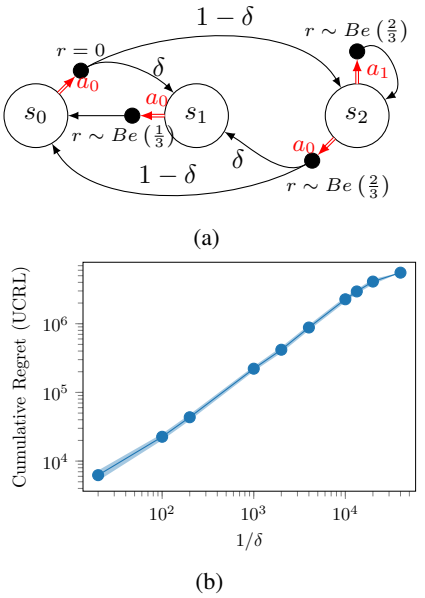

Figure 4. (upper) Simple three-state domain. *(lower)* Cumulative regret incurred by UCRL after $T = 2.5 \cdot 10^7$ steps as a function of the diameter $D \approx 1/\delta$ (averaged over 20 runs).

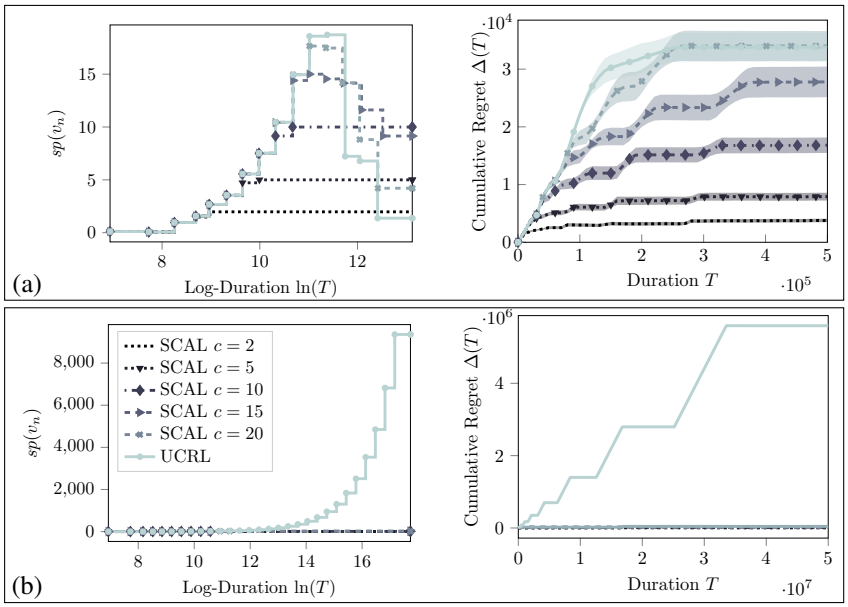

Figure 5. Results in the three-states domain with $\delta = 0.005$ *(top)* and $\delta = 0$ *(bottom)*. We report the span of the optimistic bias *(left)* and the cumulative regret *(right)* as a function of T. Results are averaged over 20 runs and 95% confidence intervals are shown.

ing a communicating MDP. With such a small δ, visiting state s_1 is rather unlikely. Nonetheless, since UCRL is based on the OFU principle, it keeps trying to visit s_1 (i.e., play a_0 in s_2) until it collects enough samples to understand that s_1 is actually a *bad* state (before that, UCRL *"optimistically"* assumes that s_1 is a *highly rewarding* state). Therefore, UCRL plays a_0 in s_2 for a long time and suffers large regret. This problem is particularly challenging for any learning algorithm solely employing *optimism* like UCRL (cf. (Ortner, 2008) for a more detailed discussion on the intrinsic limitations of optimism in RL). In contrast, SCAL is able to mitigate this issue when an appropriate constraint c is used. More precisely, whenever s_1 is believed to be the most rewarding state, the value function (bias) is maximal in s_1 and SCOPT applies a "truncation" in that state and "mixes" deterministic actions. In other words, SCAL leverages on the prior knowledge of the optimal bias span to understand that s_1 cannot be as good as predicted (from optimism). The exploration of the MDP is greatly affected as SCAL quickly discovers that action a_0 in s_2 is suboptimal. Therefore, SCAL is always performing better than UCRL (Fig. 5(a)) and the smaller c, the better the regret. Surprisingly the *actual* policy played by SCAL in this particular MDP is always deterministic. SCOPT mixes actions in s_1 where only one *true* action is available but the mixing happens in the *extended* MDP $\widetilde{\mathcal{M}}_k^{\ddagger}$ where the action set is compact. The policy that SCOPT outputs is thus *stochastic* in the *extended* MDP but *deterministic* in the *true* MDP.

Infinite Diameter. By selecting $\delta = 0$ the diameter becomes infinite ($D = +\infty$) but the MDP is still *weakly* communicating (with transient state s_1). UCRL is not able to handle this setting and suffers linear regret. On the contrary, SCAL is able to quickly recover the optimal policy (see Fig. 5(b) and App. G).

8. Conclusion

In this paper we introduced SCAL, a UCRL-like algorithm that is able to efficiently balance exploration and exploitation in any *weakly communicating* MDP for which a finite bound c on the optimal bias span $sp\{h^*\}$ is known. While UCRL exclusively relies on *optimism* and uses EVI to compute the exploratory policy, SCAL leverages the knowledge of c through the use of SCOPT, a new planning algorithm specifically designed to handle constraints on the bias span. We showed both theoretically and empirically that SCAL achieves smaller regret than UCRL. Although SCAL was inspired by REGAL.C, it is the only *implementable* approach so far. Therefore, this paper answers the long-standing open question of whether it is actually possible to design an *algorithm* that does not scale with the diameter D in the worst case. Moreover, SCAL paves the way for implementable algorithms able to learn in an MDP with *continuous* state space. Indeed, existing algorithms achieving regret guarantees in this framework (Ortner and Ryabko, 2012; Lakshmanan et al., 2015) all rely on REGAL.C. We also believe that our approach can easily be extended to optimistic PSRL (Agrawal and Jia, 2017) to achieve an even better regret bound of $\widetilde{\mathcal{O}}\left(\min\{c, r_{\max}D\}\sqrt{SAT}\right)$, i.e., drop the dependency in Γ. Finally, we leave it as an open question whether the assumption that c is known can be relaxed.

Acknowledgements

This research was supported in part by French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and French National Research Agency (ANR) under project ExTra-Learn (n.ANR-14-CE24-0010-01). Furthermore, this work was supported in part by the Austrian Science Fund (FWF): I 3437-N33 in the framework of the CHIST-ERA ERA-NET (DELTA project).

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence, UAI 2015*, pages 1–11. AUAI Press, 2015.

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

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory, 18th International Conference, ALT 2007*, volume 4754 of *Lecture Notes in Computer Science*, pages 150–165. Springer, 2007.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272, 2017.

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

Dimitri P Bertsekas. *Dynamic programming and optimal control. Vol II*. Athena Scientific, 1995.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems 28, NIPS 2015*, pages 2818–2826, 2015.

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

Donald E. Knuth. *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3rd edition, 1997.

K. Lakshmanan, Ronald Ortner, and Daniil Ryabko. Improved regret bounds for undiscounted continuous reinforcement learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *Proceedings of Machine Learning Research*, pages 524–532, 2015.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *COLT 2009 - The 22nd Conference on Learning Theory*, 2009.

Ronald Ortner. Optimism in the face of uncertainty should be refutable. *Minds and Machines*, 18(4):521–526, 2008.

Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems 25, NIPS 2012*, pages 1772–1780, 2012.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *CoRR*, abs/1608.02732, 2016.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26, NIPS 2013*, pages 3003–3011, 2013.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown Markov decision processes: A Thompson sampling approach. In *Advances in Neural Information Processing Systems 30, NIPS 2017*, pages 1333–1342, 2017.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

Eugene Seneta. Sensitivity of finite Markov chains under perturbation. *Statistics & Probability Letters*, 17(2):163–168, 1993.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Index of the Appendix

We start providing a brief recap of the content of the appendix:

- App. A
  - Proof of Lem. 4: If M is unichain then $\Pi_c^*(M) \neq \emptyset$. As a starting point the continuity of gain g and span h w.r.t. the policy is proved (see Lem. 13).
  - The policy associated to $T_c v$ can be interpreted as a solution of an LP problem (see App. A.2)
- App. B
  - Shows the limitations of SCOPT (T_c): non-feasibility B.1 and non-convergence B.2.
- App. C
  - We show how to compute the policy associated to the operator $T_c v$ when T_c is feasible in v. We consider both MDPs and extended MDPs.
- App. D
  - Proof of Lem. 5 i.e., when operator T_c is feasible (see App. D.1).
  - Proof of Lem. 7 i.e., that under Asm. 6, T_c is a span contraction (see App. D.2).
  - Proof of Lem. 8 i.e., existence and uniqueness of the optimality equation for T_c, convergence of SCOPT and gain dominance (see App. D.3)
  - Proof of Thm. 10 i.e., stopping condition for SCOPT and approximation guarantees (see App. D.4).
- App. E
  - A formal definition of perturbed extended MDP (see Lem. 19) and span contraction property for $\widetilde{L}$ in the perturbed MDP.
  - A formal definition of (reward) augmented extended MDP (see Lem. 20), equality of the operator in the original and augmented extend MDPs, and non-emptiness of $D(c, v)$ in the augmented MDP when $sp\{v\} \leq c$.
  - Proof of Thm. 11. We prove existence and uniqueness of the optimality equation for T_c, convergence of SCOPT and gain dominance for the perturbed and augmented extended MDP (i.e., $\mathcal{M}_K^{\ddagger}$) (see Thm. 21).
- App F
  - Proof of Thm. 12 i.e., the regret of SCOPT.
- App G
  - We test SCAL and UCRL on a larger and more challenging domain (the knight quest)

A. Optimization with bias span constraint

A.1. Existence of gain optimal policies under bias-span constraint: the unichain case (proof of Lem. 4)

In this section we provide a formal proof of Lem. 4.

In unichain MDPs, all policies $\pi \in \Pi^{\text{SR}}$ have a constant gain g^π (Puterman, 1994, section 8.4), thus the search space reduces to $\Pi_c = \{\pi \in \Pi^{\text{SR}} : sp\{h^\pi\} \leq c\}$. We assume that $\Pi_c \neq \emptyset$. We first prove the following lemma.

Lemma 13. *In a unichain MDP, $g : \pi \mapsto g^\pi$ and $h : \pi \mapsto h^\pi$ are continuous maps from Π^{SR} to $\mathbb{R}$ and Π^{SR} to $\mathbb{R}^S$ respectively.*

Proof. Let's consider two stationary policies $\pi = d^\infty \in \Pi^{\text{SR}}$ and $\widehat{\pi} = \widehat{d}^\infty \in \Pi^{\text{SR}}$. Denote by P and $\widehat{P}$ the transition matrices associated to d and $\widehat{d}$ respectively. Since the MDP is unichain by assumption, the Markov Chains characterized by P and $\widehat{P}$ each have a unique stationary distribution μ and $\widehat{\mu}$ respectively. We express the gap $\mu - \widehat{\mu}$ using the same decomposition as Seneta (1993)

$$(\mu^\intercal - \widehat{\mu}^\intercal)(I - \widehat{P} + e\widehat{\mu}^\intercal) = \mu^\intercal(P - \widehat{P}) \implies (\mu^\intercal - \widehat{\mu}^\intercal) = \mu^\intercal(P - \widehat{P})H_{\widehat{P}}$$

where I is the identity matrix, $e = (1 \dots, 1)^\intercal$ is the vector of all 1's and $H_{\widehat{P}} = (I - \widehat{P} + e\widehat{\mu}^\intercal)^{-1} - e\widehat{\mu}^\intercal$ is the *Drazin inverse* of $I - \widehat{P}$ also known as the *deviation matrix* of $\widehat{P}$ (always well-defined, see Appendix A of (Puterman, 1994)). The above equality implies that

$$\|\mu^\intercal - \widehat{\mu}^\intercal\|_1 \leq \underbrace{\|\mu^\intercal\|_1}_{=1} \|P - \widehat{P}\|_{\infty,1} \|H_{\widehat{P}}\|_{\infty,\infty} = \|P - \widehat{P}\|_{\infty,1} \|H_{\widehat{P}}\|_{\infty,\infty}$$

where $\|A\|_{\infty,1} := \max_i \sum_j |A_{ij}|$ and $\|A\|_{\infty,\infty} := \max_{i,j} |A_{ij}|$. As a consequence of the above inequality, when $P \to \widehat{P}$ we have $\mu \to \widehat{\mu}$. Moreover, when $d \to \widehat{d}$ we have $P \to \widehat{P}$ by linearity and thus by composition:

$$\lim_{d \to \widehat{d}} \mu = \widehat{\mu}$$

Denote by r and $\widehat{r}$ the reward functions associated to d and $\widehat{d}$ respectively. We have $g^\pi = \mu^\intercal r$ and $g^{\widehat{\pi}} = \widehat{\mu}^\intercal \widehat{r}$ and since r is linear (hence continuous) in d we conclude that

$$\lim_{\pi \to \widehat{\pi}} g^\pi = g^{\widehat{\pi}}$$

or in other words, $g : \pi \mapsto g^\pi$ is continuous at $\widehat{\pi}$ and since $\widehat{\pi}$ was chosen arbitrarily, g is continuous everywhere. Similarly, $h^\pi = H_P r$ and $P \mapsto H_P$ is continuous in P (the computation of H_P involves only continuous operations of P and μ like addition, multiplication and inversion of matrices) and therefore h^π is continuous too. □

Note that Lem. 13 does not hold in general when the MDP is not unichain (see Ex. 1 when $x \to 0$ and $y \to 0$).

Since $sp\{\cdot\}$ is a semi-norm, it is a continuous map from Π^{SR} to $\mathbb{R}$ and so the function $f : \pi \mapsto sp\{h^\pi\}$ is continuous by composition. Since f is continuous, Π^{SR} is compact and $\mathbb{R}$ is a Hausdorff space, we know from basic topology that f is a proper map i.e, the preimage of every compact set in $\mathbb{R}$ by f is compact in Π^{SR}. Since we can express Π_c as the preimage of the compact interval $[0, c]$ by f i.e., $\Pi_c = f^{-1}([0, c])$, it is clear that Π_c is compact. As a result, since g^π is continuous in π and Π_c is compact, by Weierstrass extreme value theorem the maximum of g^π is attained in Π_c and so $\Pi_c^* \neq \emptyset$.

A.2. Greedy policy under bias span constraint: LP formulation

In this section, we show that the policy associated to $T_c v$ can be interpreted as the solution of a Linear Programming (LP) problem.

As mentioned in Sec. 5.1, a consequence of Lem. 5 (see proof in App. D) is that whenever $D(c, v) \neq \emptyset$, there exists $\delta_v^+ \in D(c, v)$ such that $L_{\delta_v^+} v \geq L_d v$ for all $d \in D(c, v)$ *component-wise*, and moreover $T_c v = L_{\delta_v^+} v$. As a result, we can express δ_v^+ as a maximizer of the following optimization problem

$$\max_{d \in D(c,v)} \{(L_d v)^\intercal e\} \tag{16}$$

where $e = (1 \dots 1)^\intercal$ is the vector of all 1's. The maximum of (16) is then $(T_c v)^\intercal e$. Since $d \mapsto r_d$ and $d \mapsto P_d$ are linear maps, the function we maximize $d \mapsto (L_d v)^\intercal e$ is also linear in d. Moreover, the set $D(c, v)$ can be expressed as a set of $S \times (S - 1)$ linear constraints in $L_d v$:

$$L_d v(s) - L_d v(s') \leq c, \quad \forall s \neq s'$$

Therefore, optimization problem (16) can be formulated as an LP problem. But of course it is much easier to compute $T_c v$ using Def. 1. The policy δ_v^+ associated to $T_c v$ can also be computed efficiently without solving (16) (see App. C for more details).

Remark. *Recall that computing the maximal gain of an MDP can be done by solving the following* primal *LP problem (Puterman, 1994, Section 8.8)*

$$\begin{aligned} &\min_{g \in \mathbb{R}, h \in \mathbb{R}^S} \{g\} \\ &\text{s.t. } g + h(s) - \sum_{s' \in \mathcal{S}} p(s'|s,a) h(s') \geq r(s,a), \quad \forall s \in \mathcal{S}, \ \forall a \in \mathcal{A}_s \end{aligned} \tag{17}$$

One might wonder whether it is possible to reformulate optimization problem (7) presented in Sec. 4 by adapting the above

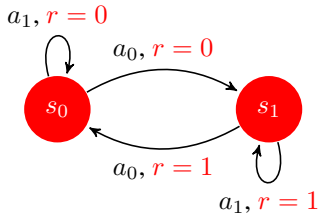

Figure 6. Example showing that T_c might not always be feasible even when $\Pi_c^* \neq \emptyset$.

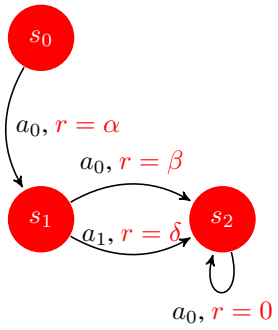

Figure 7. Example showing that T_c might not always be feasible at its fixed-point h^+.

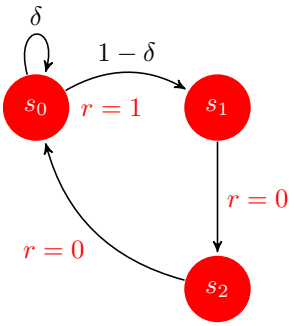

Figure 8. Example showing that the sequence $(T_c)^n v_0$ might not converge even when $\Pi_c \neq \emptyset$ and all policies are both unichain and aperiodic.

primal formulation with the addition of $S \times (S-1)$ linear constraints in h (as we did above for $L_d v$):

$$h(s) - h(s') \leq c, \quad \forall s \neq s'$$

Unfortunately it is not that simple. Indeed, the validity of LP problem (17) is a consequence of the following two properties (Puterman, 1994, Theorem 8.4.1):

$$1) \;\; ge + h - Lh \geq 0 \implies g \geq g^*$$
$$2) \;\; ge + h - Lh = 0 \implies g = g^*$$

In general, these properties no longer hold for operator T_c and optimal bias-span-constrained gain g_c^*. Therefore, the LP approach fails (one can easily try to solve the constrained LP on a simple MDP and observe the solution is incorrect). Using the dual *formulation is also tricky because the span constraint is not linear in the dual variables. Whether it is possible to formulate problem (7) as an LP problem is left as an open question.*

B. Limitations of SCOPT

In this section, we illustrate the limitations of operator T_c on some simple examples. For convenience, we introduce notation N_c to denote the (value) operator associated to policy $G_c v$ as

$$N_c v = L_{(G_c v)} v. \tag{18}$$

Trivially, if $T_c v$ is globally feasible, then $N_c v = T_c v$ and $G_c v = \delta_v^+$.

B.1. Non-feasibility of T_c

The following example shows that operator T_c may generate vectors that do not correspond to a one-step policy evaluation, i.e., there may not exists $\delta_v^+ \in D^{MR}$ such that $T_c v = L_{\delta_v^+} v$, even if $sp\{v\} \leq c$ and/or if $\Pi_c \neq \emptyset$.

Example 2. *Consider the simple MDP provided in Fig. 6. Let $v = [0, 0]$ and $c = 1/2$. In this case the policy π playing a_0 in both states matches the constraint c (i.e., $sp\{h^\pi\} \leq c \implies \pi \in \Pi_c \implies \Pi_c \neq \emptyset$) and moreover $sp\{v\} \leq c$, but clearly there exists no policy δ_v^+ achieving $L_{\delta_v^+} v = T_c v$ and moreover*

$$T_c v = [0, 1/2] \neq N_c v = [0, 1]$$

The previous example shows that there may not exists a policy associated to the *one-step application* of operator T_c. Instead, Ex. 3 shows that T_c may also not be feasible *at convergence*. In particular, we show that, surprising as it may seem, SCOPT can sometimes converge to a value h^+ that is not associated to any policy, even when $\Pi_c \neq \emptyset$ and even when $\Pi_c^* \neq \emptyset$.

