Regret Minimization in Reinforcement Learning under Bias Span Constraint

Matteo Pirotta
Facebook AI Research, Paris (FR)

Based on the joint work with Jian Qian, Ronan Fruit and Alessandro Lazaric
Reinforcement Learning

[Sutton and Barto, 1998]

learning what to do-how to map situations to actions-so as to maximize a numerical reward signal

A framework for learning by interaction
What is the difference with optimal control? Reinforcement Learning is optimal control in **unknown** MDPs

⚠️ exploration-exploitation trade-off
**GO game**
[Mnih et al., 2015]
4.9 million games of self-play

**ATARI Games**
[Mnih et al., 2013]

- train data = 10 million frames
- 1 epoch = 500000 minibatch updates (≈30 minutes of games)
GO game
[Mnih et al., 2015]
4.9 million games of self-play

ATARI Games
[Mnih et al., 2013]
train data = 10 million frames
1 epoch = 500000 minibatch updates (≈30 minutes of games)

⚠ Many RL algorithms are inefficient in the "collection" and "use" of samples
**Limitations**

**Model-free**
No explicit representation of the system

**Poor Exploration**
Non effective action selection

---

\( \epsilon \)-greedy

\[
a = \begin{cases} 
\arg \max_{a} Q^\pi(s,a) & \text{w.p. } 1 - \epsilon \\
\text{random action} & \text{w.p. } \epsilon 
\end{cases}
\]

**Softmax**

\[
P(a|s) = \frac{e^{Q^\pi(s,a)/\tau}}{\sum_{a'} e^{Q^\pi(s,a')/\tau}}
\]
Limitations

Model-free
No explicit representation of the system

$\epsilon$-greedy

$$a = \begin{cases} \arg \max_a Q^\pi(s, a) & \text{w.p. } 1 - \epsilon \\ \text{random action} & \text{w.p. } \epsilon \end{cases}$$

Softmax

$$P(a|s) = \frac{e^{Q^\pi(s, a)}/\tau}{\sum_{a'} e^{Q^\pi(s, a')}/\tau}$$

Poor Exploration
Non effective action selection
Limitations (cont’d)

- **Dithering effect**: stochastic exploration
- **Policy shift**: policy is changed at every step, no time-consistency (e.g., Q-learning)
Limitations (cont’d)

- **Dithering effect**: stochastic exploration
- **Policy shift**: policy is changed at every step, no time-consistency (e.g., Q-learning)

We need *directed* and *consistent* exploration!
Limitations (cont’d)

- **Dithering effect**: stochastic exploration
- **Policy shift**: policy is changed at every step, no time-consistency (e.g., Q-learning)

We need *directed* and *consistent* exploration!

**SOLUTION:** Optimism in face of uncertainty principle
OFU Example

Thanks Ronan Fruit for the example
The agent does not know what is at the top of the mountain

Thanks Ronan Fruit for the example
OFU Example

Maybe there is nothing interesting...

Thanks Ronan Fruit for the example
OFU Example

... or maybe there is!

Thanks Ronan Fruit for the example
Optimism pushes for exploration by assuming the best world!
The higher the mountain the more challenging the exploration!

Thanks Ronan Fruit for the example
OFU Example

The higher the mountain the more challenging the exploration!
- Intrinsic difficulty of exploration-exploitation in RL!

Thanks Ronan Fruit for the example
The higher the mountain the more challenging the exploration!
- Intrinsic difficulty of exploration-exploitation in RL!
- Unavoidable except if we can exploit some prior knowledge!

Thanks Ronan Fruit for the example
The raspberries do not grow above a certain altitude!

Thanks Ronan Fruit for the example
The raspberries do not grow above a certain altitude!

Questions of this talk:
- Can we exploit prior knowledge for exp-exp?
- Is it necessary/mandatory?

Thanks Ronan Fruit for the example
We consider a finite MDP $M = \{S, A, p, r\}$

- $S$ is the finite state space ($S = |S| < +\infty$)
- $A$ is the finite action space ($A = |A| < +\infty$)
- $p(s'|s, a)$ is the transition kernel
- $r(s, a) \in [0, 1]$ is the reward
We consider a finite MDP $M = \{S, A, p, r\}$

- $S$ is the finite state space ($S = |S| < +\infty$)
- $A$ is the finite action space ($A = |A| < +\infty$)
- $p(s'|s, a)$ is the transition kernel
- $r(s, a) \in [0, 1]$ is the reward

Unknown!

