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

Ronan Fruit (INRIA), **Matteo Pirotta** (INRIA), **Alessandro Lazaric** (FAIR), **Ronald Ortner** (Montanuniversität Leoben)

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR · Inria informatics mathematics · Région Hauts-de-France · MONTAN UNIVERSITÄT LEOBEN · facebook Artificial Intelligence Research

Motivations

- Learning in an unknown environment means to balance
  - Exploration
  - Exploitation
- Optimistic (OFU) methods
  - Construct a set of plausible MDPs
  - Execute the optimal policy of the "best" MDP in the set
- OFU may lead to over-optimism is some MDPs
  - even fail in learning
- Regularization has proved to be effective in ML ⇒ in Exp-Exp?

Online Learning in MDPs

Markov Decision Process $M = \{S, A, r, p\}$

- S: states
- $A = (A_s)_{s\in S}$: actions
- $r(s,a)$: mean rewards
- $p(s'|s,a)$: transition probabilities, $\Gamma = \max_{s,a} \|p(\cdot|s,a)\|_0 \le S$

Optimality criterion: long-term average reward

For any policy $\pi \in \Pi^{\mathrm{SR}}(M)$ starting from $s \in \mathcal{S}$:

$$\text{GAIN:} \quad g_M^\pi(s) := \lim_{T\to+\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r(s_t,a_t)\right]$$

$$\text{BIAS:} \quad h_M^\pi(s) := \underset{T\to+\infty}{C\text{-}\lim}\, \mathbb{E}\left[\sum_{t=1}^{T}\left(r(s_t,a_t) - g_M^\pi(s_t)\right)\right]$$

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 $s \in \mathcal{S}$.
Moreover, $(g^*, h^*) = (g^{\pi^*}, h^{\pi^*})$ satisfy the *optimality equation*

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

optimal Bellman operator: $Lv(s) = \max_{a\in\mathcal{A}_s}\{r(s,a) + p(\cdot|s,a)^\intercal v\}$

Cumulative regret minimization $\Delta(\mathfrak{A}, T) = \sum_{t=1}^{T}\left(g^* - r_t(s_t,a_t)\right)$

Diameter and Span: [Jaksch et al. 2010; Bartlett and Tewari, 2009]

$$D = \max_{s,s'\in\mathcal{S}}\left\{\min_{\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})} \boxed{\mathbb{E}_\pi\left[T(s')|s\right]}\right\} \quad \text{(expected time } s\to s'\text{)}$$

$$sp\{h^*\} = \max_{s\in\mathcal{S}}\{h^*(s)\} - \min_{s\in\mathcal{S}}\{h^*(s)\}$$

- $sp\{h^*\} \le D$ *(always)*
- $D = \infty$ but $sp\{h^*\} < \infty$

Regret Results

UCRL [Jaksch et al. 2010] & Optimistic PSRL [Agrawal and Jia, 2017]	$\widetilde{O}(D\sqrt{\Gamma SAT})$
REGAL.C [Bartlett and Tewari, 2009]	$\widetilde{O}(C\sqrt{\Gamma SAT})$ with $sp\{h^*\} \le C$

Lower bound: $\Omega\left(\sqrt{DSAT}\right)$ or $\Omega\left(\sqrt{sp\{h^*\}SAT}\right)$? [Jaksch et al. 2010]

References

- Jaksch, Ortner, and Auer. Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 2010.
- Bartlett and Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In UAI 2009.
- Agrawal and Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In NIPS 2017, 2017.