Example 3. *Consider the simple MDP M of Fig. 7 where we assume that $\beta < \delta < \alpha$ and we set $c = \alpha + \beta$. The MDP is unichain and all gains are equal to 0. The set of randomized decision rules can be parametrized by the probability p of*

playing a_1 in s_1 and the associated set of bias functions is

$$H = \left\{ \begin{bmatrix} \alpha + (1-p)\cdot\beta + p\cdot\delta \\ (1-p)\cdot\beta + p\cdot\delta \\ 0 \end{bmatrix} : p \in [0,1] \right\}.$$

Let's denote by $h(p)$ the bias associated to a policy parameterized by p, then $sp\{h(p)\} = \alpha + (1-p)\cdot\beta + p\cdot\delta > c$ for all $p > 0$. So there exists only one policy $\pi = d^\infty$ achieving the span constraint which plays a_0 in s_1 (i.e., $p = 0$) implying that $\Pi_c(M) = \{d^\infty\} = \Pi_c^*(M) \neq \emptyset \implies \pi_c^* = d^\infty$. It is easy to verify that:

$$\textit{Fixed point of } T_c: \ h^+ = \begin{bmatrix} \alpha+\beta \\ \delta \\ 0 \end{bmatrix} \notin H$$

$$\textit{Fixed point of } N_c: \ h^\# = \begin{bmatrix} \alpha+\delta \\ \delta \\ 0 \end{bmatrix} \in H \ \textit{ but } \ sp\left\{h^\#\right\} > c,$$

Although T_c admits a fixed point h^+, it is not globally feasible at h^+. On the other hand, while N_c is globally feasible at its fixed point $h^\#$ by definition, $h^\#$ does not satisfy the bias constraint. One might be tempted to think that the problem in this example arises from the fact that $\Pi_c(M)$ is a singleton but it is actually more subtle than that. Indeed, if we assume that $\beta > 0$ and if we add an action a_2 in s_1 that goes to s_2 with probability 1 and gives a reward 0, we face the same problem but this time $\Pi_c(M)$ contains an infinite number of policies (since we include stochastic policies). The problem is actually coming from the fact that the action played in the only state achieving maximum bias (i.e., s_0) is deterministic while the action played in state s_1 (which achieves a lower bias than s_0) is stochastic. T_c is unable to converge to such policies: by definition, it can only converge to a policy that plays a stochastic action in the states with maximal bias (it can also converge to a bias that is not associated to any policy like in this example).

The issue presented in Ex. 3 can be overcome by duplicating all actions and adjusting the rewards of the duplicated actions. More formally, denote by $\overline{a}$ the action obtained by duplicating a. The probability of transition is not modified (i.e., $p(\cdot|s,\overline{a}) = p(\cdot|s,a)$) but the reward is set to the minimal value (i.e., $r(s,\overline{a}) = r_{\min} = 0$). Denote by $M^\downarrow$ this "*augmented*" MDP. It is easy to verify that $h^+(M^\downarrow) = h^+(M) = (\alpha+\beta, \beta, 0)^\intercal$ but unlike M, $M^\downarrow$ admits a policy associated to $h^+(M^\downarrow)$ (using duplicated actions). As this example shows, augmenting the MDP never modifies the fixed point of T_c but always makes T_c globally feasible at any vector v satisfying $sp\{v\} \leq c$. Since by definition the fixed point h^+ of T_c satisfies the span constraint, T_c is globally feasible at h^+. This example gives an intuition why SCAL uses a modified MDP $\widetilde{\mathcal{M}}_k^\ddagger$ with augmented rewards (the confidence intervals B_r^k are "augmented" by below).

B.2. Non-convergence of T_c^n

It is rather easy to design an MDP for which the stopping condition of SCOPT (i.e., $sp\{v_{n+1} - v_n\} \leq \varepsilon$) is never met although all policies are unichain and aperiodic and $\Pi_c \neq \emptyset$. In contrast, for the optimal Bellman operator L, unichain and aperiodicity are sufficient conditions to ensure that the stopping condition $sp\{v_{n+1} - v_n\} \leq \varepsilon$ is met after a finite number of iterations (Puterman, 1994, Theorem 8.5.7).

Example 4. *Consider the simple MDP M provided in Fig. 8 where we assume that $1 > \delta > 0$ and $1/2 \geq c > 0$. There is only one action available in every state and thus there is only one decision rule d. In that case $L = L_d$. The contraction condition of (Puterman, 1994, Theorem 8.5.3) holds for $J = 2$, i.e., L is a 2-stage span contraction. More precisely we have:*

$$P_d = \begin{bmatrix} \delta & 1-\delta & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \implies P_d^2 = \begin{bmatrix} \delta^2 & \delta-\delta^2 & 1-\delta \\ 1 & 0 & 1 \\ \delta & 1-\delta & 0 \end{bmatrix}$$

$$\implies \gamma_d \overset{def}{=} 1 - \min_{s,u\in\mathcal{S}} \left\{ \sum_{j\in\mathcal{S}} \min\{P_d^2(j|s), P_d^2(j|u)\} \right\} = \delta < 1$$

where γ_d is the ergodic coefficient *associated to Markov Chain P_d. This implies that L is a 2-stage γ_d-span contraction: $sp\left\{L^{2n+1}v - L^{2n}v\right\} \leq \gamma_d^n sp\{Lv - v\}$ for any vector v and any integer $n \geq 0$. On the other hand, the sequence $T_c^n v_0$*

starting from $v_0 = 0$ *proceeds as follows*

$$v_0 = \begin{bmatrix}0\\0\\0\end{bmatrix},\ v_1 = \begin{bmatrix}c\\0\\0\end{bmatrix},\ v_2 = \begin{bmatrix}c\\0\\c\end{bmatrix},\ v_3 = \begin{bmatrix}2c\\c\\c\end{bmatrix} = v_1 + ce,\ \ldots,\ v_{2n} = v_2 + nce,\ v_{2n+1} = v_1 + nce$$

where $e = (1, \ldots, 1)^\intercal$ *denotes the vector of all 1's. We see that unlike* $L^n v_0$, $(T_c)^n v_0$ *is cycling with period 2 and the quantity* $sp\{v_{2n+1} - v_{2n}\} = sp\{v_2 - v_1\}$ *does not converge to* 0. *Although Lem. 7 shows that when* L *is a* J*-stage span contraction with* $J = 1$ *then* T_c *is also a span contraction (proof in App. D), surprisingly* $(T_c)^n v_0$ *might not converge when* $J > 1$. *Note that in this example* $\Pi_c(M) = \emptyset$ *and so one might wonder whether when* $\Pi_c(M) \neq \emptyset$ *the sequence* $T_c^n v_0$ *converges in span semi-norm. Unfortunately, it is not the case. Take the same MDP, duplicate the action in* s_0 *and assign a reward of* 0 *to this new action (the new action loops on* s_0 *with probability* δ *and goes to* s_1 *with probability* $1 - \delta$ *as the original action, but the reward is* 0 *instead of* 1*). In that case,* $\Pi_c(M) \neq \emptyset$ *for all* $c \geq 0$ *but we still have* $L = L_d$ *where* d *plays the original action in* s_0. *Therefore, we face exactly the same problem as before although* $\Pi_c(M) \neq \emptyset$.

C. Policy δ_v^+ associated to $T_c v$

In this section, we provide a detailed description on how to efficiently compute a policy δ_v^+ associated to $T_c v$ when T_c is feasible at v. As mentioned in Sec. 4, we say that T_c is *feasible* at $v \in \mathbb{R}^S$ and $s \in \mathcal{S}$ when there exists a distribution $\delta_v^+(s) \in \mathcal{P}(\mathcal{A})$ such that

$$T_c v(s) = \sum_{a \in \mathcal{A}_s} \delta_v^+(s,a)\big[r(s,a) + p(\cdot,s,a)^\intercal v\big]. \tag{19}$$

We distinguish between two types of states:

- **Greedy states.** When $Lv(s) \leq \min\{Lv\} + c$ i.e., $s \in \overline{\mathcal{S}}(c,v)$ (see Def. 1), $\delta_v^+(s)$ plays a deterministic *greedy* action $\overline{a} \in \arg\max_{a \in \mathcal{A}_s}\{r(s,a) + p(\cdot|s,a)^\intercal v\}$.

- **Truncated states.** When $Lv(s) > \min\{Lv\} + c$ i.e., $s \notin \overline{\mathcal{S}}(c,v)$ (see Def. 1), by definition of L there exists at least one action $\overline{a}$ (e.g., any greedy action) such that $r(s,\overline{a}) + p(\cdot|s,\overline{a})^\intercal v > \min\{Lv\} + c$. In addition, under the assumption that T_c is feasible at v and s, we know from condition (10) of Lem. 5 (proof in App. D) that there exists an action $\underline{a}$ such that $r(s,\underline{a}) + p(\cdot|s,\underline{a})^\intercal v \leq \min\{Lv\} + c$. By the intermediate value theorem, we know that there exists a convex combination $\delta_v^+(s)$ of actions $\overline{a}$ and $\underline{a}$ achieving exactly $L_{\delta_v^+} v(s) = \min\{Lv\} + c$. Note that there may exist multiple policies achieving this value (e.g., when there are multiple actions achieving higher or smaller values than $\min\{Lv\} + c$). However, to simplify the implementation, we can simply set $\delta_v^+(s)$ to play with non-zero probability only a greedy action $\overline{a} \in \arg\max_{a \in \mathcal{A}_s}\{r(s,a) + p(\cdot|s,a)^\intercal v\}$ and a minimal action $\underline{a} \in \arg\min_{a \in \mathcal{A}_s}\{r(s,a) + p(\cdot|s,a)^\intercal v\}$. Formally, let $\underline{v} = \min_{a \in \mathcal{A}_s}\{r(s,a) + p(\cdot|s,a)^\intercal v\}$ and $\overline{v} = \max_{a \in \mathcal{A}_s}\{r(s,a) + p(\cdot|s,a)^\intercal v\} = Lv(s)$. Then,

$$\delta_v^+(s,a) = \begin{cases} (\overline{v} - \min\{Lv\} - c)/(\overline{v} - \underline{v}) & \text{if } a = \underline{a} \\ (\min\{Lv\} + c - \underline{v})/(\overline{v} - \underline{v}) & \text{if } a = \overline{a} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

Bounded-Parameter MDP. When we consider a bounded-parameter MDP $\widetilde{\mathcal{M}}$, the only change is in the computation of the minimal and maximal actions. Define

$$\widetilde{L}v(s) = \max_{a \in \mathcal{A}_s, \widetilde{r} \in B_r(s,a), \widetilde{p} \in B_p(s,a)} \big[\widetilde{r} + \widetilde{p}^\intercal v\big]$$
$$\underset{\sim}{L}v(s) = \min_{a \in \mathcal{A}_s, \widetilde{r} \in B_r(s,a), \widetilde{p} \in B_p(s,a)} \big[\widetilde{r} + \widetilde{p}^\intercal v\big]$$

Then, $\overline{a}$ and $\underline{a}$ are the actions associated to $\widetilde{L}v(s)$ and $\underset{\sim}{L}v(s)$, respectively. Given $\overline{a}$ and $\underline{a}$, the policy $\delta_v^+(s)$ associated to $T_c v(s)$ is computed as in (20). The maximum and minimum of $\widetilde{r}$ for $\widetilde{r} \in B_r(s,a)$ are easy to compute (they correspond to the extreme values of the closed interval $B_r(s,a)$). The maximum of $\widetilde{p}^\intercal v$ for $\widetilde{p}(s') \in B_p(s,a,s')$ can be computed in $\mathcal{O}(S)$ operations using the algorithm described in (Dann and Brunskill, 2015, Appendix A). To compute the minimum of $\widetilde{p}^\intercal v$, the exact same algorithm can be used with input $-v$ instead of v since $\min_{\widetilde{p}}\{\widetilde{p}^\intercal v\} = -\max_{\widetilde{p}}\{\widetilde{p}^\intercal(-v)\}$.

D. Properties of operators T_c and G_c

D.1. Feasibility of T_c (proof of Lemma 5)

In this section, we prove Lem. 5.

We start by proving that for any decision rule $d \in D(c, v)$, we have $T_c v \geq L_d v$ component-wise. By definition of T_c and the optimal Bellman operator L it holds that

$$\forall s \in \overline{\mathcal{S}}(c, v),\ \forall d \in D(c, v), \quad T_c v(s) = Lv(s) \geq L_d v(s).$$

Moreover, let $m := \min_s\{Lv(s)\}$, then

$$\begin{aligned}
\forall s \in \mathcal{S} \setminus \overline{\mathcal{S}}(c, v),\ \forall d \in D(c, v), \quad T_c v(s) &= m + c && (21)\\
&\geq m + sp\left\{L_d v\right\} && (22)\\
&= m + \max\{L_d v\} - \min\{L_d v\} && (23)\\
&\geq m + L_d v(s) - m = L_d v(s), && (24)
\end{aligned}$$

Inequality (22) is a consequence of the fact that $d \in D(c, v) \Leftrightarrow sp\left\{L_d v\right\} \leq c$. Denote by $\hat{s} \in \arg\max_s\{Lv(s)\}$ any state achieving minimum value for $Lv(s)$. Inequality (24) follows by noticing that $m = Lv(\hat{s}) \geq L_d v(\hat{s}) \geq \min\{L_d v\}$. In conclusion, for any $d \in D(c, v)$ and any $s \in \mathcal{S}$, $T_c v(s) \geq L_d v(s)$. This immediately implies that whenever T_c is globally feasible at v, $T_c v = \max_{\delta \in D(c,v)} L_\delta v$ and $\delta_v^+ \in \arg\max_{\delta \in D(c,v)} L_\delta v$.

Let's now prove the equivalence between the feasibility of T_c at $v \in \mathbb{R}^S$ and $s \in \mathcal{S}$ and condition (10) i.e.,

$$\min_{a \in \mathcal{A}_s}\{r(s, a) + p(\cdot, s, a)^\mathsf{T} v\} \leq \min_s\{Lv(s)\} + c.$$

If condition (10) holds, we can use the constructive procedure described in App. C to construct a stochastic action $\delta_v^+(s) \in \mathcal{P}(\mathcal{A})$ such that $L_{\delta_v^+} v(s) = \min_s\{Lv(s)\} + c = T_c v(s)$ and thus T_c is feasible at v and s. On the other hand, if condition (10) does not hold i.e., $\min_{a \in \mathcal{A}_s}\{r(s, a) + p(\cdot|s, a)^\mathsf{T} v\} > \min_s\{Lv(s)\} + c$ then it is clear that any $d(s) \in \mathcal{P}(\mathcal{A})$ will be such that $L_d v(s) > \min_s\{Lv(s)\} + c$ and so T_c is not feasible at v and s. By contraposition, if T_c is feasible at v and s then condition (10) holds thus proving the equivalence.

Finally, T_c is globally feasible at v if and only condition (10) holds in every state $s \in \mathcal{S}$. Trivially, if T_c is globally feasible then $sp\left\{T_c v\right\} = sp\left\{L_{\delta_v^+} v\right\} \leq c$ and so $\delta_v^+ \in D(c, v)$ implying that $D(c, v) \neq \emptyset$. If $D(c, v) \neq \emptyset$ then there exists $\delta \in D^{\text{MR}}$ such that $sp\left\{L_\delta v\right\} \leq c$. Assume that condition (10) does not hold in at least one state $\overline{s} \in \mathcal{S}$ meaning that $\min_{a \in \mathcal{A}_{\overline{s}}}\{r(\overline{s}, a) + p(\cdot|\overline{s}, a)^\mathsf{T} v\} > \min_s\{Lv(s)\} + c$ which implies that for any $d \in D^{\text{MR}}$, $L_d v(\overline{s}) > \min_s\{Lv(s)\} + c$. This contradicts the fact that

$$sp\left\{L_\delta v\right\} \leq c \implies L_\delta v(\overline{s}) \leq \max_s\{L_d v(s)\} \leq \min_s\{L_d v(s)\} + c \leq \min_s\{Lv(s)\} + c$$

where we used the definition of the span $sp\left\{u\right\} := \max\{u\} - \min\{u\}$ and the fact that $L_d v \leq Lv$ component-wise by definition of L implying that $\min_s\{L_d v(s)\} \leq \min_s\{Lv(s)\}$. Therefore, condition (10) must hold in every state. In conclusion, T_c is globally feasible at v if and only $D(c, v) \neq \emptyset$.

D.2. Contraction property of T_c (proof of Lemma 7)

The purpose of this section is to prove Lem. 7. We first reinterpret operator T_c as the composition of a *projection* Γ_c and the optimal Belmman operator L ($T_c = \Gamma_c L$) and we prove interesting properties for Γ_c and T_c.

Recall that T_c can be seen as the *truncation* of the optimal Bellman operator i.e., $T_c v(s) = \min\{Lv(s), \min_x\{Lv(x)\} + c\}$. The following lemma shows that the truncation step is actually a projection in span semi-norm. Let $V_c = \{v : sp\left\{v\right\} \leq c\}$ be the *"semi-ball"* of span constrained value functions. For any vector $v \in \mathbb{R}^S$ and any $c \geq 0$, we define the truncation operator Γ_c: $\mathbb{R}^S \to V_c$ as $\Gamma_c v(s) = \min\left\{v(s), \min_x\{v(x)\} + c\right\}$.

Lemma 14. *For any vector $v \in \mathbb{R}^S$ and $c \geq 0$, $\Gamma_c v$ is the projection of v on the semi-ball V_c in span semi-norm i.e.,*

$$\Gamma_c v = \min_{z \in V_c} sp\left\{z - v\right\}.$$

Proof. Let $w = \Gamma_c v$. If $sp\left\{v\right\} \leq c$, then by definition of Γ_c, $w = \Gamma_c v = v \in \arg\min_{z \in V_c} sp\left\{z - v\right\}$. Otherwise, using

again the definition of Γ_c we have that $w \leq v$ component-wise. As a result, we have

$$\max_s\{w(s) - v(s)\} = 0$$

Moreover, the difference between w and v is maximal in the states $\overline{s} \in \arg\max_s v(s)$ and thus

$$\min_s\{w(s) - v(s)\} = -\max_s\{v(s)\} + \min_s\{v(s)\} + c,$$

Therefore: $sp\{w - v\} = \max\{w - v\} - \min\{w - v\} = sp\{v\} - c$. Furthermore, by reverse triangle inequality[7], for any vector z such that $sp\{z\} \leq c$ we have that

$$sp\{z - v\} \geq sp\{v\} - sp\{z\} \geq sp\{v\} - c = sp\{w - v\},$$

thus proving the lemma. □

We prove the following useful properties for Γ_c:

Lemma 15. *Let v and u be vectors in $\mathbb{R}^S$, then:*

(a) Monotonicity

$$v \geq u \implies \Gamma_c v \geq \Gamma_c u.$$

(b) For any $s \in \mathcal{S}$

$$\min\{v - u\} \leq \Gamma_c v(s) - \Gamma_c u(s) \leq \max\{v - u\}. \tag{25}$$

(c) Γ_c is non-expansive in span semi-norm

$$sp\{\Gamma_c v - \Gamma_c u\} \leq sp\{v - u\}.$$

(d) Γ_c is non-expansive in ℓ_∞-norm

$$\|\Gamma_c v - \Gamma_c u\|_\infty \leq \|v - u\|_\infty.$$

(e) Linearity

$$\forall \lambda \in \mathbb{R},\ \ \Gamma_c(v + \lambda e) = \Gamma_c v + \lambda e.$$

Proof. For any state $s \in \mathcal{S}$, the difference $\Gamma_c v(s) - \Gamma_c u(s)$ can only take four different values depending on the configuration of $v(s)$ and $u(s)$

$$\Gamma_c v(s) - \Gamma_c u(s) = \begin{cases} v(s) - u(s) & \text{if } u(s) \leq \min\{u\} + c \text{ and } v(s) \leq \min\{v\} + c & \text{(26a)} \\ \min\{v\} + c - u(s) & \text{if } u(s) \leq \min\{u\} + c \text{ and } v(s) > \min\{v\} + c & \text{(26b)} \\ v(s) - \min\{u\} - c & \text{if } u(s) > \min\{u\} + c \text{ and } v(s) \leq \min\{v\} + c & \text{(26c)} \\ \min\{v\} - \min\{u\} & \text{if } u(s) > \min\{u\} + c \text{ and } v(s) > \min\{v\} + c & \text{(26d)} \end{cases}$$

(a) We need to show that for all $s \in \mathcal{S}$, the difference $\Gamma_c v(s) - \Gamma_c u(s)$ is bigger or equal than zero in all four cases. Case (26a) follows directly from the assumption $v \geq u$, while case (26d) is trivially proved since $v \geq u$ implies $\min\{v\} \geq \min\{u\}$. Case (26c) follows from $v(s) - \min\{u\} - c > v(s) - u(s) \geq 0$ (by assumption $u(s) > \min\{u\} + c$ in this case). Finally, case (26b) reduces to case (26d) since we assume that $u(s) \leq \min\{u\} + c$ implying that $\min\{v\} + c - u(s) \geq \min\{v\} - \min\{u\} \geq 0$.

