# 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. 

[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 $\varepsilon$-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 $\Gamma$ 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 $D^{\text{MR}}$, while the subset of Markov deterministic decision rules is $D^{\text{MD}}$. 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 $g^\pi$ and bias $h^\pi$ 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 $\varepsilon_k$-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* $\varepsilon_k$-*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 $g^\pi_M$ and $sp\{h^\pi_M\}$ are in general non-continuous functions of $(M, \pi)$, 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

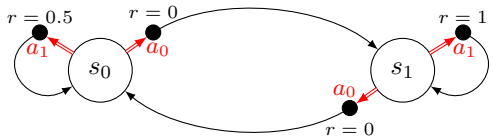


*Figure 2.* Toy example with deterministic transitions and reward for all actions.

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 $g_M^\pi$ 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 $\pi_D$ 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 $\pi_R$ 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 $\pi$ in $\Pi^{SR}$ 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 $\pi_R$ with constant gain has $g_c^* = [c, c]$. On the other hand, the only policy $\pi$ 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 $\Pi_c$ 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 $\Pi_c$ (i.e., $\Pi_c$ is not closed) and $g^\pi$ is not continuous at this point. Notice that when the MDP is *unichain* (Puterman, 1994, Sec. 8.3), $\Pi_c$ is compact, $g^\pi$ 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 $\delta_v^+$ 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 $\delta_v^+$ 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 $\gamma$-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 $\gamma$-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 $\delta_v^+$ 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* $\gamma$ *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* $\pi_n$ *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 $\varepsilon$-optimal policy $\pi_n$. 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 $\pi^+$ is unichain then $sp\{h^+\} = sp\{h^{\pi^+}\}$, hence the second part of the theorem. Notice also that even if $\pi^+$ is unichain, we cannot guarantee that $\pi_n$ 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* $\eta$ *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 $\gamma$-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 $\pi$ implying that $\pi$ is unichain and so $\widetilde{\mathcal{M}}^\ddagger$ is unichain. As is shown in Thm. 11, the $\eta$-perturbation of $B_p$ introduces a small bias $\eta c$ 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 $\gamma$-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 $\pi^*$ 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 $\delta$ 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-

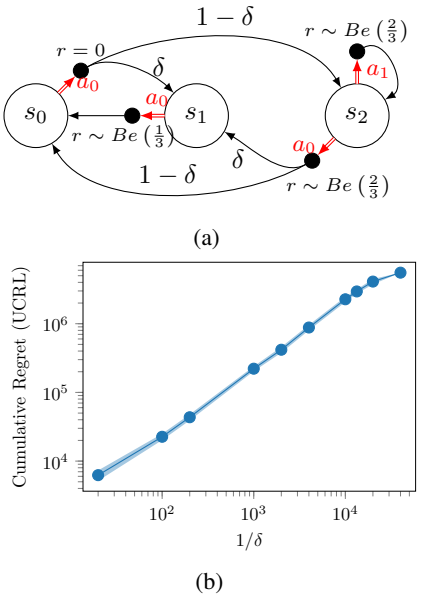


*Figure 4. (upper)* Simple three-state domain. *(lower)* Cumulative regret incurred by UCRL after $T = 2.5 \cdot 10^7$ steps as a function of the diameter $D \approx 1/\delta$ (averaged over 20 runs).

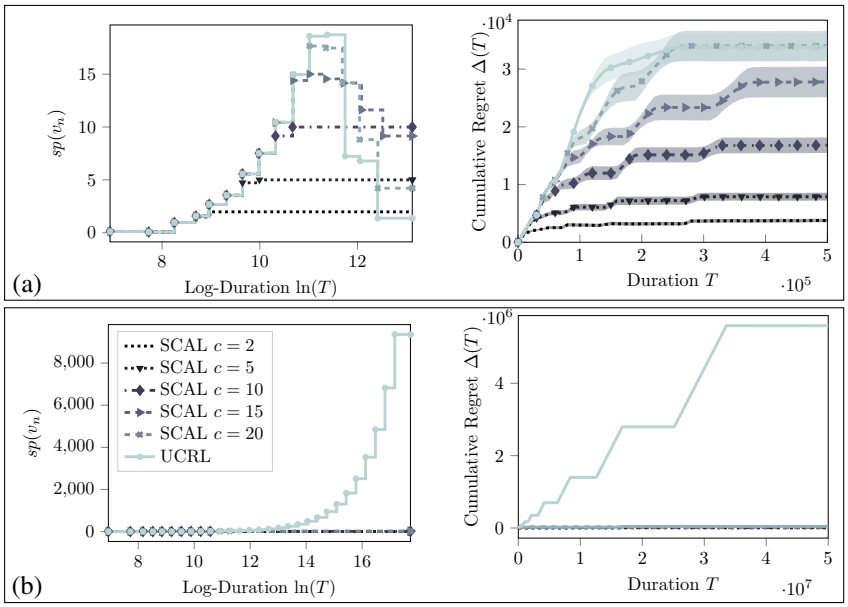


*Figure 5.* Results in the three-states domain with $\delta = 0.005$ *(top)* and $\delta = 0$ *(bottom)*. We report the span of the optimistic bias *(left)* and the cumulative regret *(right)* as a function of $T$. Results are averaged over 20 runs and 95% confidence intervals are shown.

ing a communicating MDP. With such a small $\delta$, 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 $\Gamma$. 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.