Optimism in Face of Uncertainty

For episodes $k = 1, 2, \dots$

1. Compute a set of plausible MDPs $\mathcal{M}_k$
2. Planning: Compute an optimistic policy $\widetilde{\pi}_k$
3. Execute policy $\widetilde{\pi}_k$

Confidence set:
$\mathcal{M}_k = \left\{M = \langle \mathcal{S}, \mathcal{A}, \widetilde{r}, \widetilde{p}\rangle : \widetilde{r}(s,a) \in B_r^k(s,a),\ \widetilde{p}(s'|s,a) \in B_p^k(s,a,s')\right\}$
B_r^k, B_p^k = high-probability confidence intervals around empirical estimates

$$|\widehat{r}(s,a) - \widetilde{r}(s,a)| \le \beta_{r,k}^{sa}, \qquad |\widehat{p}(s'|s,a) - \widetilde{p}(s'|s,a)| \le \beta_{p,k}^{sas'}$$

Planning:

- ▶ UCRL [Jaksch et al., 2010] $\quad (\widetilde{M}_k^*, \widetilde{\pi}_k^*) \in \underset{M\in\mathcal{M}_k, \pi\in\Pi^{\mathrm{SD}}(M)}{\arg\max}\, g_M^\pi$
  - No prior knowledge ⇒ might be over-optimistic
- ▶ REGAL.C [Bartlett and Tewari, 2009] $\quad (\widetilde{M}_{\mathrm{R}}^*, \widetilde{\pi}_{\mathrm{R}}^*) = \underset{M\in\mathcal{M}_{\mathrm{RC}}, \pi\in\Pi^{\mathrm{SD}}(M)}{\arg\max}\, g_M^\pi$

  where $\mathcal{M}_{\mathrm{RC}} := \{M \in \mathcal{M}_k : sp\{h_M^*\} \le c\}$
  - Constrains the set of possible MDPs ⇒ intractable
  - Overall problem is ill-posed
- ▶ SCAL $\quad (\widetilde{M}_c^*, \widetilde{\pi}_c^*) \in \underset{M\in\mathcal{M}_k,\ \pi\in\Pi_c(M)}{\arg\max}\, g_M^\pi$

  where $\Pi_c(M) := \left\{\pi \in \Pi^{\mathrm{SR}} : sp\{h_M^\pi\} \le c \ \wedge\ sp\{g_M^\pi\} = 0\right\}$
  - Relaxation of REGAL.C: see PROP. 1
  - Overall problem is well-posed

Use prior knowledge about $sp\{h_{M^*}^*\} \le c$

Planning under bias-span constraint (SCOPT)

Optimization problem: $g_c^* := \sup_{\pi\in\Pi_c(M)}\{g_M^\pi\}$

where $\Pi_c(M) := \left\{\pi \in \Pi^{\mathrm{SR}} : sp\{h_M^\pi\} \le C \ \wedge\ sp\{g_M^\pi\} = 0\right\}$

Why $\pi \in \Pi^{\mathrm{SR}}$?
the maximizer $\pi_c^*(M)$ may be a randomize policy

Why $sp\{g_M^\pi\} = 0$?
there may be no *dominating* policy $\pi \in \Pi^{\mathrm{SR}}$ with constant bias span
$\pi \in \Pi^{\mathrm{SR}}$ is *dominating* $\Rightarrow \forall \pi' \in \Pi^{\mathrm{SR}}, \forall s \in \mathcal{S}, g^\pi(s) \ge g^{\pi'}(s)$

The supremum always exists, the set $\Pi_c^*(M)$ of maximizers?

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

Value Operator: Given $v \in \mathbb{R}^S$ and $C \ge 0$, we define $\overline{\mathcal{S}}(C,v) = \left\{s \in \mathcal{S} | Lv(s) \le \min_s\{Lv(s)\} + C\right\}$ and

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

T_c is *feasible* at (v,s) i.i.f.
$s \in \widetilde{\mathcal{S}}(C,v) = \{\min_{a\in\mathcal{A}_s}\{r(s,a) + p(\cdot|s,a)^\intercal v\} \le \min_{s'}\{Lv(s')\} + C$
$\Rightarrow \exists \delta_v^+\ T_c v(s) = \sum_{a\in\mathcal{A}_s} \delta_v^+(s,a)\left[r(s,a) + p(\cdot|s,a)^\intercal v\right]$