(b) We treat all four cases separately as we did to prove *(a)*.

- Case (26a): it is straightforward to see that $\min\{v - u\} \leq \Gamma_c v(s) - \Gamma_c u(s) = v(s) - u(s) \leq \max\{v - u\}$
- Case (26b): we have that $\min\{v\} + c - u(s) < v(s) - u(s) \leq \max\{v - s\}$. To prove the other inequality we start by noticing that $\min\{v\} + c - u(s) > \min\{v\} + c - \min\{u\} - c = \min\{v\} - \min\{u\}$. Since moreover

$$\min_s\{v(s) - u(s)\} \leq \min_s\left\{v(s) - \min_{s'}\{u(s')\}\right\} = \min\{v\} - \min\{u\} \tag{27}$$

the inequality holds.

[7] The triangle inequality for the span is proved in (Puterman, 1994, Section 6.6.1).

- Case (26c): by definition $v(s) - \min\{u\} - c \leq \min\{v\} + c - \min\{u\} - c = \min\{v\} - \min\{u\}$. Then:

$$\max_s\{v(s) - u(s)\} \geq \max_s \left\{\min_{s'}\{v(s')\} - u(s)\right\} = \min\{v\} - \min\{u\} \tag{28}$$

  The other inequality trivially follows by noticing that $v(s) - \min\{u\} - c > v(s) - u(s) \geq \min\{v - u\}$.
- Case (26d): inequalities (25) is a consequence of inequalities (27) and (28).

(c) This is easy to prove exploiting inequality (25) (property *(b)*):

$$\begin{aligned} sp\left\{\Gamma_c v - \Gamma_c u\right\} &= \max\{\Gamma_c v - \Gamma_c u\} - \min\{\Gamma_c v - \Gamma_c u\} \\ &\leq \max\{v - u\} - \min\{v - u\} = sp\left\{v - u\right\} \end{aligned}$$

(d) We can again use inequality (25)

$$\begin{aligned} \left\|\Gamma_c v - \Gamma_c u\right\|_\infty &= \max_s \left\{\left|\Gamma_c v(s) - \Gamma_c u(s)\right|\right\} \\ &= \max\left\{\max_s\{\Gamma_c v(s) - \Gamma_c u(s)\}, \max_s\{\Gamma_c u(s) - \Gamma_c v(s)\}\right\} \\ &\leq \max\left\{\max_s\{v(s) - u(s)\}, \max_s\{u(s) - v(s)\}\right\} = \left\|v - u\right\|_\infty \end{aligned}$$

(e) By definition of Γ_c, for any $s \in \mathcal{S}$:

$$\begin{aligned} \Gamma_c(v + \lambda e)(s) &= \min\left\{v(s) + \lambda, \min\{v + \lambda e\} + c\right\} \\ &= \min\left\{v(s) + \lambda, \min\{v\} + \lambda + c\right\} \\ &= \min\left\{v(s), \min\{v\} + c\right\} + \lambda \\ &= \Gamma_c v(s) + \lambda. \end{aligned}$$

□

We are now ready to prove the following lemma:

Lemma 16. *Let v and u be vectors in $\mathbb{R}^S$. Operator T_c enjoys the following properties:*

(a) Monotonicity

$$v \geq u \implies T_c v \geq T_c u.$$

(b) $T_c v \leq Lv$ and if in addition $sp\left\{v\right\} \leq c$ and $v \leq Lv$ then $v \leq T_c v$.

(c) Linearity

$$\forall \lambda \in \mathbb{R},\ \ T_c\left(v + \lambda e\right) = T_c v + \lambda e.$$

(d) T_c is non-expansive both in span semi-norm and ℓ_∞-norm

$$sp\left\{T_c v - T_c u\right\} \leq sp\left\{v - u\right\} \ \ \textit{and} \ \ \left\|T_c v - T_c u\right\|_\infty \leq \left\|v - u\right\|_\infty.$$

Moreover, if L is a γ-span contraction then T_c is also a γ-span contraction (Lem. 7).

Proof. We rely on the properties proved in Lem. 15.

(a) The monotonicity of T_c is a direct consequence of the monotonicity of both L (Puterman, 1994) and Γ_c (property *(a)* of Lem. 15) and the fact that monotonicity is preserved by composition.

(b) $T_c v = \Gamma_c L v \leq Lv$ by definition of Γ_c. If $v \leq Lv$ then using the fact that Γ_c is monotone we have $\Gamma_c v \leq \Gamma_c L v = T_c v$ and if we assume that $sp\left\{v\right\} \leq c$ then $v = \Gamma_c v$ and so $v \leq T_c v$.

(c) The linearity of T_c is a direct consequence of the linearity of both L (Puterman, 1994) and Γ_c (property *(e)* of Lem. 15) and the fact that linearity is preserved by composition.

(d) Using the fact that L (Puterman, 1994) and Γ_c (properties *(c)* and *(d)* of Lem. 15) are non-expansive both in span semi-norm and ℓ_∞-norm we show the following:

$$sp\{T_c v - T_c u\} = sp\{\Gamma_c L v - \Gamma_c L u\} \le sp\{Lv - Lu\} \le sp\{v-u\}$$
$$\text{and } \|T_c v - T_c u\|_\infty = \|\Gamma_c L v - \Gamma_c L u\|_\infty \le \|Lv - Lu\|_\infty \le \|v-u\|_\infty$$

If L is a γ-span contraction then:

$$sp\{T_c v - T_c u\} = sp\{\Gamma_c L v - \Gamma_c L u\} \le sp\{Lv - Lu\} \le \gamma sp\{v-u\}$$

meaning that T_c is also a γ-span contraction.

□

Lem. 7 immediately follows from property *(d)* of Lem. 16.

D.3. Convergence properties of T_c (proof of Lemma 8)

In this section we provide a detailed proof of Lem. 8.

We assume that Asm. 6 holds which implies that T_c is a γ-span contraction by Lem. 7.

1. *Existence and uniqueness of the solution of optimality equation* (14)*:*
Consider the quotient vector space $W = \mathbb{R}^S/\text{Span}(e)$ where $\text{Span}(e)$ is the linear span of vector e i.e., the intersection of all vector spaces containing e: $\text{Span}(e) = \{\lambda e : \lambda \in \mathbb{R}\}$. The quotient space W is a vector space with dimension $S-1$ (it is in bijection with $\mathbb{R}^{S-1} \times \{0\}$, where one coordinate is set to 0 and the others are free real variables). Since $\text{Span}(e)$ is the null space of the *semi-norm* $sp\{\cdot\}$, then $sp\{\cdot\}$ is indeed a *norm* on W and thus $(W, sp\{\cdot\})$ is a normed vector space. The operator T_c is well-defined also on $(W, sp\{\cdot\})$ because of property *(c)* of Lem. 16 (linearity of T_c): $\forall h \in \mathbb{R}^S,\ T_c(h + \lambda e) = T_c h + \lambda e$ implying that for any given $w \in W$, the vector $T_c w \in W$ is uniquely defined (i.e., there is no ambiguity in the definition of T_c). Moreover, if $h \in \mathbb{R}^S$ maps to $w \in W$ then $T_c h \in \mathbb{R}^S$ maps to $T_c w \in W$. Since T_c is a span contraction, then T_c has a unique fixed point w^+ in W by Banach fixed-point theorem, which corresponds to the optimality equation $T_c w^+ = w^+$ (in W). Let $h^+ \in \mathbb{R}^S$ be an arbitrary (bounded) vector in the original space that maps to $w^+ \in W$. Since $T_c h^+ \in \mathbb{R}^S$ maps to $T_c w^+ \in W$ and $T_c w^+ = w^+$ we have that $T_c h^+$ and h^+ differ only by a constant vector i.e., $sp\{T_c h^+ - h^+\} = 0$ or in other words, there exists a constant $g^+ \in \mathbb{R}$ such that $T_c h^+ = h^+ + g^+ e$ which proves the existence of the solution of optimality equation (14). Any other solution $h' \in \mathbb{R}^S$ to this equation will necessarily map to $w^+ \in W$ by uniqueness of the solution in W and so $sp\{h^+ - h'\} = 0$. As a result, the fixed point property of T_c in W translates into a fixed point *up to a constant vector* in $\mathbb{R}^S$, which leads to the optimality equation $T_c h^+ = h^+ + g^+ e$ as in (14) where h^+ is defined up to a constant. Furthermore, let (g_1^+, h_1^+) and (g_2^+, h_2^+) be two solutions of (14). Since there exists a $\lambda \in \mathbb{R}$ such that $h_2^+ = h_1^+ + \lambda e$ we have

$$g_2^+ e = T_c h_2^+ - h_2^+ = T_c(h_1^+ + \lambda e) - (h_1^+ + \lambda e) = T_c h_1^+ + \lambda e - h_1^+ - \lambda e = T_c h_1^+ - h_1^+ = g_1^+ e,$$

where we used property *(c)* of Lemma 16, which leads to $g_1^+ = g_2^+$ and thus the uniqueness of g^+ in (14).

2. *Convergence of (relative) value iteration:*
Fix an arbitrary state $\overline{s} \in \mathcal{S}$ and any initial vector $v_0 = v \in \mathbb{R}^S$, the relative value iteration algorithm implemented by ScOpt proceeds through iterations as

$$v_n = T_c v_{n-1} - (T_c v_{n-1})(\overline{s})e = T_c^n v - (T_c^n v)(\overline{s})e. \tag{29}$$

The last equality in (29) can be proved by induction on $n \ge 1$: it is trivially true for $n = 1$ and assuming that for a given $n \ge 1$ it holds that $v_n = T_c^n v - (T_c^n v)(\overline{s})e$, then

$$\begin{aligned} v_{n+1} = T_c v_n - (T_c v_n)(\overline{s})e &= T_c(T_c^n v - (T_c^n v)(\overline{s})e) - (T_c(T_c^n v - (T_c^n v)(\overline{s})e))(\overline{s})e \\ &= T_c^{n+1} v - (T_c^n v)(\overline{s})e - \left(T_c^{n+1} v\right)(\overline{s})e + (T_c^n v)(\overline{s})e \\ &= T_c^{n+1} v - \left(T_c^{n+1} v\right)(\overline{s})e \end{aligned}$$

where we used the linearity of T_c (property *(c)* of Lem. 16).

Denote by $q_n = v_{n+1} - v_n$. Since $v_{n+1}(\overline{s}) = v_n(\overline{s}) = 0$, then $q_n(\overline{s}) = 0$ and the absolute value of any component $q_n(s)$ can be upper-bounded by its span. [8] As a result, we have

$$|q_n(s)| \leq \|q_n\|_\infty \leq sp\{q_n\}.$$

Using the span contraction property of T_c we have that

$$\begin{aligned}
|v_{n+1}(s) - v_n(s)| &\leq sp\{v_{n+1} - v_n\} = sp\{T_c v_n - (T_c v_n)e - T_c v_{n-1} + (T_c v_{n-1})e\} \\
&\overset{(a)}{=} sp\{T_c v_n - T_c v_{n-1}\} \leq \gamma sp\{v_n - v_{n-1}\} = \gamma sp\{q_{n-1}\} \\
&\overset{(b)}{\leq} \gamma^n sp\{T_c v - v\},
\end{aligned}$$

where *(a)* follows from the fact that $sp\{f\} = sp\{f + \lambda e\}$ for any $\lambda \in \mathbb{R}$ and *(b)* is obtained by iterating the first inequality. Since $sp\{T_c v - v\}$ is bounded and $\gamma < 1$, we can conclude that $\{v_{n+1} - v_n\}_n$ is a convergent sequence. Now we show that $\{v_n\}_n$ is a Cauchy sequence. Let $m > n$, then the following inequalities hold

$$\begin{aligned}
|v_m(s) - v_n(s)| &\leq sp\{v_m - v_n\} = sp\{v_m - v_{m-1} + v_{m-1} - v_{m-2} + \ldots + v_{n+1} - v_n\} \\
&\leq sp\{v_m - v_{m-1}\} + sp\{v_{m-1} - v_{m-2}\} + \ldots + sp\{v_{n+1} - v_n\} \\
&\overset{(a)}{\leq} \gamma^{m-1} sp\{T_c v - v\} + \gamma^{m-2} sp\{T_c v - v\} + \ldots + \gamma^n sp\{T_c v - v\} \\
&= \gamma^n sp\{T_c v - v\} \sum_{k=0}^{m-n-1} \gamma^k \leq \gamma^n sp\{T_c v - v\} \sum_{k=0}^{\infty} \gamma^k = \frac{\gamma^n}{1-\gamma} sp\{T_c v - v\},
\end{aligned} \tag{30}$$

where *(a)* is the application of the previous inequality $sp\{v_{n+1} - v_n\} \leq \gamma^n sp\{T_c v - v\}$. Since $\gamma < 1$, for any arbitrary $\varepsilon > 0$, there exists a N_ε, so that for any $m > n > N_\varepsilon$, $\|v_m - v_n\|_\infty \leq \varepsilon$. As a result $\{v_n\}_n$ is a Cauchy sequence and since $(\mathbb{R}^S, \|\cdot\|_\infty)$ is a Banach space, v_n converges to a vector that we denote by $h(v, \overline{s})$. We now show that $h(v, \overline{s})$ satisfies the optimality equation. Using property *(c)* in Lem. 16 and Eq. 29, we can write

$$T_c v_n = T_c\big(T_c^n v - (T_c^n v)(\overline{s})e\big) = T_c^{n+1} v - (T_c^n v)(\overline{s})e = v_{n+1} + \big(T_c^{n+1} v - T_c^n v\big)(\overline{s})e. \tag{31}$$

Then,

$$sp\{T_c v_n - v_{n+1}\} = sp\big\{\big(T_c^{n+1} v - T_c^n v)\big)(\overline{s})e\big\} = 0. \tag{32}$$

By continuity of the semi-norm $sp\{\cdot\}$ and uniqueness of the limit this implies that

$$\lim_{n \to +\infty} sp\{T_c v_n - v_{n+1}\} = sp\{T_c h(v, \overline{s}) - h(v, \overline{s})\} = 0,$$

where we used the fact that the sequences v_n and v_{n+1} converge to $h(v, \overline{s})$. Since $T_c h(v, \overline{s}) - h(v, \overline{s})$ has zero span, we conclude that there exists a constant value $g \in \mathbb{R}$ such that $T_c h(v, \overline{s}) = h(v, \overline{s}) + ge$, which is indeed optimality equation (14) and by uniqueness of the solution, $g = g^+$. This proves that relative value iteration using T_c does converge to a solution of the optimality equation.

Alternatively, we can prove that standard value iteration converges in the sense that $\lim_{n \to +\infty} T_c^{n+1} v - T_c^n v = g^+ e$. Using the continuity of T_c and the fact that $\lim_{n \to +\infty} v_n = \lim_{n \to +\infty} v_{n+1} = h(v, \overline{s})$, we can write

$$\lim_{n \to +\infty} T_c v_n - v_{n+1} = T_c h(v, \overline{s}) - h(v, \overline{s}) = g^+ e. \tag{33}$$

Using the definition of relative value iteration ($v_n = T_c^n v - T_c^n v(\overline{s})e$) and the linearity of T_c we have

$$\begin{aligned}
T_c^{n+1} v - T_c^n v &= T_c(v_n + (T_c^n v)(\overline{s})e) - (v_n + (T_c^n v)(\overline{s})e) \\
&= T_c v_n - v_n \\
&= \underbrace{v_{n+1} - v_n}_{\underset{n \to +\infty}{\longrightarrow} 0} + \underbrace{T_c v_n - v_{n+1}}_{\underset{n \to +\infty}{\longrightarrow} g^+ e} \underset{n \to +\infty}{\longrightarrow} g^+ e,
\end{aligned}$$

where the limits rely on the convergence of v_n and Eq. 33.

3. *Dominance $g^+ \geq g_c^*$:*

[8] Since q_n takes the value 0 in $\overline{s}$, there are only three possible scenarios: *1)* q_n is non-negative (with 0 included) and then $\max\{q_n\} = sp\{q_n\}$, *2)* q_n is non-positive (with 0 included) and then $\max\{|q_n|\} = 0 - \min\{q_n\} = sp\{q_n\}$, *3)* q_n has both positive and negative values and then both its maximal and minimal value are smaller than the span.

Let $\pi = d^\infty \in \Pi_c$ be a policy with constant gain and bounded bias span. The evaluation Bellman equation gives $L_d h^\pi = h^\pi + g^\pi e$ and $sp\{L_d h^\pi\} = sp\{h^\pi\} \leq c$. This implies that $d \in D(c, h^\pi) \neq \emptyset$ and so from Lemma 5 we have

$$T_c h^\pi \geq h^\pi + g^\pi e.$$

By monotonicity and linearity of T_c (properties *(a)* and *(c)*) of Lemma 16) we have

$$T_c h^\pi \geq h^\pi + g^\pi e \implies T_c^2 h^\pi \geq T_c(h^\pi + g^\pi e) = T_c h^\pi + g^\pi e \geq h^\pi + 2g^\pi e,$$

As a result, we can iterate the inequality and obtain for all $n \in \mathbb{N}$:

$$T_c^{n+1} h^\pi \geq T_c^n h^\pi + g^\pi e \implies \underbrace{T_c^{n+1} h^\pi - T_c^n h^\pi}_{\underset{n\to+\infty}{\longrightarrow} g^+ e} \geq g^\pi e \implies g^+ \geq g^\pi,$$

where we used property *2.* of Lem. 8 proved above. Since the inequality holds for any $\pi \in \Pi_c$, it also holds for the supremum $\sup_{\pi\in\Pi_c} g^\pi = g_c^*$ (solution to problem (7)) i.e., $g^+ \geq g_c^*$.

D.4. Approximation guarantees of SCOPT (proof of Theorem 10)

In this section we prove a slightly more general statement than Thm. 10.

We use operators G_c and N_c defined in Def. 2 and App. B respectively. We recall that when T_c is globally feasible at v, then $N_c v = T_c v$ and $G_c v = \delta_v^+$. We first slightly relax Asm. 9 (Asm. 17 below) and then prove a generalisation of Thm. 10 (Thm. 18 below).