On-line learning problem
Setting

We consider a finite MDP $M = \{S, A, p, r\}$

- $S$ is the finite state space ($S = |S| < +\infty$)
- $A$ is the finite action space ($A = |A| < +\infty$)
- $p(s'|s, a)$ is the transition kernel
- $r(s, a) \in [0, 1]$ is the reward

**Goal:** Learn the optimal policy $\pi^* : S \rightarrow \mathcal{P}(A)$
We consider a finite MDP $M = \{S, A, p, r\}$

- $S$ is the finite state space ($S = |S| < +\infty$)
- $A$ is the finite action space ($A = |A| < +\infty$)
- $p(s'|s, a)$ is the transition kernel
- $r(s, a) \in [0, 1]$ is the reward

**Goal:** Learn the optimal policy $\pi^* : S \rightarrow \mathcal{P}(A)$
Average Reward (the gain)

Average expected reward or gain

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

Optimal gain $g^*$ and optimal policy $\pi^*$

$$\pi^* := \arg \max_{\pi} g^\pi_M(s)$$

$$g^* := g^{\pi^*}_M(s) = \max_{\pi} g^\pi_M(s)$$
Average Reward (the bias)

\[ h^\pi_M(s) := \lim_{T \to +\infty} \mathbb{E} \left[ \sum_{t=1}^{T} (r(s_t, \pi(s_t)) - g^\pi_M(s_t)) \right] \]
Average Reward (the bias)

\[ h^\pi_M(s) := \lim_{T \to +\infty} E \left[ \sum_{t=1}^{T} \left( r(s_t, \pi(s_t)) - g^\pi_M(s_t) \right) \right] \]

"transient" reward

difference between immediate reward and asymptotic reward

"stationary" reward

Optimality Equation

\[ h^* + g^* e = Lh^* \]

\[ = \max_a \{ r(s, a) + p(\cdot|s, a)^T h^* \} \]
Optimal gain and bias span

- Remember the “fruity” example!

- Gain $g^*$ $\iff$ preferred fruit (raspberry $\gg$ apple)

- Bias span $sp \{h^*\}$ $\iff$ altitude of the mountain

Thanks Ronan Fruit for the example
Optimal gain and bias span

- Remember the “fruity” example!
- Gain $g^*$ $\iff$ preferred fruit (raspberry $\gg$ apple)
- Bias span $sp \{h^*\} \iff$ altitude of the mountain

Thanks Ronan Fruit for the example
Optimal gain and bias span

- Remember the "fruity" example!
- Gain $g^*$ $\iff$ preferred fruit (raspberry $\gg$ apple)
- Bias span $sp \{h^*\}$ $\iff$ altitude of the mountain

Thanks Ronan Fruit for the example
Optimal gain and bias span

- Remember the “fruity” example!
- Gain $g^*$ $\iff$ preferred fruit (raspberry $\gg$ apple)
- Bias span $sp \{h^*\} \iff$ altitude of the mountain

Thanks Ronan Fruit for the example

Matteo Pirotta
Optimal gain and bias span

- Remember the “fruity” example!
- Gain $g^*$ $\iff$ preferred fruit (raspberry $\gg$ apple)
- Bias span $sp\{h^*\} \iff$ altitude of the mountain

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

$s^p\{h^*\}$ characterizes the complexity of the problem!
Optimal gain and bias span

- Remember the “fruity” example!
- Gain $g^*$ $\iff$ preferred fruit (raspberry $\gg$ apple)
- Bias span $sp \{h^*\}$ $\iff$ altitude of the mountain
- Prior knowledge $c \geq sp \{h^*\}$ $\iff$ maximum altitude where raspberries can grow

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

$sp \{h^*\}$ characterizes the complexity of the problem!
Optimism in Face of Uncertainty (OFU)

When you are uncertain, consider the **best possible world**

OFU in RL

\[ t = 0 \]

\[ \text{for} \ \text{episode } k = 1, 2, \ldots \ \text{do} \]

Optimistic Planning $\rightarrow \pi_k$

\[ \mathcal{H}_{k+1} = \mathcal{H}_k \]

\[ \text{while } \text{not enough knowledge } \text{do} \]

Take action $a_t \sim \pi_k(\cdot|s_t)$

Observe reward $r_t$ and next state $s_{t+1}$

Update $\mathcal{H}_{k+1} = \mathcal{H}_{k+1} \cup (s_t, a_t, r_t, s_{t+1})$

end

\[ \text{Execute policy} \]

Matteo Pirotta
OFU in RL

\[
t = 0
\]
\[
\text{for } \text{episode } k = 1, 2, \ldots \text{ do}
\]
\[
\quad \text{Optimistic Planning } \rightarrow \pi_k
\]
\[
\quad \mathcal{H}_{k+1} = \mathcal{H}_k
\]
\[
\quad \text{while } \text{not enough knowledge } \text{ do}
\]
\[
\quad \quad \text{Take action } a_t \sim \pi_k(\cdot|s_t)
\]
\[
\quad \quad \text{Observe reward } r_t \text{ and next state } s_{t+1}
\]
\[
\quad \quad \text{Update } \mathcal{H}_{k+1} = \mathcal{H}_{k+1} \cup (s_t, a_t, r_t, s_{t+1})
\]
\[
\quad \text{end}
\]
\[
\text{end}
\]
\[
\text{ Execute policy}
\]