Greedy operator:

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

Planning algorithm: SCOPT →

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

relative value iteration
c-span truncation (from above)
$\forall n,\ sp\{v_n\} \le c$

Asm. 1. L is a γ-span contraction
$\forall u, v \in \mathbb{R}^S,\ sp\{Lu - Lv\} \le \gamma sp\{u - v\}$

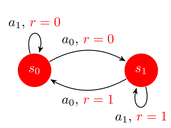
Asm. 2. T_c is *globally feasible*:
$\exists \delta_v^+,\ T_c v = L_{\delta_v^+} v$ (feasible at each s)

SCOPT PROPERTIES

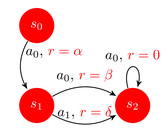
- *Span contraction:* T_c is a γ-span contraction
- *Optimality equation (OE):* $\exists (g^+, h^+) \in \mathbb{R}\times\mathbb{R}^S$ s.t. $T_c h^+ = h^+ + g^+ e$.
- *Uniqueness:* if (g,h) solution of OE, then $g = g^+$ and $h = h^+ + \lambda e$
- *Convergence:* $\forall v_0 \in \mathbb{R}^S$, the sequence (v_n) converges to h^+ and $\lim_{n\to+\infty} T_c^{n+1} v_0 - T_c^n v_0 = g^+ e$.
- *Dominance:* the gain g^+ is an upper-bound on the supremum, i.e., $g^+ \ge g_c^*$.

Asm. 1 + Asm. 2

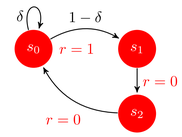
- *ϵ-optimality:* the policy returned by SCOPT is ϵ-optimal: $\|g^+ e - g^{\pi_n}\|_\infty \le \varepsilon$
- *Convergence to g_c^*:* if $\pi^+ = G_c(h^+)$ is unichain, then $g^+ = g_c^*$ and $\pi^+ \in \Pi_c^*$



T_c might not always be feasible even when $\Pi_c^* \neq \emptyset$



T_c might not always be feasible at its fixed-point h^+



$(T_c)^n v_0$ might not converge even when $\Pi_c \neq \emptyset$ and all policies are both unichain and aperiodic

Learning under bias-span constraint (SCAL)

Equivalent planning problem:

$$\widetilde{\mu}^* \in \underset{\mu\in\Pi_c(\widetilde{\mathcal{M}}_k)}{\arg\max}\, \{g_{\widetilde{\mathcal{M}}_k}^\mu\}$$

$\widetilde{\mathcal{M}}_k$ is an *extended MDP*
⇒ $\mathcal{A}$ is "extended" to a compact space $\widetilde{\mathcal{A}}$ by considering every possible value in B_r^k and B_p^k

Define

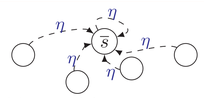
$$\widetilde{L}v = \max_{a\in\mathcal{A}_s}\left\{\max_{\widetilde{r}\in B_r^k(s,a)} \widetilde{r} + \max_{\widetilde{p}\in B_p^k(s,a,\cdot)} \{\widetilde{p}^\intercal v\}\right\}$$

and $\widetilde{T}_c = \Gamma_c \widetilde{L} v$
⇒ use SCOPT for planning

No convergence for SCOPT in $\widetilde{\mathcal{M}}_k$

CAN WE STILL USE SCOPT?