Assumption 17. *Operator T_c is globally feasible at h^+, i.e., the decision rule $d^+ = G_c h^+$ is such that $T_c h^+ = L_{d^+} h^+ = N_c h^+ = h^+ + g^+ e$.*

Theorem 18. *Assume Asm. 6 and 17 hold and let*

$$\begin{aligned}
M_n^+ &= \frac{2\gamma^n}{1-\gamma} sp\{v_1 - v_0\} + \max_s \{T_c v_n(s) - v_n(s)\} \\
m_n^+ &= \frac{-2\gamma^n}{1-\gamma} sp\{v_1 - v_0\} + \min_s \{T_c v_n(s) - v_n(s)\} \\
M_n &= \max_s \{N_c v_n(s) - v_n(s)\} \\
m_n &= \min_s \{N_c v_n(s) - v_n(s)\},
\end{aligned}$$

and $d_n = G_c v_n$ be the decision rule computed after n iterations and $\pi_n = (d_n)^\infty$ the corresponding policy. Then we have

$$1) \left\| g^{\pi_n} - \frac{1}{2}(M_n + m_n) e \right\|_\infty \leq \frac{1}{2}(M_n - m_n) = \frac{1}{2} sp\{N_c v_n - v_n\} \tag{34}$$

$$2) \left| g^+ - \frac{1}{2}\left(M_n^+ + m_n^+\right) \right| \leq \frac{1}{2}\left(M_n^+ - m_n^+\right) = \frac{2\gamma^n}{1-\gamma} sp\{v_1 - v_0\} + \frac{1}{2} sp\{T_c v_n - v_n\} \tag{35}$$

$$3) \left\| g^+ e - g^{\pi_n} \right\|_\infty \leq \max\left\{M_n^+ - m_n,\ M_n - m_n^+\right\} \tag{36}$$

Moreover, if in addition $\pi^+ = (d^+)^\infty$ is unichain then g^+ is the solution to optimization problem (7), i.e., $g^+ = g_c^*$ and $\pi^+ \in \Pi_c^*$.

Proof. We first analyse the convergence error. Let π_n be the policy associated to the decision rule $d_n = G_c v_n$. We prove the three convergence statements of the theorem.

1. We recall that operator N_c is such that $L_{d_n} v_n = L_{G_c v_n} v_n = N_c v_n$. By definition of the gain g^{π_n}, there exists a *stationary transition matrix* $P_{d_n}^* := C\text{-}\lim_{k\to+\infty} (P_{d_n})^k$ (Puterman, 1994, Appendix A) such that $g^{\pi_n} = P_{d_n}^* r_{d_n}$. Furthermore, since $P_{d_n}^* P_{d_n} = P_{d_n}^*$, for any vector v_n, $g^{\pi_n} = P_{d_n}^*(r_{d_n} + P_{d_n} v_n - v_n) = P_{d_n}^*(L_{d_n} v_n - v_n)$. Since $\min\{L_{d_n} v_n - v_n\} e \leq L_{d_n} v_n - v_n \leq \max\{L_{d_n} v_n - v_n\} e$ and by multiplying these two inequalities by $P_{d_n}^*$ (which is a stochastic matrix) we obtain $m_n e \leq g^{\pi_n} \leq M_n e$ and the result holds.

2. Using Eq. 30 and letting $m \to \infty$, we have $\|h^+ - v_n\|_\infty \leq \frac{\gamma^n}{1-\gamma} sp\{v_1 - v_0\}$. Therefore by monotonicity and linearity

of T_c (property *(a)* and *(c)* of Lem. 16) and using optimality equation $T_c h^+ = h^+ + g^+ e$:

$$\begin{aligned}
&-\frac{\gamma^n}{1-\gamma} sp\{v_1 - v_0\} e + h^+ \le v_n \le h^+ + \frac{\gamma^n}{1-\gamma} sp\{v_1 - v_0\} e \\
\implies &-\frac{\gamma^n}{1-\gamma} sp\{v_1 - v_0\} e + h^+ + g^+ e \le T_c v_n \le h^+ + g^+ e + \frac{\gamma^n}{1-\gamma} sp\{v_1 - v_0\} e \\
\implies &-\frac{2\gamma^n}{1-\gamma} sp\{v_1 - v_0\} e + g^+ e \le T_c v_n - v_n \le g^+ e + \frac{2\gamma^n}{1-\gamma} sp\{v_1 - v_0\} e \\
\implies &-\frac{2\gamma^n}{1-\gamma} sp\{v_1 - v_0\} e + T_c v_n - v_n \le g^+ e \le T_c v_n - v_n + \frac{2\gamma^n}{1-\gamma} sp\{v_1 - v_0\} e \\
\implies &m_n^+ \le g^+ \le M_n^+
\end{aligned}$$

and the result holds.

3. The last inequality is a direct consequence of the two inequalities previously proved:

$$m_n e \le g^{\pi_n} \le M_n e \text{ and } m_n^+ \le g^+ \le M_n^+.$$

We now prove optimality. Under the global feasibility assumption at h^+ (Asm. 17), we have that there exists a decision rule d^+ and an associated policy $\pi^+ = (d^+)^\infty$ such that

$$T_c h^+ = L_{d^+} h^+ = h^+ + g^+ e, \tag{37}$$

Since (g^+, h^+) is a solution of the Bellman evaluation equations (1) associated to $\pi^+ = (d^+)^\infty$ and since by assumption π^+ is unichain, Corollary 8.2.7. of Puterman (1994) holds and so $g^+ = g^{\pi^+}$ and there exists $\lambda \in \mathbb{R}$ such that $h^{\pi^+} = h^+ + \lambda e$ implying that

$$sp\left\{h^{\pi^+}\right\} = sp\left\{h^+ + \lambda e\right\} = sp\left\{h^+\right\} = sp\left\{h^+ + g^+ e\right\} = sp\left\{T_c h^+\right\} \le c,$$

where we used the invariance of the span by translation, Eq. 37 and the definition of T_c. As a result, $\pi^+ \in \Pi_c$ and by property *3.* of Lem. 8 we can conclude that

$$g^{\pi^+} = g^+ \ge g^\pi, \quad \forall \pi \in \Pi_c,$$

which implies that $\pi^+ \in \Pi_c^*$ and $g^{\pi^+} = g_c^*$. Note that if π^+ is not unichain then we might have $sp\{h^{\pi^+}\} > sp\{h^+\}$ in which case it is possible that $\pi^+ \notin \Pi_c$ and so the result does not hold. □

We now relate Asm. 17 and Thm. 18 to (respectively) Asm. 9 and Thm. 10. As we just showed in the proof of Thm. 18, we always have $sp\{h^+\} \le c$ and therefore, whenever Asm. 9 holds, Asm. 17 holds too. As a result, if Asm. 6 also holds then the first part of Thm. 18 holds too. Moreover, if $sp\{v_0\} \le c$ it is straightforward to see that $sp\{v_n\} \le c$ for any $n \ge 1$ and so due to Asm. 9, $T_c v_n = N_c v_n$ for all $n \ge 1$. This implies that

$$\max\left\{M_n^+ - m_n,\ M_n - m_n^+\right\} = sp\{T_c v_n - v_n\} + \frac{2\gamma^n}{1-\gamma} sp\{v_1 - v_0\} = sp\{v_{n+1} - v_n\} + \frac{2\gamma^n}{1-\gamma} sp\{v_1 - v_0\}$$

and as a consequence Thm. 10 holds.

E. Modified bounded-parameter (extended) MDPs (Proof of Thm. 11)

In this section we prove Thm. 11.

We analyse separately the two modifications introduced in Def. 3 on the rewards and transition probabilities (transition "*kernel*"). For a given bounded-parameter (extended) MDP $\widetilde{\mathcal{M}}$, we denote by $\widetilde{\mathcal{M}}_\eta$ the *perturbed* bounded-parameter MDP whose transition kernel is an η-perturbation of the original one (see formal definition in Lem. 19 below), and by $\widetilde{\mathcal{M}}^\downarrow$ the *augmented* bounded-parameter MDP whose reward intervals are extended from below compared to the original ones (the maximum is not changed while the lower bound of the interval is set to zero, see formal definition in Lem. 20 below). We first prove interesting properties for operators $\widetilde{T}_c^\eta$ associated to $\widetilde{\mathcal{M}}_\eta$ (Lem. 19) and $\widetilde{T}_c^\downarrow$ associated to $\widetilde{\mathcal{M}}^\downarrow$ (Lem. 20). We then consider the MDP $\widetilde{\mathcal{M}}_\eta^\downarrow$ that is both *augmented* and *perturbed* and we present the properties of the corresponding operator $\widetilde{T}_c^{\eta,\downarrow}$ in Thm. 21. Note that in Sec. 6 we called the augmented and perturbed MDP "*modified MDP*" for simplicity, and we

used the notation $\widetilde{\mathcal{M}}^{\ddagger}$ instead of $\widetilde{\mathcal{M}}_{\eta}^{\downarrow}$ for clarity. Thm. 11 in Sec. 6 is thus equivalent to Thm. 21 stated below.

In the following, for any closed interval $[a,b] \subset \mathbb{R}$ we use the notations $\min\{[a,b]\} := a$ and $\max\{[a,b]\} := b$.

Lemma 19. *Let $\widetilde{\mathcal{M}}$ be a bounded-parameter MDP defined for all $s, s' \in \mathcal{S}$ and all $a \in \mathcal{A}_s$ by*

$$r(s,a) \in B_r(s,a) \;\; \textit{and} \;\; p(s'|s,a) \in B_p(s,a,s')$$

*where $\mathcal{S}$ and $\mathcal{A}_s$ are finite, $B_r(s,a)$ and $B_p(s,a,s')$ are closed intervals of $[0, r_{\max}]$ and $[0,1]$ respectively. Let $1 \geq \eta > 0$ and $\overline{s} \in \mathcal{S}$ and consider the "*perturbed*" bounded-parameter MDP $\widetilde{\mathcal{M}}_{\eta}$ defined $\forall s, s' \in \mathcal{S}$ and $\forall a \in \mathcal{A}_s$ by:*

$$B_r^{\eta}(s,a) = B_r(s,a) \;\; \textit{and} \;\; B_p^{\eta}(s,a,s') = \begin{cases} B_p(s,a,s') \;\; \textit{if} \;\; s' \neq \overline{s} \\ \textit{and} \;\; B_p(s,a,\overline{s}) \cap [\eta, 1] \;\; \textit{if} \;\; s' = \overline{s} \end{cases}$$

where we assume that η is small enough so that $\forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}_s$,

1. $B_p(s,a,\overline{s}) \cap [\eta, 1] \neq \emptyset$
2. $\sum_{s' \in \mathcal{S}} \min\{B_p^{\eta}(s,a,s')\} \leq 1$ *and* $\sum_{s' \in \mathcal{S}} \max\{B_p(s,a,s')\} \geq 1$

If $\widetilde{L}$ denotes the optimal Bellman operator of $\widetilde{\mathcal{M}}$ and $\widetilde{L}_{\eta}$ the optimal Bellman operator of $\widetilde{\mathcal{M}}_{\eta}$, then $\forall v \in \mathbb{R}^S$:

$$\left\| \widetilde{L} v - \widetilde{L}_{\eta} v \right\|_{\infty} \leq sp\{v\}\, \eta \tag{38}$$

Moreover, $\widetilde{L}_{\eta}$ is a γ-span contraction with $\gamma \leq 1 - \eta < 1$ and $\widetilde{\mathcal{M}}_{\eta}$ is unichain.

Proof. For all states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}_s$ we use the following notations

$$\widetilde{r}(s,a) := \max\{B_r(s,a)\} \;\; \text{and} \;\; \widetilde{p}(\cdot|s,a) := \underset{p(s') \in B_p(s,a,s')}{\arg\max} \sum_{s' \in \mathcal{S}} p(s') v(s') \tag{39}$$

and we define $\widetilde{r}_{\eta}(s,a)$ and $\widetilde{p}_{\eta}(\cdot|s,a)$ similarly with $B_r(s,a)$ and $B_p(s,a,s')$ replaced by $B_r^{\eta}(s,a)$ and $B_p^{\eta}(s,a,s')$.

$$\begin{aligned} \forall s \in \mathcal{S}, \; |Lv(s) - \widetilde{L}_{\eta} v(s)| &= \left| \max_{a \in \mathcal{A}_s} \left\{ \widetilde{r}(s,a) + \sum_{s' \in \mathcal{S}} \widetilde{p}(s'|s,a) v(s') \right\} - \max_{a \in \mathcal{A}_s} \left\{ \widetilde{r}_{\eta}(s,a) + \sum_{s' \in \mathcal{S}} \widetilde{p}_{\eta}(s'|s,a) v(s') \right\} \right| \\ &\leq \max_{a \in \mathcal{A}_s} \left| \underbrace{\widetilde{r}(s,a) - \widetilde{r}_{\eta}(s,a)}_{=0} + \sum_{s' \in \mathcal{S}} (\widetilde{p}(s'|s,a) - \widetilde{p}_{\eta}(s'|s,a)) v(s') \right| \end{aligned}$$

where we used the fact that $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$ and $B_r^{\eta}(s,a) = B_r(s,a)$ by definition. Since $\widetilde{p}(\cdot|s,a)$ and $\widetilde{p}_{\eta}(\cdot|s,a)$ are probability distributions (i.e., sum to 1), for any real λ:

$$\sum_{s' \in \mathcal{S}} (\widetilde{p}(s'|s,a) - \widetilde{p}_{\eta}(s'|s,a)) v(s') = \sum_{s' \in \mathcal{S}} (\widetilde{p}(s'|s,a) - \widetilde{p}_{\eta}(s'|s,a))(v(s') + \lambda)$$

Taking $\lambda = -(\max_s v(s) + \min_s v(s))/2$ we obtain:

$$\begin{aligned} \forall s \in \mathcal{S}, \; |Lv(s) - \widetilde{L}_{\eta} v(s)| &\leq \max_{a \in \mathcal{A}_s} \sum_{s' \in \mathcal{S}} |\widetilde{p}(s'|s,a) - \widetilde{p}_{\eta}(s'|s,a)| \cdot \max_s \{v(s) + \lambda\} \\ &= \max_{a \in \mathcal{A}_s} \|\widetilde{p}(\cdot|s,a) - \widetilde{p}_{\eta}(\cdot|s,a)\|_1 \cdot (\max_s \{v(s)\} + \lambda) \\ &= \max_{a \in \mathcal{A}_s} \|\widetilde{p}(\cdot|s,a) - \widetilde{p}_{\eta}(\cdot|s,a)\|_1 \cdot \frac{sp\{v\}}{2} \end{aligned}$$

We now need to upper-bound $\|\widetilde{p}(\cdot|s,a) - \widetilde{p}_{\eta}(\cdot|s,a)\|_1$. $\widetilde{p}(\cdot|s,a)$ and $\widetilde{p}_{\eta}(\cdot|s,a)$ can be computed using the following procedure (Dann and Brunskill, 2015, Appendix A):

1. Assume without loss of generality that the coordinates of v are sorted in decreasing order: $v(s_1) \geq v(s_2) \geq ... \geq v(s_n)$
2. Initialise $\widetilde{p}^0(s'|s,a) = \min\{B_p(s,a,s')\}$, $\Delta^0 = 1 - \sum_{s' \in \mathcal{S}} \widetilde{p}^0(s'|s,a)$ for all $s' \in \mathcal{S}$, and $i = 1$
3. While $\Delta^{i-1} > 0$ do
   - $\delta^i \leftarrow \min\left\{\Delta^{i-1}; \max\{B_p(s,a,s_i)\} - \widetilde{p}^{i-1}(s_i|s,a)\right\}$

- $\widetilde{p}^i(s_i|s,a) \leftarrow \widetilde{p}^{i-1}(s_i|s,a) + \delta^i$
- $\Delta^i \leftarrow \Delta^{i-1} - \delta^i$
- $i \leftarrow i+1$

4. Return $\widetilde{p}(\cdot|s,a) = \widetilde{p}^{i-1}(\cdot|s,a)$

Let's now show that at any iteration of the above procedure $\widetilde{p}(\cdot|s,a)$ and $\widetilde{p}_\eta(\cdot|s,a)$ are at most 2η-far in ℓ_1-norm. Notice that at the end of iteration i, the vector $\widetilde{p}^i(\cdot|s,a)$ differs from $\widetilde{p}^{i-1}(\cdot|s,a)$ only in state s_i. In the following we use index η to denote the quantities obtained when the above procedure is applied with $B_p^\eta(s,a,s')$ instead of $B_p(s,a,s')$ (the output is then $\widetilde{p}_\eta(\cdot|s,a)$). The conditions $B_p(s,a,\overline{s}) \cap [\eta,1] \neq \emptyset$, $\sum_{s'\in\mathcal{S}} \min\{B_p^\eta(s,a,s')\} \leq 1$ and $\sum_{s'\in\mathcal{S}} \max\{B_p(s,a,s')\} \geq 1$ ensure that the procedure doesn't stop prematurely when $B_p^\eta(s,a,s')$ replaces $B_p(s,a,s')$. Indeed, when they hold there exists a vector p satisfying $\sum_{s'\in\mathcal{S}} p(s') = 1$ and $\forall s' \in \mathcal{S},\ p(s') \in B_p^\eta(s,a,s')$.

- $\boldsymbol{i=0}$ **(initialization):** By definition, for any $s' \neq \overline{s}$, $B_p^\eta(s,a,s') = B_p(s,a,s')$ implying that $\widetilde{p}^0(s'|s,a) = \widetilde{p}_\eta^0(s'|s,a)$ and moreover $\eta \leq \min\{B_p^\eta(s,a,\overline{s})\} \leq \eta + \min\{B_p(s,a,\overline{s})\}$ implying that $\eta \leq \widetilde{p}_\eta^0(\overline{s}|s,a) \leq \eta + \widetilde{p}^0(\overline{s}|s,a)$ and thus $\Delta_\eta^0 + \eta \geq \Delta^0 \geq \Delta_\eta^0$ and

$$\|\widetilde{p}^0(\cdot|s,a) - \widetilde{p}_\eta^0(\cdot|s,a)\|_1 = |\widetilde{p}^0(\overline{s}|s,a) - \widetilde{p}_\eta^0(\overline{s}|s,a)| \leq \eta$$

When the procedure to compute $\widetilde{p}_\eta(\cdot|s,a)$ stops, there are only two possibilities: state $\overline{s}$ is updated either before or after $\Delta_\eta = 0$. In the following we analyze separately the two cases.