- provides consistency
- avoids policy shift
OFU in RL

$t = 0$
for episode $k = 1, 2, \ldots$ do

- Optimistic Planning $\rightarrow \pi_k$

$\mathcal{H}_{k+1} = \mathcal{H}_k$

while not enough knowledge do

- Take action $a_t \sim \pi_k(\cdot | s_t)$
- Observe reward $r_t$ and next state $s_{t+1}$
- Update $\mathcal{H}_{k+1} = \mathcal{H}_{k+1} \cup (s_t, a_t, r_t, s_{t+1})$

end

end

Plausible MDPs

1. Construct a set of plausible MDPs (high-confidence)
2. Select the MDP with highest gain

- e.g., UCRL [Jaksch et al., 2010], REGAL [Bartlett and Tewari, 2009], SCAL [Fruit, P., Lazaric Ortner; 2018b], TUCRL [Fruit, P., Lazaric, 2018a]

Execute policy

- provides consistency
- avoids policy shift

Matteo Pirotta
OFU in RL

\[ t = 0 \]

**for episode** \( k = 1, 2, \ldots \) **do**

**Optimistic Planning** → \( \pi_k \)

\[ \mathcal{H}_{k+1} = \mathcal{H}_k \]

**while** not enough knowledge **do**

- Take action \( a_t \sim \pi_k(\cdot|s_t) \)
- Observe reward \( r_t \) and next state \( s_{t+1} \)
- Update \( \mathcal{H}_{k+1} = \mathcal{H}_{k+1} \cup (s_t, a_t, r_t, s_{t+1}) \)

**end**

**end**

---

**Plausible MDPs**

1. Construct a set of plausible MDPs (high-confidence)
2. Select the MDP with highest gain

   e.g., UCRL [Jaksch et al., 2010], REGAL [Bartlett and Tewari, 2009], SCAL [Fruit, P., Lazaric Ortner; 2018b], TUCRL [Fruit, P., Lazaric, 2018a]

**Exploration Bonus**

- Compute the optimal policy of the empirical MDP plus *bonus*

- The bonus is an additive term to the reward

   e.g., MBIE-EB [Strehl and Littman, 2008], UCBV-1 [Azar et al., 2017], vUCQ [Kakade et al., 2018], SCAL\(^+\) [Qian, Fruit, P., Lazaric; 2018]

---

Matteo Pirotta
Plausible MDPs: Confidence intervals

Estimated trans. (MLE): \( \bar{p}_k(s'|s, a) = \frac{N_k(s, a, s')}{N_k(s, a)} \)

\[
\| \bar{p}_k(\cdot|s, a) - p_k(\cdot|s, a) \|_1 \leq \beta_{p,k}(s, a) \approx \sqrt{\frac{S \ln(1/\delta)}{N_k(s, a)}}
\]

Admissible transitions

number of visits in \((s, a)\)

Based on Hoeffding [Klenke and Loève, 2013] or empirical Bernstein concentration inequalities [Audibert et al., 2007]
Plausible MDPs: Confidence intervals

Estimated trans. (MLE):  $\tilde{p}_k(s'|s,a) = \frac{N_k(s,a,s')}{N_k(s,a)}$

$$\left\| \tilde{p}_k(\cdot|s,a) - \bar{p}_k(\cdot|s,a) \right\|_1 \leq \beta_{p,k}(s,a) \approx \sqrt{S \frac{\ln(1/\delta)}{N_k(s,a)}}$$

Admissible transitions

Based on Hoeffding [Klenke and Lo`eve, 2013] or empirical Bernstein concentration inequalities [Audibert et al., 2007]
Plausible MDPs: Optimistic Planning

- **UCRL** [Jaksch et al., 2010]

\[(M_k, \pi_k) \in \arg \max_{M \in \mathcal{M}_t, \pi : S \rightarrow \mathcal{P}(A)} g^\pi_M\]

**MDP space**

- \(\tilde{\mathcal{M}}_1\)
- \(\mathcal{M}_1\)
- \(M^*\)
Plausible MDPs: Optimistic Planning

UCRL [Jaksch et al., 2010]

\[(M_k, \pi_k) \in \arg \max_{M \in \mathcal{M}_t, \pi: \mathcal{S} \to \mathcal{P}(A)} g^\pi_M\]

MDP space

\[\tilde{\mathcal{M}}_1, M_1, M^*, \overline{M}_1\]
Plausible MDPs: Optimistic Planning

UCRL [Jaksch et al., 2010]

\[(M_k, \pi_k) \in \arg \max_{M \in \mathcal{M}_t, \pi : S \to \mathcal{P}(A)} g^\pi_M\]
Plausible MDPs: Optimistic Planning