We can alter $\widetilde{\mathcal{M}}_k \Rightarrow \widetilde{\mathcal{M}}_k^\dagger$



1. η-perturbation of the transition model
   we enforce that the *"attractive"* state $\overline{s}$ is reached with non-zero probability from any state-action pair

$$\gamma = 1 - \min_{\substack{s,u\in\mathcal{S},\ a,b\in\mathcal{A} \\ \widetilde{p},\ \widetilde{q}\in B_p^\ddagger}}\left\{\sum_{j\in\mathcal{S}} \min\left\{\widetilde{p}(j|s,a), \widetilde{q}(j|u,b)\right\}\right\}$$

$$\le 1 - \eta < 1 \quad \Longrightarrow \quad \widetilde{L}^\dagger \text{ is } \gamma\text{-contractive}$$

SCOPT in $\widetilde{\mathcal{M}}_k^\dagger$ converges to $g_c^* \ge \max_{\pi\in\Pi_c(\widetilde{\mathcal{M}}_k)} g_{\widetilde{\mathcal{M}}_k}^\pi - \eta c$

There might not be any policy associated to g_c^*

Augment the reward in $\widetilde{\mathcal{M}}_k^\eta \Rightarrow \widetilde{\mathcal{M}}_k^\ddagger$

2. $\mathcal{A}_k$ is expanded by duplicating every action i.e., $B_{r,k}^\ddagger(s,a) = [0, \max\{B_r^k(s,a)\}]$
   - $\widetilde{L}^\eta v = \widetilde{L}^\ddagger v$ and $\widetilde{T}_c^\eta v = \widetilde{T}_c^\ddagger v$
     $g_c^*(\widetilde{\mathcal{M}}_k^\eta) = g_c^*(\widetilde{\mathcal{M}}_k^\ddagger)$
   - $\widetilde{T}_c$ is *globally* feasible
     $\forall v$ s.t. $sp\{v\} \le c$, $\exists \delta$ s.t. $sp\left\{\widetilde{L}_\delta^\ddagger v\right\} \le c$
   - $\widetilde{\mathcal{M}}_k^\ddagger$ is unichain
     $\Longrightarrow\ g_c^* \longmapsto \widetilde{\pi}_k$ (a policy exists)

SCAL computes

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

well defined problem and admits a maximizer $\pi_c^*(\widetilde{\mathcal{M}}_k^\ddagger)$
$\Longrightarrow$ efficiently computed using SCOPT

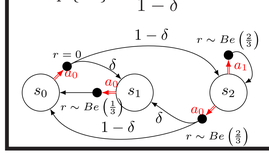
Regret

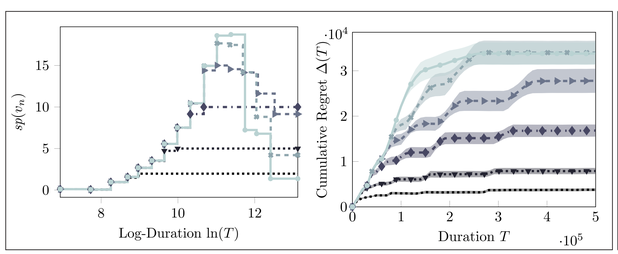
$$\Delta(\text{SCAL}, T) = \widetilde{O}\left(\min\{D, c\}\sqrt{\Gamma SAT}\right)$$

Numerical Experiments

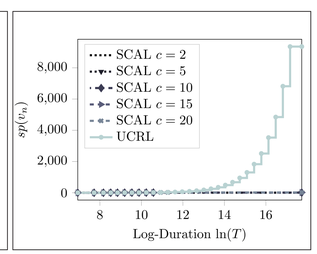
SIMPLE 2D DOMAIN

- ▶ δ regulates the complexity
  - $g^* = 2/3$
  - $D \approx 1/\delta$
  - $sp\{h^*\} = \frac{1}{1-\delta}$



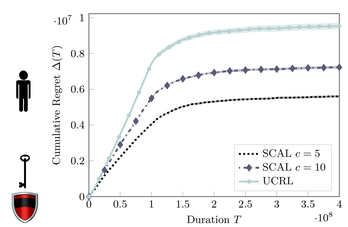


$\delta = 0.005 \Longrightarrow$ communicating MDP



$\delta = 0.0 \Longrightarrow s_1$ transient and $D = \infty$

WAY TO FREEDOM



$S = 360$, $A = 8$, $D \approx 250$ (communicating) and $sp\{h^*\} \approx 3.2$