- $\boldsymbol{\Delta_\eta = 0}$ **occurs first:** i.e., $\Delta_\eta^i = 0$ and for $k = 1, ..., i$, $s_k \neq \overline{s}$ and $\Delta_\eta^{k-1} > 0$.
  As a consequence, we have that $\widetilde{p}_\eta(\cdot|s,a) = \widetilde{p}_\eta^i(\cdot|s,a)$ and by triangle inequality:

$$\|\widetilde{p}(\cdot|s,a) - \widetilde{p}_\eta(\cdot|s,a)\|_1 \leq \|\widetilde{p}^i(\cdot|s,a) - \widetilde{p}_\eta^i(\cdot|s,a)\|_1 + \|\widetilde{p}^i(\cdot|s,a) - \widetilde{p}(\cdot|s,a)\|_1 \tag{40}$$

  By assumption for $k = 1, ..., i-1$, $\Delta_\eta^{k-1} > 0$, we have that $\widetilde{p}_\eta^k(s_k|s,a) = \max\{B_p^\eta(s,a,s_k)\} = \max\{B_p(s,a,s_k)\}$. Moreover, for all $k = 1, ..., i-1$, $s_k \neq \overline{s}$ and thus by trivial induction we have that $\Delta^k - \Delta_\eta^k = \Delta^0 - \Delta_\eta^0$, $\Delta^{k-1} \geq \Delta_\eta^{k-1} > 0$ and $\widetilde{p}^k(s_k|s,a) = \max\{B_p(s,a,s_k)\} = \widetilde{p}_\eta^k(s_k|s,a)$. Then:

$$\forall k = 1, ..., i-1, \quad \|\widetilde{p}^k(\cdot|s,a) - \widetilde{p}_\eta^k(\cdot|s,a)\|_1 = \Delta^0 - \Delta_\eta^0 \leq \eta$$

  and thus $\|\widetilde{p}^i(\cdot|s,a) - \widetilde{p}_\eta^i(\cdot|s,a)\|_1 = \Delta^0 - \Delta_\eta^0 + \delta^i - \delta_\eta^i$. Since $\|\widetilde{p}^i(\cdot|s,a) - \widetilde{p}(\cdot|s,a)\|_1 = \Delta^i$, after incorporating everything into (40) we obtain:

$$\begin{aligned}
\|\widetilde{p}(\cdot|s,a) - \widetilde{p}_\eta(\cdot|s,a)\|_1 &\leq \Delta^0 - \Delta_\eta^0 - \delta_\eta^i + \underbrace{\delta^i + \Delta^i}_{=\Delta^{i-1}} \\
&= \Delta^0 - \Delta_\eta^0 + \underbrace{\Delta^{i-1}}_{=\Delta_\eta^{i-1}+\Delta^0-\Delta_\eta^0} - \delta_\eta^i = 2(\Delta^0 - \Delta_\eta^0) + \underbrace{\Delta_\eta^{i-1} - \delta_\eta^i}_{=\Delta_\eta^i = 0} \\
&= 2(\Delta^0 - \Delta_\eta^0) \leq 2\eta
\end{aligned}$$

- $\boldsymbol{\overline{s}}$ **is updated first:** i.e., $s_i = \overline{s}$ and for $k = 1, ..., i-1$, $s_k \neq \overline{s}$ and $\Delta_\eta^k > 0$. By trivial induction we have that for all $k < i$, $\Delta^k - \Delta_\eta^k = \Delta^0 - \Delta_\eta^0$ and $\widetilde{p}^k(s_k|s,a) = \max\{B_p(s,a,s_k)\} = \max\{B_p^\eta(s,a,s_k)\} = \widetilde{p}_\eta^k(s_k|s,a)$. Since δ_η^i is defined as the minimum between two values, there are only two possible cases:
  - If $\delta_\eta^i = \max\{B_p^\eta(s,a,s_i)\} - \widetilde{p}_\eta^{i-1}(s_i|s,a) \leq \Delta_\eta^{i-1}$ we have that

$$\begin{aligned}
\max\{B_p(s,a,s_i)\} - \widetilde{p}^{i-1}(s_i|s,a) &= \max\{B_p^\eta(s,a,s_i)\} - \widetilde{p}_\eta^{i-1}(s_i|s,a) + \Delta^0 - \Delta_\eta^0 \\
&\leq \Delta_\eta^{i-1} + \Delta^0 - \Delta_\eta^0 = \Delta^{i-1}
\end{aligned}$$

    which implies $\delta^i = \max\{B_p(s,a,s_i)\} - \widetilde{p}^{i-1}(s_i|s,a)$. As a consequence,
    * $\widetilde{p}^i(s_i|s,a) = \max\{B_p(s,a,s_i)\} = \max\{B_p^\eta(s,a,s_i)\} = \widetilde{p}_\eta^i(s_i|s,a)$
    * $\Delta^i = \Delta_\eta^i$

    Thus for $k > i$, $\widetilde{p}^k(s_k|s,a) = \widetilde{p}_\eta^k(s_k|s,a)$ and so $\widetilde{p}(\cdot|s,a) = \widetilde{p}_\eta(\cdot|s,a)$.

– If $\delta_\eta^i = \Delta_\eta^{i-1} \leq \max\{B_p^\eta(s,a,s_i)\} - \widetilde{p}_\eta^{i-1}(s_i|s,a)$ we have that

$$\begin{aligned}\widetilde{p}_\eta^{i-1}(s_i|s,a) = \widetilde{p}_\eta^0(s_i|s,a) = \widetilde{p}_\eta^0(\overline{s}|s,a) = \min\{B_p^\eta(s,a,\overline{s})\} &= \min\{B_p(s,a,\overline{s})\} + \Delta^0 - \Delta_\eta^0 \\ &= \widetilde{p}^0(\overline{s}|s,a) + \Delta^0 - \Delta_\eta^0 \\ &= \widetilde{p}^0(s_i|s,a) + \Delta^0 - \Delta_\eta^0 \\ &= \widetilde{p}^{i-1}(s_i|s,a) + \Delta^0 - \Delta_\eta^0\end{aligned}$$

implying that

$$\begin{aligned}\Delta^{i-1} = \Delta_\eta^{i-1} + \Delta^0 - \Delta_\eta^0 &\leq \max\{B_p^\eta(s,a,s_i)\} - \widetilde{p}_\eta^{i-1}(s_i|s,a) + \Delta^0 - \Delta_\eta^0 \\ &= \max\{B_p(s,a,s_i)\} - \widetilde{p}^{i-1}(s_i|s,a)\end{aligned}$$

which implies $\delta^i = \Delta^{i-1}$. As a consequence,

$* \ \widetilde{p}^i(s_i|s,a) = \widetilde{p}^{i-1}(s_i|s,a) + \delta^i = \underbrace{\widetilde{p}^{i-1}(s_i|s,a) + \Delta^0 - \Delta_\eta^0}_{\widetilde{p}_\eta^{i-1}(\overline{s}|s,a)} + \Delta_\eta^{i-1} = \widetilde{p}_\eta^i(s_i|s,a)$

$* \ \Delta^i = \Delta_\eta^i = 0$

Thus for $k > i$, $\widetilde{p}^k(s_k|s,a) = \widetilde{p}_\eta^k(s_k|s,a)$ and so $\widetilde{p}(\cdot|s,a) = \widetilde{p}_\eta(\cdot|s,a)$.

In conclusion: $\max_{a\in\mathcal{A}_s} \|\widetilde{p}(\cdot|s,a) - \widetilde{p}_\eta(\cdot|s,a)\|_1 \leq 2\eta$ implying $\|Lv - \widetilde{L}_\eta v\|_\infty \leq sp\{v\}\,\eta$.

From (Puterman, 1994, Theorem 6.6.6), we know that $\widetilde{L}_\eta$ is Lipschitz continuous in span semi-norm with Lipschitz constant:

$$\begin{aligned}\gamma &= 1 - \min_{s\in\mathcal{S},u\in\mathcal{S},a\in\mathcal{A}_s,b\in\mathcal{A}_u} \min_{\widetilde{p}_\eta,\overline{p}_\eta} \left\{ \sum_{j\in\mathcal{S}} \min\left\{\widetilde{p}_\eta(j|s,a), \overline{p}_\eta(j|u,b)\right\} \right\} \\ &= 1 - \min_{s\in\mathcal{S},u\in\mathcal{S},a\in\mathcal{A}_s,b\in\mathcal{A}_u} \min_{\widetilde{p}_\eta,\overline{p}_\eta} \left\{ \underbrace{\sum_{j\neq\overline{s}} \min\left\{\widetilde{p}_\eta(j|s,a), \overline{p}_\eta(j|u,b)\right\}}_{\geq 0} + \underbrace{\min\left\{\widetilde{p}_\eta(\overline{s}|s,a), \overline{p}_\eta(\overline{s}|u,b)\right\}}_{\geq\eta} \right\} \leq 1-\eta < 1\end{aligned}$$

Thus $\widetilde{L}_\eta$ is a γ-span-contraction with $\gamma \leq 1-\eta < 1$. The term γ is often referred to as "*ergodic coefficient*" in the literature.

Finally, by definition of $\widetilde{\mathcal{M}}_\eta$, for any decision rule $d \in D^{\text{MR}}$ and for any state $s \in \mathcal{S}$: $p(\overline{s}|s,d(s)) > 0$. Assume that the policy $\pi = d^\infty$ associated to d has more than one *recurrent class* and pick $s_1, s_2 \in \mathcal{S}$ belonging to two different recurrent classes. Since, $p(\overline{s}|s_1,d(s_1)) > 0$ and $p(\overline{s}|s_2,d(s_2)) > 0$, necessarily $\overline{s}$ must belong to both recurrent classes which is impossible as two distinct recurrent classes have disjoint state spaces by definition. Therefore π is unichain. Since π was chosen arbitrarily, $\widetilde{\mathcal{M}}_\eta$ is unichain which concludes the proof. □

We now consider the perturbation of the reward intervals.

Lemma 20. *Let $\widetilde{\mathcal{M}}$ be a bounded-parameter MDP defined $\forall s,s' \in \mathcal{S}$ and $\forall a \in \mathcal{A}_s$ by:*

$$r(s,a) \in B_r(s,a) \ \text{ and } \ p(s'|s,a) \in B_p(s,a,s')$$

where $\mathcal{S}$ and $\mathcal{A}_s$ are finite, $B_r(s,a)$ and $B_p(s,a,s')$ are closed intervals of $[0, r_{\max}]$ and $[0,1]$ respectively. Consider the "augmented" bounded-parameter MDP $\widetilde{\mathcal{M}}^\downarrow$ defined $\forall s,s' \in \mathcal{S}$ and $\forall a \in \mathcal{A}_s$ by:

$$B_r^\downarrow(s,a) = [r_{\min}, \max\{B_r(s,a)\}] \ \text{ and } \ B_p^\downarrow(s,a,s') = B_p(s,a,s')$$

Let $\widetilde{L}_d$ ($\widetilde{L}$) and $\widetilde{L}_d^\downarrow$ ($\widetilde{L}^\downarrow$) be the (optimal) Bellman operators of $\widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{M}}^\downarrow$, respectively. Then

1. *$\forall v \in \mathbb{R}^S$, $\widetilde{L}v = \widetilde{L}^\downarrow v$ and $\widetilde{T}_c v = \widetilde{T}_c^\downarrow v$.*
2. *$\forall v \in \mathbb{R}^S$ s.t. $sp\{v\} \leq c$, $\widetilde{D}^\downarrow(c,v) = \{d \in D^{MR}(\widetilde{\mathcal{M}}^\downarrow) \mid sp\{\widetilde{L}_d^\downarrow v\} \leq c\} \neq \emptyset$*

Proof. For all states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}_s$ we use the following notations

$$\widetilde{r}(s,a) := \max\{B_r(s,a)\} \text{ and } \widetilde{p}(\cdot|s,a) := \underset{p(s')\in B_p(s,a,s')}{\arg\max} \sum_{s'\in\mathcal{S}} p(s')v(s') \tag{41}$$

$$\underset{\sim}{r}(s,a) := \min\{B_r(s,a)\} \text{ and } \underset{\sim}{p}(\cdot|s,a) := \underset{p(s')\in B_p(s,a,s')}{\arg\min} \sum_{s'\in\mathcal{S}} p(s')v(s') \tag{42}$$

and we define $\widetilde{r}_\downarrow(s,a)$, $\underset{\sim}{r}_\downarrow(s,a)$ and $\widetilde{p}_\downarrow(\cdot|s,a)$, $\underset{\sim}{p}_\downarrow(s,a)$ similarly with $B_r(s,a)$ and $B_p(s,a,s')$ replaced by $B_r^\downarrow(s,a)$ and $B_p^\downarrow(s,a,s')$.

Notice that the bounded-parameter MDP $\widetilde{\mathcal{M}}^\downarrow$ is just augmented from below, i.e., the maximum value of the reward is not altered: $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$, $\widetilde{r}(s,a) = \widetilde{r}_\downarrow(s,a)$. Moreover, by definition: $\forall s,s' \in \mathcal{S},\ \forall a \in \mathcal{A},\ \widetilde{p}(s'|s,a) = \widetilde{p}_\downarrow(s'|s,a)$. As a consequence, $\forall v \in \mathbb{R}^S$ and $\forall s \in \mathcal{S}$, $\widetilde{L}v(s) = \max_{a\in\mathcal{A}_s}\left\{\widetilde{r}(s,a) + \sum_{s'\in\mathcal{S}} \widetilde{p}(s'|s,a)v(s')\right\} = \widetilde{L}^\downarrow v(s)$. Since $\widetilde{T}_c v = \Gamma_c \widetilde{L} v$ and $\widetilde{T}_c^\downarrow v = \Gamma_c \widetilde{L}^\downarrow v$ by definition, it follows that $\widetilde{T}_c v = \widetilde{T}_c^\downarrow v$

To prove the second statement, for any $v \in \mathbb{R}^S$ we define $\delta_v \in D^{\text{MR}}(\widetilde{\mathcal{M}}^\downarrow)$ achieving the component-wise minimal value of $\widetilde{L}_d^\downarrow v$ for $d \in D^{\text{MR}}(\widetilde{\mathcal{M}}^\downarrow)$. Formally:

$$\delta_v := \underset{d\in D^{\text{MR}}(\widetilde{\mathcal{M}}^\downarrow)}{\arg\min} \left\{\widetilde{L}_d^\downarrow v\right\} \implies \forall s \in \mathcal{S},\ L_{\delta_v} v(s) = \min_{a\in\mathcal{A}_s} \left\{ \underbrace{\underset{\sim}{r}_\downarrow(s,a)}_{=r_{\min}} + \sum_{s'\in\mathcal{S}} \underset{\sim}{p}_\downarrow(s'|s,a)v(s') \right\}.$$

As a consequence, if v satisfies $sp\{v\} \le c$ then:

$$sp\left\{\widetilde{L}_{\delta_v}^\downarrow v\right\} = sp\left\{r_{\min}\boldsymbol{e} + \underset{\sim}{P}_{\delta_v}^\downarrow v\right\} = sp\left\{\underset{\sim}{P}_{\delta_v}^\downarrow v\right\} \le sp\{v\} \le c$$

where we used (Puterman, 1994, Poposition 6.6.1) applied to the stochastic matrix $\underset{\sim}{P}_{\delta_v}^\downarrow$. Then, $\delta_v \in \widetilde{D}^\downarrow(c,v) \neq \emptyset$. □

We can finally "merge" Lem. 19 and 20 and provide properties for operator $T_c^{\eta,\downarrow}$ associated to the augmented and perturbed bounded-parameter MDP $\widetilde{\mathcal{M}}_\eta^\downarrow$.

Theorem 21 (Equivalent to Thm. 11)**.** *Let $\widetilde{\mathcal{M}}$ be a bounded-parameter MDP and $\widetilde{\mathcal{M}}_\eta^\downarrow$ be the associated both "augmented" and "perturbed" bounded-parameter MDP (see Lem. 19 and 20). Then,*

1. *Lem. 8 applies to $\widetilde{T}_c^{\eta,\downarrow}$. Denote by (g^+, h^+) a solution to equation (14) for $\widetilde{T}_c^{\eta,\downarrow}$.*
2. *$\widetilde{\mathcal{M}}_\eta^\downarrow$ is unichain and Thm. 10 applies to $\widetilde{T}_c^{\eta,\downarrow}$. Denote by $\pi^+ \in \Pi^{SR}(\widetilde{\mathcal{M}}_\eta^\downarrow)$ any policy achieving:*
$$\widetilde{L}_{\pi^+}^{\eta,\downarrow} h^+ = \widetilde{T}_c^{\eta,\downarrow} h^+ = \widetilde{N}_c^{\eta,\downarrow} h^+ = h^+ + g^+ e.$$
3. *$\forall \mu \in \Pi_c(\widetilde{\mathcal{M}}),\quad g^+ = g_{\widetilde{\mathcal{M}}_\eta^\downarrow}^{\pi^+} = g_c^*(\widetilde{\mathcal{M}}_\eta^\downarrow) \ge g_{\widetilde{\mathcal{M}}}^{\mu\infty} - \eta c$.*

Proof. Lem. 19 states that $\widetilde{L}_\eta$ is a γ-span contraction ($\gamma < 1$). Since for any $v \in \mathbb{R}^S$, $\widetilde{L}_\eta^\downarrow v = \widetilde{L}_\eta v$ (property *1.* of Lem. 20), $\widetilde{L}_\eta^\downarrow$ is also a γ-span contraction. As a consequence, Asm. 6 holds and so Lem. 8 applies to $\widetilde{T}_c^{\eta,\downarrow}$ thus proving the first statement.

To prove the second statement, notice that $\widetilde{L}_\eta^\downarrow$ satisfies Asm. 9 due to property *2.* of Lem. 20 and Lem. 5. Moreover, $\widetilde{\mathcal{M}}_\eta^\downarrow$ is unichain by Lem. 19 meaning that all policies (and in particular π^+) are unichain. Therefore, Thm. 10 applies to $\widetilde{T}_c^{\eta,\downarrow}$.

Finally, we prove the third statement. We first prove by induction that the two sequences $(v_n)_{n\in\mathbb{N}}$ and $(\widehat{v}_n)_{n\in\mathbb{N}}$ defined by $v_0 = \widehat{v}_0$ such that $sp\{v_0\} \le c$ and for all $n \in \mathbb{N}$, $v_{n+1} = \widetilde{T}_c^\downarrow v_n$ and $\widehat{v}_{n+1} = \widetilde{T}_c^{\eta,\downarrow}\widehat{v}_n$, satisfy $\|v_n - \widehat{v}_n\|_\infty \le n\eta c$:

1. The result trivially holds for $n = 0$.

2. Assume the result holds for $n \in \mathbb{N}$. Let's show that it is also true for $n+1$:

$$\begin{aligned}\|v_{n+1} - \widehat{v}_{n+1}\|_\infty &= \left\|\widetilde{T}_c^{\downarrow} v_n - \widetilde{T}_c^{\eta,\downarrow}\widehat{v}_n\right\|_\infty = \left\|\Gamma_c \widetilde{L}^{\downarrow} v_n - \Gamma_c \widetilde{L}_\eta^{\downarrow}\widehat{v}_n\right\|_\infty \\ &\le \left\|\widetilde{L}^{\downarrow} v_n - \widetilde{L}_\eta^{\downarrow}\widehat{v}_n\right\|_\infty \\ &\le \underbrace{\left\|\widetilde{L}^{\downarrow} v_n - \widetilde{L}_\eta^{\downarrow} v_n\right\|_\infty}_{\le sp\{v_n\}\eta \le \eta c} + \underbrace{\left\|\widetilde{L}_\eta^{\downarrow} v_n - \widetilde{L}_\eta^{\downarrow}\widehat{v}_n\right\|_\infty}_{\le \|v_n - \widehat{v}_n\|_\infty \le n\eta c} \\ &\le (n+1)\eta c\end{aligned}$$

The first inequality comes from the fact that Γ_c is non-expansive (property *(d)* of Lem. 15). The second inequality is just the triangle inequality. The last inequality follows from Lem. 19, the fact that $sp\{v_n\} \le c$ by definition, the fact that $\widetilde{L}_\eta$ is non-expansive and the induction assumption. Let $\mu \in \Pi_c(\widetilde{\mathcal{M}})$ and for simplicity denote by h^μ (respectively g^μ) the bias $h^{\mu^\infty}_{\widetilde{\mathcal{M}}}$ associated to policy μ^∞ in $\widetilde{\mathcal{M}}$ (respectively the gain $g^{\mu^\infty}_{\widetilde{\mathcal{M}}}$). Since $sp\{h^\mu\} \le c$ by definition of $\Pi_c(\widetilde{\mathcal{M}})$, we can apply the result we just proved with $v_0 = \widehat{v}_0 = h^\mu$:

$$\forall n \in \mathbb{N},\ \ (\widetilde{T}_c^{\eta,\downarrow})^n h^\mu \ge (\widetilde{T}_c^{\downarrow})^n h^\mu - n\eta c e$$

where $e = (1, \dots, 1)^\intercal$ is the vector of all 1's and $(\widetilde{T}_c^{\eta,\downarrow})^n$ denotes n consecutive applications of operator $\widetilde{T}_c^{\eta,\downarrow}$. By property *1* of Lem. 20 we have that $(\widetilde{T}_c^{\downarrow})^n h^\mu = (\widetilde{T}_c)^n h^\mu$ implying that:

$$\forall n \in \mathbb{N},\ \ (\widetilde{T}_c^{\eta,\downarrow})^n h^\mu \ge (\widetilde{T}_c)^n h^\mu - n\eta c e \tag{43}$$

Now using the Bellman evaluation equation of μ we have:

$$\widetilde{L}_\mu h^\mu = h^\mu + g^\mu e \implies sp\left\{\widetilde{L}_\mu h^\mu\right\} = sp\{h^\mu\} \le c \implies \mu \in \widetilde{D}(c, h^\mu)$$

Therefore, by Lem. 5 we have that $\widetilde{T}_c h^\mu \ge \widetilde{L}_\mu h^\mu = h^\mu + g^\mu e$ and using the monotonicity of $\widetilde{T}_c$ (property *(a)* of Lem. 16) we obtain by induction that:

$$\forall n \in \mathbb{N},\ \ (\widetilde{T}_c)^n h^\mu \ge h^\mu + n g^\mu e \tag{44}$$

Combining (43) and (44) we have that

$$\forall n \ge 1,\ \ \frac{1}{n}\sum_{k=1}^{n}\left[(\widetilde{T}_c^{\eta,\downarrow})^k h^\mu - (\widetilde{T}_c^{\eta,\downarrow})^{k-1} h^\mu\right] = \frac{1}{n}\left[(\widetilde{T}_c^{\eta,\downarrow})^n h^\mu - h^\mu\right] \ge (g^\mu - \eta c)e \tag{45}$$

The term on the left-hand side of (44) is the *Cesaro mean* of the sequence $\left((\widetilde{T}_c^{\eta,\downarrow})^n h^\mu - (\widetilde{T}_c^{\eta,\downarrow})^{n-1} h^\mu\right)_{n\in\mathbb{N}}$. By property *2.* of Lem. 8 we know that this sequence converges to g^+ and thus by Cesaro theorem we know that the Cesaro mean has the same limit. Therefore, taking the limit on both sides of the inequality in (44) yields:

$$g^+ \ge g^\mu - \eta c$$

which concludes the proof. □

F. Regret Analysis of SCAL (Proof of Thm. 12)

We follow the proof structure in (Jaksch et al., 2010) and use similar notations. The main differences with Jaksch et al. (2010)'s regret proof are the following:

1. We use empirical Bernstein confidence bounds for both the rewards and the transition probabilities and not Hoeffding bounds.
2. The actual confidence bounds used by extended value iteration needs to be adapted in order to insure both convergence of the algorithm and feasibility of the policy (the MDP is "modfied", see Def 3).
3. The policy returned by extended value iteration may be stochastic.

F.1. Splitting into episodes

The regret after T time steps is defined as:

$$\Delta(\text{SCAL}, T) = Tg^* - \sum_{t=1}^{T} r_t(s_t, a_t)$$

Define the filtration $\mathcal{F}_t = \sigma(s_1, a_1, r_1, \ldots, s_{t+1})$ and the stochastic process $X_t = r_t(s_t, a_t) - \sum_{a \in \mathcal{A}_{s_t}} r(s_t, a)\widetilde{\pi}_{k_t}(s_t, a)$ where k_t is the episode at time t and $\widetilde{\pi}_{k_t}$ is the stochastic policy being executed at time t. Note that $\widetilde{\pi}_{k_t}$ is a random variable that is $\mathcal{F}_{t-1}$-measurable. Moreover, $(X_t, \mathcal{F}_t)_{t \geq 0}$ is a Martingale Difference Sequence (MDS) since $|X_t| \leq r_{\max}$ and $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = 0$. Using Azuma's inequality (see for example Jaksch et al. (2010, Lemma 10)):

$$\mathbb{P}\left(\sum_{t=1}^{T} r_t(s_t, a_t) \leq \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_{s_t}} r(s_t, a)\widetilde{\pi}_{k_t}(s_t, a) - r_{\max}\sqrt{\frac{5}{2} T \ln\left(\frac{11T}{\delta}\right)} \right) \leq \left(\frac{\delta}{11T}\right)^{5/4} < \frac{\delta}{20T^{5/4}}$$

For any episode k, we denote by t_k the starting time of that episode. Let's also denote by $\nu_k(s)$ (resp. $\nu_k(s,a)$) the total number of visits in state s (resp. state-action pair (s,a)) during episode k (i.e., before time t_{k+1}, t_{k+1} *not* included, and after time t_k, t_k included):

$$\begin{aligned} \nu_k(s,a) &:= \big| \{ t_k \leq \tau < t_{k+1} : (s_\tau, a_\tau) = (s,a) \} \big| \\ \nu_k(s) &:= \big| \{ t_k \leq \tau < t_{k+1} : s_\tau = s \} \big| = \sum_{a \in \mathcal{A}_s} \nu_k(s,a) \end{aligned}$$

Defining $\Delta_k = \sum_{s \in \mathcal{S}} \nu_k(s) \left(g^* - \sum_{a \in \mathcal{A}_{s_t}} r(s,a)\widetilde{\pi}_k(s,a) \right)$, it holds with probability at least $1 - \frac{\delta}{20T^{5/4}}$ that:

$$\begin{aligned} \Delta(\text{SCAL}, T) &\leq Tg^* - \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_{s_t}} r(s_t, a)\widetilde{\pi}_{k_t}(s_t, a) + r_{\max}\sqrt{\frac{5}{2} T \ln\left(\frac{11T}{\delta}\right)} \\ &= \sum_{k=1}^{m} \sum_{s \in \mathcal{S}} \nu_k(s) g^* - \sum_{k=1}^{m} \sum_{s \in \mathcal{S}} \nu_k(s) \sum_{a \in \mathcal{A}_s} r(s,a)\widetilde{\pi}_k(s,a) + r_{\max}\sqrt{\frac{5}{2} T \ln\left(\frac{11T}{\delta}\right)} \\ &= \sum_{k=1}^{m} \Delta_k + r_{\max}\sqrt{\frac{5}{2} T \ln\left(\frac{11T}{\delta}\right)} \end{aligned} \tag{46}$$

F.2. Dealing with failing confidence regions

We start by bounding the term $\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \widetilde{\mathcal{M}}_k}$ corresponding to the regret suffered in episodes where the true MDP M is not contained in the original set of plausible MDPs $\mathcal{M}_k$ (and not the modified set $\mathcal{M}_k^\ddagger$). We use exactly the same proof as in (Jaksch et al., 2010).

$$\begin{aligned} \sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} &\leq r_{\max} \sum_{k=1}^{m} \sum_{s} \nu_k(s) \mathbb{1}_{M \notin \mathcal{M}_k} = r_{\max} \sum_{k=1}^{m} \sum_{s,a} \nu_k(s,a) \mathbb{1}_{M \notin \mathcal{M}_k} \\ &\leq r_{\max}\sqrt{T} + r_{\max} \sum_{t=\lfloor T^{1/4} \rfloor + 1}^{T} t \mathbb{1}_{\{\exists k \geq 1 : t = t_k \text{ and } M \notin \mathcal{M}_k\}} \end{aligned}$$

Provided $\mathbb{P}(M \notin \mathcal{M}_k) \leq \frac{\delta}{15 t_k^6}$ for all $k \geq 1$ (see Thm. 22 below), we conclude as in Jaksch et al. (2010) that with probability at least $1 - \frac{\delta}{20T^{5/4}}$:

$$\sum_{k=1}^{m} \Delta_k \mathbb{1}_{M \notin \mathcal{M}_k} \leq r_{\max}\sqrt{T} \tag{47}$$

We recall the upper confidence bounds used for the reward function and the transition kernel in the algorithm:

$$|\widetilde{r}(s,a) - \widehat{r}_k(s,a)| \leq \overbrace{\sqrt{\frac{14\alpha_r \widehat{\sigma}^2_{r,k}(s,a)\ln(2SAt_k/\delta)}{\max\{1,N_k(s,a)\}}} + \frac{49\alpha_r r_{\max}\ln(2SAt_k/\delta)}{3\max\{1,N_k(s,a)-1\}}}^{\beta^{sa}_{r,k}} \tag{48}$$

$$|\widetilde{p}(s'|s,a) - \widehat{p}_k(s'|s,a)| \leq \underbrace{\sqrt{\frac{14\alpha_p \widehat{\sigma}^2_{p,k}(s'|s,a)\ln(2SAt_k/\delta)}{\max\{1,N_k(s,a)\}}} + \frac{49\alpha_p \ln(2SAt_k/\delta)}{3\max\{1,N_k(s,a)-1\}}}_{\beta^{sas'}_{p,k}} \tag{49}$$

where for the theoretical analysis we set[9] $\alpha_r = \alpha_p = 1$ and the *unbiased* estimates of the variances are

$$\widehat{\sigma}^2_{r,k}(s,a) = \frac{\sum_{t=1}^{t_k-1}\Big(r_t(s_t,a_t) - \widehat{r}_k(s,a)\Big)^2 \mathbb{1}_{\{(s_t,a_t)=(s,a)\}}}{N_k(s,a)-1} \tag{50}$$

$$\text{and } \widehat{\sigma}^2_{p,k}(s'|s,a) = \widehat{p}_k(s'|s,a)\left(1-\widehat{p}_k(s'|s,a)\right) \tag{51}$$

Note that although the definition of the sample variance of the reward r involves a sum, it can be computed dynamically using the following well-known recurrence relation:

$$\widehat{\sigma}^2_{n+1} = \frac{(n-1)}{n}\widehat{\sigma}^2_n + \frac{1}{n+1}(r_{n+1} - \widehat{r}_n)^2$$

where n denotes the number of samples ($N_k(s,a)$ in our case), r_{n+1} is the $(n+1)$-th sample observed, and $\widehat{r}_n = 1/n\sum_{i=1}^n r_i$ and $\widehat{\sigma}^2_n = 1/(n-1)\sum_{i=1}^n (r_i - \widehat{r}_n)^2$ are respectively the empirical average and sample variance obtained with the first n samples. The previous formula is subject to numerical unstability because the second term becomes negligible compared to the first term as n grows. A better approach for computing the variance is to exploit the following iterative scheme known as Welford's method (Knuth, 1997, p. 232):

$$\begin{aligned}
\widehat{r}_n &= \widehat{r}_{n-1} + \frac{r_n - \widehat{r}_{n-1}}{n} \\
S_n &= S_{n-1} + (r_n - \widehat{r}_{n-1})(r_n - \widehat{r}_n) \\
\widehat{\sigma}^2_n &= \frac{S_n}{n-1}
\end{aligned}$$

for $n \geq 2$, with $\widehat{r}_1 = r_1$ and $S_1 = 0$. This approach is less prone to numerical instability and its accuracy is comparable to the one of two-pass methods.

Theorem 22. *For any $k \geq 1$, the probability that the true MDP M is not contained in the set of plausible MDPs $\mathcal{M}_k$ at time t_k (as given by the confidence intervals in (48) and (49)) is at most $\frac{\delta}{15t_k^6}$, that is $\mathbb{P}(M \notin \mathcal{M}_k) \leq \frac{\delta}{15t_k^6}$.*

Proof. First note the following equality

$$\mathbb{P}\left(M \notin \mathcal{M}_k\right) = \mathbb{P}\left(\bigcup_{s,a,s'} \left\{\widetilde{r}_k(s,a) \notin B^k_r(s,a)\right\} \cup \left\{\widetilde{p}_k(s'|s,a) \notin B^k_p(s,a,s')\right\}\right)$$

Using Theorem 4 in (Maurer and Pontil, 2009)[10] we have that given an episode $k \geq 1$, a state action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$, a number of visits $N_k(s,a)$ in (s,a) before time t_k and (similarly to (Jaksch et al., 2010)):

$$\epsilon_{r,k} = \widehat{\sigma}_{r,k}\sqrt{\frac{2\ln(120SAt_k^7/\delta)}{\max\{1,N_k(s,a)\}}} + \frac{7r_{\max}\ln(120SAt_k^7/\delta)}{3\max\{1,N_k(s,a)-1\}} \leq \beta^{sa}_{r,k}$$

then

$$\mathbb{P}\left(|r(s,a) - \widehat{r}_k(s,a)| \geq \beta^{sa}_{r,k}\right) \leq \mathbb{P}\left(|r(s,a) - \widehat{r}_k(s,a)| \geq \epsilon_{r,k}\right) \leq \frac{\delta}{60t_k^7 SA}$$

[9] α_r and α_p are coefficients used to shrink the confidence intervals in the implementation in order to speed up the learning in practice. However, to insure that $M \in \mathcal{M}_k$ holds with high probability they should both be set equal to 1.

[10] Also known as "Empirical Bernstein" concentration inequality.

Similarly, $\mathbb{P}\left(|p(s'|s,a) - \widehat{p}_k(s'|s,a)| \geq \beta_{p,k}^{sas'}\right) \leq \frac{\delta}{20t_k^7 SA}$.

Note that when $N_k(s,a) = 0$ (i.e., there hasn't been any observation), the bound holds trivially with probability 1 both for rewards and transition probabilities. Recall the definitions of $B_r^k(s,a)$ and $B_p^k(s,a,s')$

$$B_r^k(s,a) = [\widehat{r}_k(s,a) - \beta_{r,k}^{sa}, \widehat{r}_k(s,a) + \beta_{r,k}^{sa}] \cap [0, r_{\max}]$$
$$\text{and } B_p^k(s,a,s') = [\widehat{p}_k(s'|s,a) - \beta_{p,k}^{sas'}, \widehat{p}_k(s'|s,a) + \beta_{p,k}^{sas'}] \cap [0,1]$$

Therefore it is clear that

$$\mathbb{P}\Big(\widetilde{r}_k(s,a) \notin B_r^k(s,a)\Big) \leq \frac{\delta}{60t_k^7 SA}$$
$$\text{and } \mathbb{P}\Big(\widetilde{p}_k(s'|s,a) \notin B_p^k(s,a,s')\Big) \leq \frac{\delta}{20t_k^7 S^2 A}$$

By taking a union bound over all state-action pairs (s,a) and all possible values for $N_k(s,a) = 0, \dots, t_k - 1$ we obtain

$$\mathbb{P}(M \notin \mathcal{M}_k) \leq \sum_{s,a} \sum_{N_k(s,a)=1}^{t_k - 1} \left(\frac{\delta}{60t_k^7 SA} + \sum_{s'} \frac{\delta}{20t_k^7 S^2 A} \right) \leq \frac{\delta}{15t_k^6}$$

□

F.3. Episodes whith $M \in \mathcal{M}_k$

Now we assume that $M \in \mathcal{M}_k$ and we first bound Δ_k. Note that we do *not* assume that M belongs to the modified set of MDPs $\mathcal{M}_k^\ddagger$. Denote by $\widetilde{g}_k := 1/2(\max\{v_{n+1} - v_n\} + \min\{v_{n+1} - v_n\})$ where v_n is the value function returned by SCOPT$(0, \overline{s}, \gamma_k, \varepsilon_k)$ (see Sec. 6). SCOPT recursively applies operator $\widetilde{T}_c^\ddagger := \widetilde{T}_c^{\eta_k,\downarrow}$ (see App. E) with a perturbation $\eta_k = 1/c\sqrt{t_k}$ until the stopping condition is reached. Moreover, the stopping condition is such that when the algorithm stops, the accuracy of the gain $\widetilde{g}_k$ with respect to $g_c^*(\widetilde{\mathcal{M}}_k^\ddagger)$ is $\varepsilon_k = 1/\sqrt{t_k}$ (see convergence guarantee *2)* of Thm. 18). Therefore, due to the fact that $M \in \mathcal{M}_k$ and since by assumption there exists an optimal policy π^* such that $sp\left\{h^{\pi^*}(M)\right\} \leq c$ we can apply Thm. 11 which implies that

$$\widetilde{g}_k \overset{\text{Thm. 18}}{\geq} g_c^*(\widetilde{\mathcal{M}}_k^\ddagger) - \underbrace{\varepsilon_k}_{=r_{\max}/\sqrt{t_k}} \overset{\text{Thm. 11}}{\geq} g_c^*(\widetilde{\mathcal{M}}_k) - \underbrace{\frac{c \cdot r_{\max}}{c \cdot t_k}}_{=r_{\max}/t_k} - \frac{r_{\max}}{\sqrt{t_k}} \overset{M \in \mathcal{M}_k}{\geq} g^* - \frac{r_{\max}}{t_k} - \frac{r_{\max}}{\sqrt{t_k}} \geq g^* - \frac{2r_{\max}}{\sqrt{t_k}}$$

implying:

$$\Delta_k \leq \sum_{s \in \mathcal{S}} \nu_k(s) \left(\widetilde{g}_k - \sum_{a \in \mathcal{A}_s} r(s,a) \widetilde{\pi}_k(s,a) \right) + 2r_{\max} \sum_{s \in \mathcal{S}} \frac{\nu_k(s)}{\sqrt{t_k}}$$

F.4. Extended Value Iteration

A direct consequence of Thm. 18 is that when the convergence criterion holds at iteration n then:

$$\forall s \in \mathcal{S}, \ \ |v_{n+1}(s) - v_n(s) - \widetilde{g}_k| \leq \frac{r_{\max}}{\sqrt{t_k}}$$

As is shown in the proof of Thm. 21, operator $\widetilde{T}_c^{\eta_k,\downarrow}$ is feasible at v_n for every $n \in \mathbb{N}$ (Asm. 9 holds) and we can expand v_{n+1} as

$$\forall s \in \mathcal{S}, \ \ v_{n+1}(s) = \sum_{a \in \mathcal{A}_s} \widetilde{r}_k(s,a) \widetilde{\pi}_k(s,a) + \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \widetilde{p}_k(s'|s,a) \widetilde{\pi}_k(s,a) v_n(s')$$

implying:

$$\forall s \in \mathcal{S}, \quad \left| \left(\widetilde{g}_k - \sum_{a \in \mathcal{A}_s} \widetilde{r}_k(s,a) \widetilde{\pi}_k(s,a) \right) - \left(\sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \widetilde{p}_k(s'|s,a) \widetilde{\pi}_k(s,a) v_n(s') - v_n(s) \right) \right| \leq \frac{r_{\max}}{\sqrt{t_k}} \tag{52}$$

Setting $\boldsymbol{\nu}_k := (\nu_k(s))_{s\in\mathcal{S}}$ the row vector of visit counts for each state and $\widetilde{\boldsymbol{P}}_k := \left(\sum_{a\in\mathcal{A}_s} \widetilde{p}_k(s'|s,a)\widetilde{\pi}_k(s,a)\right)_{s,s'\in\mathcal{S}}$ the "optimistic" transition matrix of $\widetilde{\pi}_k$ we obtain (using (52)):

$$\begin{aligned}
\Delta_k &\le \sum_{s\in\mathcal{S}} \nu_k(s)\left(\widetilde{g}_k - \sum_{a\in\mathcal{A}_s} r(s,a)\widetilde{\pi}_k(s,a)\right) + 2r_{\max}\sum_{s\in\mathcal{S}}\frac{\nu_k(s)}{\sqrt{t_k}} \\
&= \sum_{s\in\mathcal{S}} \nu_k(s)\left(\widetilde{g}_k - \sum_{a\in\mathcal{A}_s} \widetilde{r}_k(s,a)\widetilde{\pi}_k(s,a)\right) + \sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\Big(\widetilde{r}_k(s,a) - r(s,a)\Big) + 2r_{\max}\sum_{s\in\mathcal{S}}\frac{\nu_k(s)}{\sqrt{t_k}} \\
&\le \boldsymbol{\nu}_k(\widetilde{\boldsymbol{P}}_k - I)\boldsymbol{v}_n + \sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\Big(\widetilde{r}_k(s,a) - r(s,a)\Big) + 3r_{\max}\sum_{s\in\mathcal{S}}\frac{\nu_k(s)}{\sqrt{t_k}}
\end{aligned}$$

Since the rows of $\widetilde{\boldsymbol{P}}_k$ sum to 1 (i.e., $\widetilde{\boldsymbol{P}}_k\boldsymbol{e} = \boldsymbol{e}$), we can replace $\boldsymbol{v}_n$ by $\boldsymbol{w}_k$ where we set

$$\boldsymbol{w}_k := \boldsymbol{v}_n - \frac{\max_s v_n(s) + \min_s v_n(s)}{2}\boldsymbol{e}$$

In conclusion,

$$\Delta_k \le \boldsymbol{\nu}_k(\widetilde{\boldsymbol{P}}_k - I)\boldsymbol{w}_k + \sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\Big(\widetilde{r}_k(s,a) - r(s,a)\Big) + 3r_{\max}\sum_{s\in\mathcal{S}}\frac{\nu_k(s)}{\sqrt{t_k}} \tag{53}$$

By definition of operator $\widetilde{T}_c^{\eta_k,\downarrow}$, we have that $sp\{w_k\} = sp\{v_n\} = sp\left\{\widetilde{T}_c^{\eta_k,\downarrow}v_{n-1}\right\} \le c$ and since w_k is obtained by "recentering" v_n around 0 we have that $\|w_k\|_\infty = sp\{w_k\}/2 \le c/2$.

F.5. Bounding the reward

To guarantee the *feasibility* of operator $\widetilde{T}_c$ we had to *augment* the MDP (see Lem. 20), i.e., allow the rewards $\widetilde{r}_k$ to be as small as 0 even when $\widehat{r}_k - \beta_{r,k} > 0$. Nevertheless, the upper-bound of the reward was not modified (only the lower-bound) and so $\widetilde{r}_k \le \min\{r_{\max}, \widehat{r}_k + \beta_{r,k}\} \le \widehat{r}_k + \min\{r_{\max}, \beta_{r,k}\}$. Therefore:

$$\sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\Big(\widetilde{r}_k(s,a) - r(s,a)\Big) \le \sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\min\left\{r_{\max}, \beta_{r,k}^{sa}\right\} + \sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\Big(\widehat{r}_k(s,a) - r(s,a)\Big)$$

Moreover, since we assumed that $M\in\mathcal{M}_k$ the bound $\widehat{r}_k \le \min\{r_{\max}, r + \beta_{r,k}\} \le r + \min\{r_{\max}, \beta_{r,k}\}$ holds and thus

$$\sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\Big(\widetilde{r}_k(s,a) - r(s,a)\Big) \le 2\sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\min\left\{r_{\max}, \beta_{r,k}^{sa}\right\}$$

Note that when summing over all episodes $k \ge 1$, we can rewrite

$$\sum_{k=1}^{m}\sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\min\left\{r_{\max}, \beta_{r,k}^{sa}\right\} = \sum_{t=1}^{T}\sum_{a\in\mathcal{A}_{s_t}}\widetilde{\pi}_{k_t}(s_t,a)\min\left\{r_{\max}, \beta_{r,k_t}^{s_t a}\right\}$$

Define the filtration $\mathcal{F}_t = \sigma(s_1, a_1, r_1, \dots, s_{t+1})$ and the stochastic process

$$X_t = \sum_{a\in\mathcal{A}_{s_t}}\widetilde{\pi}_{k_t}(s_t,a)\min\left\{r_{\max}, \beta_{r,k_t}^{s_t a}\right\} - \min\left\{r_{\max}, \beta_{r,k_t}^{s_t a_t}\right\}$$

Note that $\widetilde{\pi}_{k_t}$ is a random variable that is $\mathcal{F}_{t-1}$-measurable. Moreover, $(X_t, \mathcal{F}_t)_{t\ge0}$ is an MDS since $|X_t| \le r_{\max}$ and $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$. Using Azuma's inequality:

$$\mathbb{P}\left(\sum_{t=1}^{T}\min\left\{r_{\max}, \beta_{r,k_t}^{s_t a_t}\right\} \le \sum_{t=1}^{T}\sum_{a\in\mathcal{A}_{s_t}}\widetilde{\pi}_{k_t}(s_t,a)\min\left\{r_{\max}, \beta_{r,k_t}^{s_t a}\right\} - r_{\max}\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}\right) \le \left(\frac{\delta}{11T}\right)^{5/4} < \frac{\delta}{20T^{5/4}}$$

or in other words, with probability at least $1 - \frac{\delta}{20T^{5/4}}$:

$$\sum_{k=1}^{m}\sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\min\left\{r_{\max},\beta_{r,k}^{sa}\right\} \le \sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\underbrace{\min\left\{r_{\max},\beta_{r,k}^{sa}\right\}}_{\le\beta_{r,k}^{sa}} + r_{\max}\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}$$

In conclusion, with probability at least $1 - \frac{\delta}{20T^{5/4}}$:

$$\sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\Big(\widetilde{r}_k(s,a) - r(s,a)\Big) \le 2\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\beta_{r,k}^{sa} + 2r_{\max}\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)} \tag{54}$$

F.6. Bounding the transition matrix

We denote by $\boldsymbol{P}_k := \left(\sum_a p(s'|s,a)\widetilde{\pi}_k(s,a)\right)_{s,s'\in\mathcal{S}}$ the true transition matrix and $\widehat{\boldsymbol{P}}_k := \left(\sum_a \widehat{p}_k(s'|s,a)\widetilde{\pi}_k(s,a)\right)_{s,s'\in\mathcal{S}}$ the estimated transition matrix. We do the following decomposition

$$\boldsymbol{\nu}_k(\widetilde{\boldsymbol{P}}_k - I)\boldsymbol{w}_k = \underbrace{\boldsymbol{\nu}_k(\widetilde{\boldsymbol{P}}_k - \widehat{\boldsymbol{P}}_k)\boldsymbol{w}_k + \boldsymbol{\nu}_k(\widehat{\boldsymbol{P}}_k - \boldsymbol{P}_k)\boldsymbol{w}_k}_{\boldsymbol{\nu}_k(\widetilde{\boldsymbol{P}}_k - \boldsymbol{P}_k)\boldsymbol{w}_k} + \boldsymbol{\nu}_k(\boldsymbol{P}_k - I)\boldsymbol{w}_k$$

Since we assumed that $M \in \mathcal{M}_k$ the difference $\widehat{\boldsymbol{P}}_k - \boldsymbol{P}_k$ concentrates. Moreover, the *perturbation* $\eta_k > 0$ applied by operator $\widetilde{T}_c^{\eta_k,\downarrow}$ to guarantee *convergence* (see Lem. 19) is only *shrinking* (and not expanding) the confidence intervals $B_p^k(s,a,s')$ and therefore by construction $\widetilde{p}_k(s'|s,a) \in B_p^k(s,a,s')$ implying that the difference $\widetilde{\boldsymbol{P}}_k - \widehat{\boldsymbol{P}}_k$ also concentrates. More formally, we have the following bounds

$$\begin{aligned}
\boldsymbol{\nu}_k(\widetilde{\boldsymbol{P}}_k - \widehat{\boldsymbol{P}}_k)\boldsymbol{w}_k &\le \sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\cdot\|\widetilde{p}_k(\cdot|s,a) - \widehat{p}_k(\cdot|s,a)\|_1\cdot\|w_k\|_\infty \\
&\le \sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\cdot\min\left\{2,\beta_{p,k}^{sa}\right\}\cdot\frac{c}{2}
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\nu}_k(\widehat{\boldsymbol{P}}_k - \boldsymbol{P}_k)\boldsymbol{w}_k &\le \sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\cdot\|\widehat{p}_k(\cdot|s,a) - p(\cdot|s,a)\|_1\cdot\|w_k\|_\infty \\
&\le \sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\cdot\min\left\{2,\beta_{p,k}^{sa}\right\}\cdot\frac{c}{2}
\end{aligned}$$

where $\beta_{p,k}^{sa} = \sum_{s'\in\mathcal{S}}\beta_{p,k}^{sas'}$. The term $\min\left\{2,\beta_{p,k}^{sa}\right\}$ appears because $\widetilde{p}_k(\cdot|s,a)$, $\widehat{p}_k(\cdot|s,a)$ and $p(\cdot|s,a)$ are probability distributions and any two probability distributions cannot be more than 2-far in ℓ_1 norm.

Similarly to what we did for the reward, when summing over all episodes $k \ge 1$, we can rewrite

$$\sum_{k=1}^{m}\sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\min\left\{2,\beta_{p,k}^{sa}\right\} = \sum_{t=1}^{T}\sum_{a\in\mathcal{A}_{s_t}}\widetilde{\pi}_{k_t}(s_t,a)\min\left\{2,\beta_{p,k_t}^{s_t a}\right\}$$

Define the filtration $\mathcal{F}_t = \sigma(s_1,a_1,r_1,\ldots,s_{t+1})$ and the stochastic process

$$X_t = \sum_{a\in\mathcal{A}_{s_t}}\widetilde{\pi}_{k_t}(s_t,a)\min\left\{2,\beta_{p,k_t}^{s_t a}\right\} - \min\left\{2,\beta_{p,k_t}^{s_t a_t}\right\}$$

Note that $\widetilde{\pi}_{k_t}$ is a random variable that is $\mathcal{F}_{t-1}$-measurable. Moreover, $(X_t,\mathcal{F}_t)_{t\ge0}$ is an MDS since $|X_t| \le 2$ and $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$. Using Azuma's inequality:

$$\mathbb{P}\left(\sum_{t=1}^{T}\min\left\{2,\beta_{p,k_t}^{s_t a_t}\right\} \le \sum_{t=1}^{T}\sum_{a\in\mathcal{A}_{s_t}}\widetilde{\pi}_{k_t}(s_t,a)\min\left\{2,\beta_{p,k_t}^{s_t a}\right\} - 2\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}\right) \le \left(\frac{\delta}{11T}\right)^{5/4} < \frac{\delta}{20T^{5/4}}$$

or in other words, with probability at least $1-\frac{\delta}{20T^{5/4}}$:

$$\sum_{k=1}^{m}\sum_{s,a}\nu_k(s)\widetilde{\pi}_k(s,a)\min\left\{2,\beta_{p,k}^{sa}\right\}\leq\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\underbrace{\min\left\{2,\beta_{p,k}^{sa}\right\}}_{\leq\beta_{p,k}^{sa}}+2\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}$$

In conclusion, with probability at least $1-\frac{\delta}{20T^{5/4}}$:

$$\sum_{k=1}^{m}\boldsymbol{\nu}_k(\widetilde{\boldsymbol{P}}_k-\boldsymbol{P}_k)\boldsymbol{w}_k\leq c\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\beta_{p,k}^{sa}+2c\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}\tag{55}$$

We now show that the remaining term $\boldsymbol{\nu}_k(\boldsymbol{P}_k-I)\boldsymbol{w}_k$ is an MDS. Let's denote by $\boldsymbol{e}_i$ the unit row vector with i-th coordinate 1 and all other coordinates 0.

$$\begin{aligned}\boldsymbol{\nu}_k(\boldsymbol{P}_k-I)\boldsymbol{w}_k&=\sum_{t=t_k}^{t_{k+1}-1}\left(\sum_{a\in\mathcal{A}_{s_t}}p(\cdot|s_t,a)\widetilde{\pi}_k(s_t,a)-\boldsymbol{e}_{s_t}\right)\boldsymbol{w}_k\\&=\sum_{t=t_k}^{t_{k+1}-1}\underbrace{\left(\sum_{a\in\mathcal{A}_{s_t}}p(\cdot|s_t,a)\widetilde{\pi}_k(s_t,a)-\boldsymbol{e}_{s_{t+1}}\right)\boldsymbol{w}_k}_{:=X_t}+\sum_{t=t_k}^{t_{k+1}-1}\left(\boldsymbol{e}_{s_{t+1}}-\boldsymbol{e}_{s_t}\right)\boldsymbol{w}_k\\&=\sum_{t=t_k}^{t_{k+1}-1}X_t+\underbrace{w_k(s_{t_{k+1}})-w_k(s_{t_k})}_{\leq sp\{w_k\}\leq c}\end{aligned}$$

Since $\|\boldsymbol{w}_k\|_\infty\leq\frac{c}{2}$ we have $|X_t|\leq\left(\|\sum_{a\in\mathcal{A}_{s_t}}p(\cdot|s_t,a)\widetilde{\pi}_k(s_t,a)\|_1+\|\boldsymbol{e}_{s_{t+1}}\|_1\right)\cdot\frac{c}{2}\leq c$. If we define the filtration $\mathcal{F}_t=\sigma\left(s_1,a_1,r_1,\ldots,s_{t+1}\right)$ then $\mathbb{E}[X_t|\mathcal{F}_{t-1}]=0$ since $\widetilde{\pi}_{k_t}$ is $\mathcal{F}_{t-1}$-measurable. Using Azuma's inequality:

$$\mathbb{P}\left(\sum_{t=1}^{T}\left(\sum_{a\in\mathcal{A}_{s_t}}p(\cdot|s_t,a)\widetilde{\pi}_k(s_t,a)-\boldsymbol{e}_{s_{t+1}}\right)\boldsymbol{w}_k\geq c\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}\right)\leq\left(\frac{\delta}{11T}\right)^{5/4}<\frac{\delta}{20T^{5/4}}$$

In conclusion, with probability at least $1-\frac{\delta}{20T^{5/4}}$:

$$\sum_{k=1}^{m}\boldsymbol{\nu}_k(\boldsymbol{P}_k-I)\boldsymbol{w}_k\mathbb{1}_{M\in\mathcal{M}_k}\leq c\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}+mc\tag{56}$$

In Appendix C.2 of (Jaksch et al., 2010) it is proved that given the stopping condition used for episodes, when $T\geq SA$ we can bound m as $m\leq SA\log_2\left(\frac{8T}{SA}\right)$.

F.7. Summing over episodes with $M\in\mathcal{M}_k$

We now gather inequalities (54), (55) and (56) into inequality (53) summed over all episodes k for which $M\in\mathcal{M}_k$ which yields (after taking a union bound) that with probability at least $1-\frac{3\delta}{20T^{5/4}}$ (for $T\geq SA$)

$$\begin{aligned}\sum_{k=1}^{m}\Delta_k\mathbb{1}_{M\in\mathcal{M}_k}\leq\;&c\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\beta_{p,k}^{sa}+c\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}+cSA\log_2\left(\frac{8T}{SA}\right)\\&+2\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\beta_{r,k}^{sa}+2r_{\max}\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}+3r_{\max}\sum_{k=1}^{m}\sum_{s\in\mathcal{S}}\frac{\nu_k(s)}{\sqrt{t_k}}\end{aligned}\tag{57}$$

The first and fourth terms appearing in the bound of Eq. 57 can be expanded as follows

$$\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\beta_{r,k}^{sa} = \sum_{k=1}^{m}\sum_{s,a}\Bigg[\underbrace{\sqrt{14\widehat{\sigma}_{r,k}^2\ln\left(\frac{2SAt_k}{\delta}\right)}}_{\leq r_{\max}\sqrt{14\ln\left(\frac{2SAT}{\delta}\right)}}\frac{\nu_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}} + \frac{49}{3}r_{\max}\underbrace{\ln\left(\frac{2SAt_k}{\delta}\right)}_{\leq\ln\left(\frac{2SAT}{\delta}\right)}\frac{\nu_k(s,a)}{\max\{1,N_k(s,a)-1\}}\Bigg]$$
$$\leq r_{\max}\sqrt{14\ln\left(\frac{2SAT}{\delta}\right)}\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}} + \frac{49}{3}r_{\max}\ln\left(\frac{2SAT}{\delta}\right)\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\max\{1,N_k(s,a)-1\}}$$

and similarly using the fact that $\beta_{p,k}^{sa} = \sum_{s'\in\mathcal{S}}\beta_{p,k}^{sas'}$

$$\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\beta_{p,k}^{sa} \leq \sqrt{14\ln\left(\frac{2SAT}{\delta}\right)}\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}}\sum_{s'\in\mathcal{S}}\sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))} + \frac{49}{3}S\ln\left(\frac{2SAT}{\delta}\right)\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\max\{1,N_k(s,a)-1\}}$$

By Cauchy-Schwartz inequality[11]

$$\sum_{s'\in\mathcal{S}}\sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))} = \sum_{s'\in\mathcal{S}:\,\widehat{p}_k(s'|s,a)>0}\sqrt{\widehat{p}_k(s'|s,a)(1-\widehat{p}_k(s'|s,a))}$$
$$\leq \sqrt{\left(\sum_{s'\in\mathcal{S}:\,\widehat{p}_k(s'|s,a)>0}\widehat{p}_k(s'|s,a)\right)\cdot\left(\sum_{s'\in\mathcal{S}:\,\widehat{p}_k(s'|s,a)>0}1-\widehat{p}_k(s'|s,a)\right)}$$
$$\leq \sqrt{\left|\operatorname{supp}\{\widehat{p}_k(\cdot|s,a)\}\right|-1}$$

where $\operatorname{supp}\{\widehat{p}_k(\cdot|s,a)\} = \{s'\in\mathcal{S}:\,\widehat{p}_k(s'|s,a)>0\}$ is the support of $\widehat{p}_k(\cdot|s,a)$ and $|\cdot|$ denotes the cardinal of a set. Note that by definition of $\widehat{p}_k$, $\operatorname{supp}\{\widehat{p}_k(\cdot|s,a)\}\subseteq\operatorname{supp}\{p(\cdot|s,a)\}$ and so $\left|\operatorname{supp}\{\widehat{p}_k(\cdot|s,a)\}\right|\leq\left|\operatorname{supp}\{p(\cdot|s,a)\}\right|$. Let's denote by Γ the maximal support over all state-action pairs (s,a):

$$\Gamma = \max_{s,a\in\mathcal{S}\times\mathcal{A}}\left|\operatorname{supp}\{p(\cdot|s,a)\}\right|$$

Therefore

$$\sum_{k=1}^{m}\sum_{s,a}\nu_k(s,a)\beta_{p,k}^{sa} \leq \sqrt{14(\Gamma-1)\ln\left(\frac{2SAT}{\delta}\right)}\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}} + \frac{49}{3}S\ln\left(\frac{2SAT}{\delta}\right)\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\max\{1,N_k(s,a)-1\}}$$

As proved by Jaksch et al. (2010, Sections 4.3.1 and 4.3.3), the stopping condition used for episodes implies that

$$\sum_{k=1}^{m}\sum_{s\in\mathcal{S}}\frac{\nu_k(s)}{\sqrt{t_k}} = \sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{t_k}} \leq \sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\sqrt{\max\{1,N_k(s,a)\}}} \leq \left(\sqrt{2}+1\right)\sqrt{SAT}$$

[11] The inequality obtained is somehow tight since when $\widehat{p}_k(\cdot|s,a)$ is uniform on its support, it becomes an equality.

Finally we bound the term

$$\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\max\{1,N_k(s,a)-1\}}=\sum_{s,a}\sum_{t=1}^{T}\frac{\mathbb{1}_{\{(s_t,a_t)=(s,a)\}}}{\max\{1,N_{k_t}(s,a)-1\}}$$

The stopping condition of episodes ensures that for all $t\geq 1$, $N_t(s,a)\leq 2N_{k_t}(s,a)$ where $N_t(s,a)$ is the total number of visits in (s,a) before time t, t *not* included:

$$N_t(s,a)=\#\{1\leq\tau<t:(s_\tau,a_\tau)=(s,a)\}$$

Therefore, similarly to what is done in (Ouyang et al., 2017, Proof of Lemma 5)

$$\begin{aligned}\sum_{k=1}^{m}\sum_{s,a}\frac{\nu_k(s,a)}{\max\{1,N_k(s,a)-1\}}&\leq 2\sum_{s,a}\sum_{t=1}^{T}\frac{\mathbb{1}_{\{(s_t,a_t)=(s,a)\}}}{\max\{1,N_t(s,a)-1\}}\\&=2\sum_{s,a}\left[\mathbb{1}_{\{N_{T+1}(s,a)\geq 1\}}+\mathbb{1}_{\{N_{T+1}(s,a)\geq 2\}}+\underbrace{\sum_{j=2}^{N_{T+1}(s,a)-1}\frac{1}{j-1}}_{\leq 1+\ln(N_{T+1}(s,a))\mathbb{1}_{\{N_{T+1}(s,a)\geq 1\}}}\right] &(58)\\&\leq 6SA+2\sum_{s,a}\ln\left(N_{T+1}(s,a)\right)\mathbb{1}_{\{N_{T+1}(s,a)\geq 1\}} &(59)\\&\leq 6SA+2SA\ln\left(\frac{\sum_{s,a}N_{T+1}(s,a)\mathbb{1}_{\{N_{T+1}(s,a)\geq 1\}}}{\sum_{s,a}\mathbb{1}_{\{N_{T+1}(s,a)\geq 1\}}}\right)\leq 6SA+2SA\ln(T) &(60)\end{aligned}$$

where (59) follows from the rate of divergence of an harmonic series and (60) is derived by applying Jensen's inequality to the concave function $\ln(\cdot)$ in the penultimate inequality (with a normalization factor $\sum_{sa}\mathbb{1}_{\{N_{T+1}(s,a)\geq 1\}}\leq SA$).

In conclusion, for $T\geq SA$, with probability at least $1-\frac{3\delta}{20T^{5/4}}$

$$\begin{aligned}\sum_{k=1}^{m}\Delta_k\mathbb{1}_{M\in\mathcal{M}_k}\leq&\left(\sqrt{28}+\sqrt{14}\right)c\sqrt{(\Gamma-1)SAT\ln\left(\frac{2SAT}{\delta}\right)}+\frac{98}{3}cS^2A\ln\left(\frac{2SAT}{\delta}\right)(3+\ln(T))\\&+2\left(\sqrt{28}+\sqrt{14}\right)r_{\max}\sqrt{SAT\ln\left(\frac{2SAT}{\delta}\right)}+\frac{196}{3}r_{\max}SA\ln\left(\frac{2SAT}{\delta}\right)(3+\ln(T)) &(61)\\&+(3c+2r_{\max})\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}+cSA\log_2\left(\frac{T}{SA}\right)+3(\sqrt{2}+1)r_{\max}\sqrt{SAT}\end{aligned}$$

F.8. Completing the regret bound

From (46) we have that with probability at least $1-\frac{\delta}{20T^{5/4}}$

$$\Delta(\text{SCAL},T)\leq\sum_{k=1}^{m}\Delta_k\mathbb{1}_{M\in\mathcal{M}_k}+\sum_{k=1}^{m}\Delta_k\mathbb{1}_{M\notin\mathcal{M}_k}+r_{\max}\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)} \tag{62}$$

By gathering (47) and (61) into (62) (using a union bound) we have that with probability at least $1-\frac{\delta}{20T^{5/4}}-\frac{3\delta}{20T^{5/4}}-\frac{\delta}{20T^{4/5}}=1-\frac{\delta}{4T^{5/4}}$ (for $T\geq SA$)

$$\begin{aligned}\Delta(\text{SCAL},T)\leq&\left(\sqrt{28}+\sqrt{14}\right)c\sqrt{(\Gamma-1)SAT\ln\left(\frac{2SAT}{\delta}\right)}+\frac{98}{3}cS^2A\ln\left(\frac{2SAT}{\delta}\right)(3+\ln(T))\\&+2\left(\sqrt{28}+\sqrt{14}\right)r_{\max}\sqrt{SAT\ln\left(\frac{2SAT}{\delta}\right)}+\frac{196}{3}r_{\max}SA\ln\left(\frac{2SAT}{\delta}\right)(3+\ln(T)) &(63)\\&+3(c+r_{\max})\sqrt{\frac{5}{2}T\ln\left(\frac{11T}{\delta}\right)}+cSA\log_2\left(\frac{8T}{SA}\right)+3(\sqrt{2}+1)r_{\max}\sqrt{SAT}+r_{\max}\sqrt{T}\end{aligned}$$

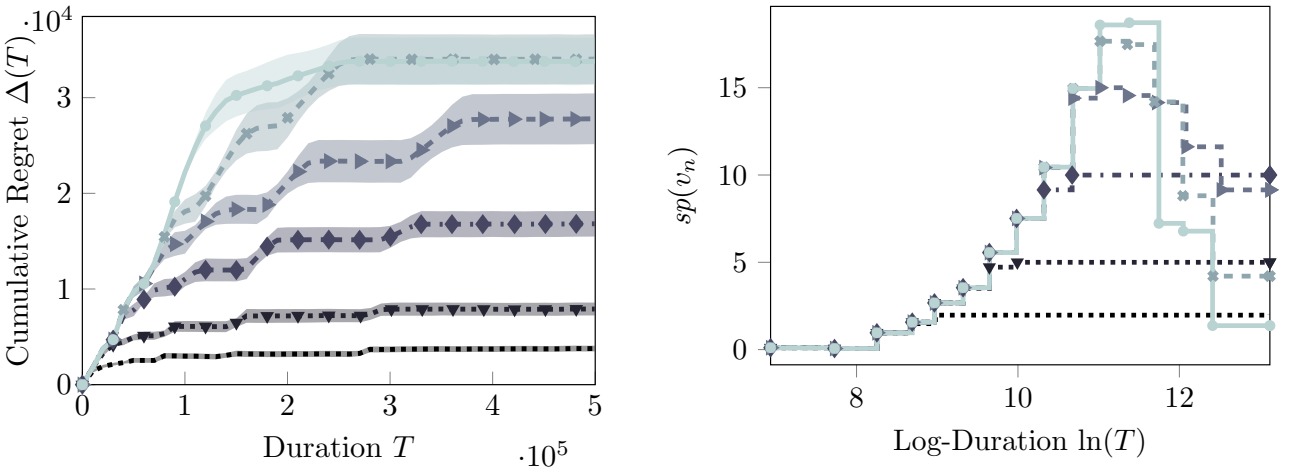

Figure 9. UCRL and SCAL behaviour with $\delta = 0.005$ in the three-states MDP.

For $T \leq SA$ the regret can be bounded with probability 1 as

$$\Delta(\text{SCAL}, T) \leq r_{\max} T = r_{\max} \sqrt{T} \cdot \sqrt{T} \leq r_{\max} \sqrt{SAT}$$

Finally, we take a union bound over all possible values of T and use the fact that $\sum_{T=2}^{+\infty} \frac{\delta}{4T^{5/4}} < \delta$.

In conclusion, there exists a numerical constant α such that for any MDP M, with probability at least $1 - \delta$ our algorithm SCAL has a regret bounded by

$$\Delta(\text{SCAL}, T) \leq \alpha \cdot \left(\max\{r_{\max}, c\} \sqrt{\Gamma SAT \ln\left(\frac{T}{\delta}\right)} + \max\{r_{\max}, c\} S^2 A \ln^2\left(\frac{T}{\delta}\right) \right) \tag{64}$$

The second term of the upper-bound in Eq. 64 is negligible when T is big enough and so

$$\mathbb{P}\left(\Delta(\text{SCAL}, T) = \mathcal{O}\left(\max\{r_{\max}, c\} \sqrt{\Gamma SAT \ln\left(\frac{T}{\delta}\right)} \right) \right) \geq 1 - \delta$$

G. Additional Experiments

In this section we provide clearer figures for the three-states domain and we present a more challenging domain: Knight Quest.

G.1. Three-States MDP

We simply restate the results presented in the main paper on bigger figures (see Fig. 9 and 10).

G.2. Knight Quest

The second environment takes inspiration from classical arcade games. The goal is to rescue a princess in the shortest time without being killed by the dragon. To achieve this task, the knight needs to collect gold, buy the magic key and reach the princess location. A representation of the environment is provided in Fig. 11.

The elements of the game are: I) the knight; II) the princess; III) a dragon patrolling the princess; IV) a gold mine and V) a town.

Town, Princess and Gold Mine. These elements are special states of the environment. The town (T) is the place where the knight can buy objects and where it is reset when he rescues the princess or he is killed by the dragon. The princess (P) is the terminal state, while the gold mine (G) is the place where the knight can collect gold.

Dragon. The dragon (D) is the enemy and it is randomly moving around the princess's location. Let's denote with

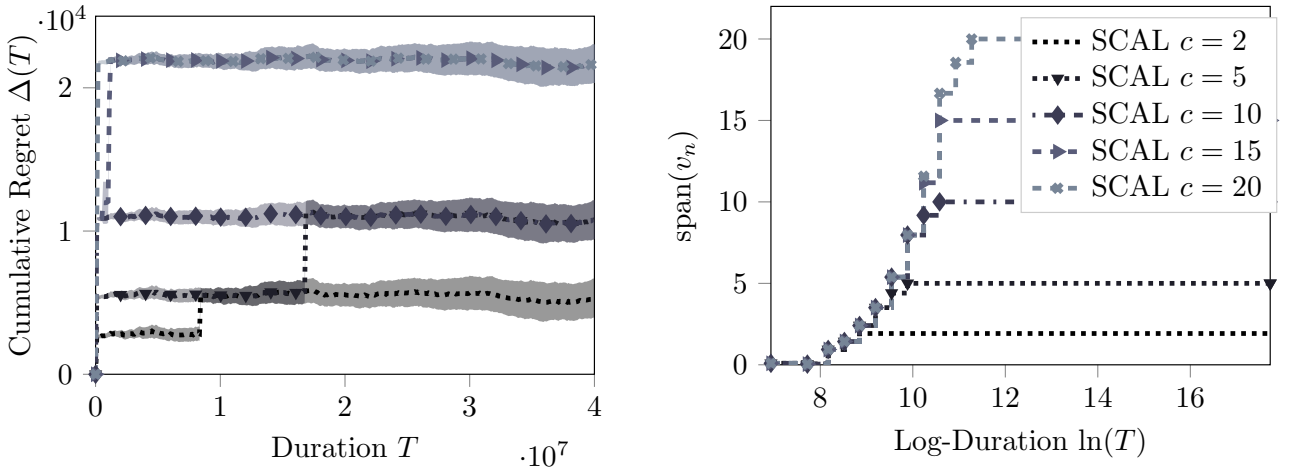

Figure 10. SCAL behaviour with $\delta = 0$ in the three-states MDP. In this setting, the MDP is weakly communicating and UCRL is not able to learn. We have omitted UCRL since it is out-of-scale (see Fig. 5).

$d \in \{0, 1, 2\}$ the position of the dragon such that: $d = 0 := (0, 1)$, $d = 1 := (1, 0)$ and $d = 2 := (1, 1)$. The transition probabilities of the dragon are:

$$p_d(\cdot|0) = [0.4, 0, 0.6]; \quad p_d(\cdot|1) = [0, 0.4, 0.6]; \quad p_d(\cdot|2) = [0.4, 0.2, 0.4].$$

When the dragon can kill the knight when they are at the same position and the knight does not have the armour.

Knight. The knight is the only player of the game. He moves in the environment using the four cardinal actions (i.e.,, *right*, *down*, *left* and *up*) plus an action to keep the current position (*stay*). We refer to these 5 actions as *movement actions*. Additionally, the knight can collect the gold (action *CG*), buy a key (action *BK*) or buy an armour (action *BA*).

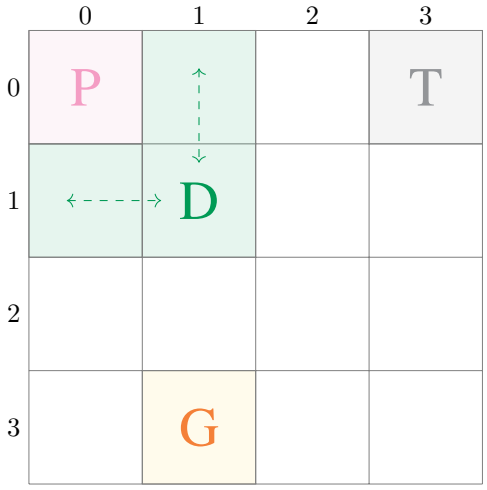

Figure 11. Representation of the Knight Quest 4×4 map. The green shadowed cells are the locations where the dragon can move.

State representation, action effect and reward. The state s_t of the game is represented by the following elements:

- Knight position: coordinates of the grid (row, col), $row, col \in 0, 1, 2, 3$;
- Gold level: the amount of gold own by the knight, $g \in \{0, 1\}$;
- Dragon position: $d \in \{0, 1, 2\}$;
- Object identifier: the object(s) own by the knight, $o = \{0, 1, 2, 3\}$ where $0 :=$ nothing, $1 :=$ key, $2 :=$ armour and $3 :=$ key and armour.

Now we can finally explain the effects of the actions, i.e., how states s_{t+1} is generated. The movement actions have the trivial effect of changing the knight position. The action CG changes the state only when the knight is at the mine. In this case the level of gold is incremented by one, formally $g_{t+1} = \min\{1, g_t + 1\}$. Actions BK and BA alter the state only when are executed in the town with gold-level equal to 1,

i.e.,

$$a_t = BK \wedge (row_t, col_t) = T \wedge g_t = 1 \implies o_{t+1} = \begin{cases} 1 & \text{if } o_t = 0 \\ 3 & \text{otherwise} \end{cases}$$

$$a_t = BA \wedge (row_t, col_t) = T \wedge g_t = 1 \implies o_{t+1} = \begin{cases} 2 & \text{if } o_t = 0 \\ 3 & \text{otherwise} \end{cases}$$

All the actions are deterministic when the knight does not own the armour. When the knight has the armour:

- The movement actions result in a normal (correct) transition with probability 0.5, otherwise the current position is kept;
- The CG action fails with probability 0.99, i.e., with probability 0.01 the gold level is increment by 1.
- Actions BK and BA are not modified.

Notice that when the knight is equipped with the armour it cannot be killed by the dragon (i.e., knight and dragon can occupy the same cell). However, due to the weight of the armour, knight's gait is unsteady. At the same time, the armour makes the collection of the goal very challenging (i.e., success probability is 0.01). You can imagine that mining with a metal armour can be very difficult!

The basic reward signal is -1 at each time step. Nevertheless, the knight receives a reward of -10 when he executes CG, BK or BA outside the designed location (i.e., mine and town). Finally, he obtains a reward of 20 when he reaches the princess with the key and -20 when he is killed by the dragon (i.e., knight and dragon are in the same cell and the knight does not have the armour). For the experiments, we have scaled the reward in the range $[0, 1]$.

Finally, when the episode ends (i.e., the knight reaches the princess with the key or he is killed), the knight is reset at town location with no gold or object ($g, o = 0$) and the dragon position is randomly drawn ($d \sim \mathcal{U}(\{0, 1, 2\})$).

Properties of the game. The state and action space size are $S = 360$ and $A = 8$, while the diameter of the MDP is $D \approx 250$. The diameter is due to the following path: start from the town with no gold but the armour and reach the princess with one unit of gold, key and armour. However, the optimal strategy is simply to collect the gold, buy the magic key from the town and rescue the princess.[12] The optimal policy is such that $g^* \approx 0.5$, $sp\{h^*\} \approx 3.28$.

This game is challenging for OFU approaches since the policy suffering the diameter is orthogonal to the optimal one. This is due to the presence of actions that are not relevant to the final objective and simply mess up the navigability of the environment. We think that this characteristic is shared by common real-world applications where the agent can face several choices (actions) and most of them are useless. This property can be interpreted also as a hierarchical structure that has been proved to be at the core of many applications.

More practically, the high diameter induces UCRL to explore remote states that are seen as promising states (in contrast with the real importance). On contrary, SCAL can leverage on the knowledge of "simple" game (i.e., small span) in order to condition the exploration. Notice that the span constraint c can be interpreted as the difficulty of the game. This game becomes difficult only if I want reach extreme states that are nevertheless useless to the final goal (rescue the princess). By giving small c values to the algorithm, we are implicitly saying that the game is simple, do not trust states that are too promising (i.e., generate a high span).

Results. We have tested UCRL and SCAL with different constraints over an horizon $T = 4 \cdot 10^8$ with Bernstein's bound. The code is available on GitHub (`https://github.com/RonanFR/UCRL`). SCAL is run with the reward augmented but no perturbation of the transition matrix ($\eta_k = 0$). For the terminal condition of SCOPT we set $\gamma_k = 0$.[13] In order to speed up the learning we have set $\alpha_p = \alpha_r = 0.05$ (see Eq. 49 and 48). This still guarantees that the confidence intervals at t_0 are still bigger than 1 and $r_{\max}$, respectively. Results are reported in Fig. 12. We can notice that SCAL is able to outperform UCRL by a big margin. It is interesting to notice that in the regret curves it is easy to identify the linear and logarithmic regimes, while the square-root one is almost absent. This is due to the fact that once the algorithms discover that visiting

[12] Notice that there are deterministic strategies to the princess's location avoiding the dragon.

[13] Note that in EVI and SCOPT the optimistic reward is truncated at $r_{\max}$, i.e., $\max_{\widetilde{r}}\{\widetilde{r}(s,a)\} = \min\{r_{\max}, \widehat{r}(s,a) + \max\{B^r(s,a)\}\}$.

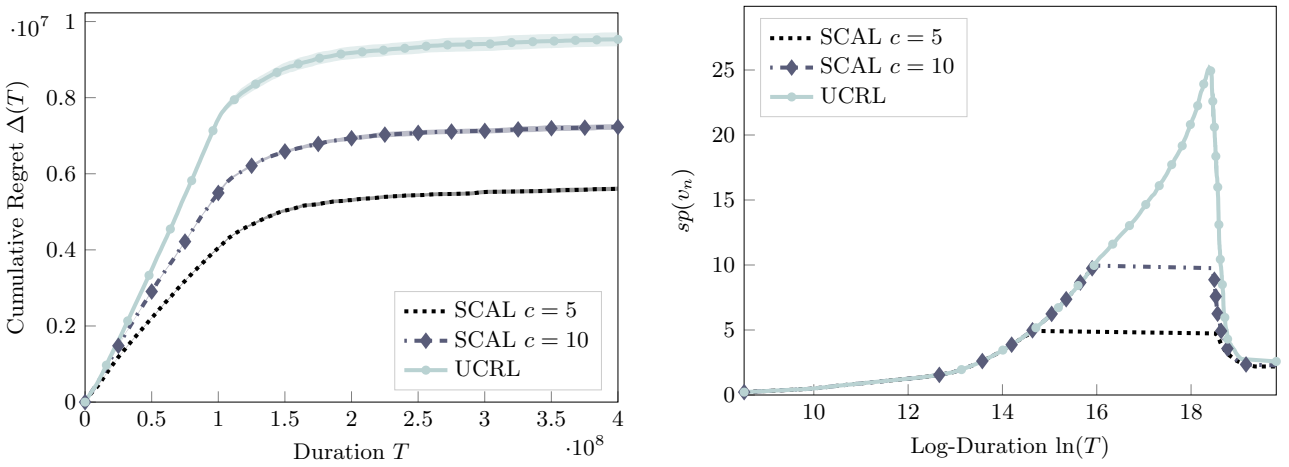

Figure 12. Algorithm behaviour in the knight quest game. Figures show the span of the optimistic bias *(right)* and the cumulative regret *(left)* as a function of T. Results are averaged over 15 runs and 95% confidence intervals of the mean are shown for the regret.

extreme states is not relevant they have almost perfectly learnt the dynamics under the optimal policy.[14]

[14]The actions executed by the optimal policy are deterministic. This means that by using Bernstein's bound we have the term involving the variance equal to zero and the second term scales linearly with the number of visits.