UCRL [Jaksch et al., 2010]

\[
(M_k, \pi_k) \in \arg\max_{M \in M_t, \pi : S \to P(A)} g_M^\pi
\]

MDP space

MDP with highest gain

\[
M_k \in \arg\max_{M \in M_k} \{g_M^*\}
\]

Optimal policy of \( M_k \):

\[
\pi_k \in \arg\max_{\pi : S \to P(A)} \{g_M^\pi\}
\]
Plausible MDPs: Optimistic Planning

UCRL [Jaksch et al., 2010]

\[
(M_k, \pi_k) \in \arg\max_{M \in \mathcal{M}_t, \pi: S \to \mathcal{P}(A)} g_M^\pi
\]

MDP space

MDP with highest gain

\[
M_k \in \arg\max_{M \in \mathcal{M}_k} \{g_M^*\}
\]

Optimal policy of \(M_k\)

\[
\pi_k \in \arg\max_{\pi: S \to \mathcal{P}(A)} \{g_{M_k}^\pi\}
\]
Plausible MDPs: Optimistic Planning

- **SCAL** [Fruit, P., Lazaric, Ortner; 2018b]

\[(M_k, \pi_k) \in \arg \max_{M \in M_k, \pi \in \Pi_C(M)} \{g^\pi_M\}\]

\[\Pi_C(M) := \{\pi : S \rightarrow \mathcal{P}(A) : \text{sp}\{h^\pi_M\} \leq c\}\]

A *regularized* version was proposed by Bartlett and Tewari [2009] but no solution algorithm is known.

⚠️ this is a *constrained* optimization problem
A regularized version was proposed by Bartlett and Tewari [2009] but no solution algorithm is known.

This is a constrained optimization problem.

Yet, it can be solved: SCOPT [Fruit, P., Lazaric, Ortner; 2018b]

Lots of technical details: e.g., stochastic policy, feasibility, convergence
Problems

1. Optimism may be a little bit *loose*
2. Need to plan on an extended MDP (i.e., on a set of MDPs)
   - Extended Value Iteration (EVI) [Strehl and Littman, 2008, Jaksch et al., 2010] for UCRL
     \[
     v_{n+1} = \tilde{L}v_n := \max_{a \in A} \left\{ \max_{r \in \beta_{r,k}(s,a)} r + \max_{p \in \beta_{p,k}(s,a)} p(s,a)^T v_n \right\}
     \]  (1)
   - ScOPT for SCAL
3. Complicated to generalize outside finite MDPs
Problems

1. Optimism may be a little bit *loose*

2. Need to plan on an *extended MDP* (i.e., on a set of MDPs)
   - Extended Value Iteration (EVI) [Strehl and Littman, 2008, Jaksch et al., 2010] for UCRL
     \[
     v_{n+1} = \tilde{L}v_n := \max_{a \in A} \left\{ \max_{r \in \beta_{r,k}(s,a)} r + \max_{p \in \beta_{p,k}(s,a)} p(\cdot|s,a)^T v_n \right\}
     \] (1)
   - ScOPT for SCAL

3. Complicated to generalize outside finite MDPs

SOLUTION
exploration bonus
Exploration Bonus: the optimistic empirical MDP

Empirical MDP: $\hat{M}_k = \{S, A, \overline{p}_k, \overline{r}_k \}$

- Consider MLE of transitions $\overline{p}_k$ and rewards $\overline{r}_k$
- Optimism is obtained by an exploration bonus

$$b_k(s, a) \approx (c + r_{\text{max}}) \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}}$$

- SCAL$^+$ [Qian, Fruit, P., Lazaric, 2018c] plans on a single MDP

$$\pi_k \in \arg \max_{\pi \in g^{\pi}_{\hat{M}_k}}$$
Exploration Bonus: the optimistic empirical MDP

Optimistic Empirical MDP:

\[ \hat{M}_k = \{S, A, \bar{p}_k, \bar{r}_k + b_k \} \]

- Consider MLE of transitions \( \bar{p}_k \) and rewards \( \bar{r}_k \)
- Optimism is obtained by an exploration bonus

\[ b_k(s, a) \approx (c + r_{\text{max}}) \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

- SCAL\(^+\) [Qian, Fruit, P., Lazaric, 2018c] plans on a single MDP

\[ \pi_k \in \arg \max_{\pi \in \pi} g^{\pi}_{\hat{M}_k} \]
Exploration Bonus: the optimistic empirical MDP

\[ \hat{M}_k = \{ S, A, \bar{p}_k, \bar{r}_k + b_k \} \]

- Consider MLE of transitions \( \bar{p}_k \) and rewards \( \bar{r}_k \)
- Optimism is obtained by an exploration bonus
  \[ b_k(s, a) \approx (c + r_{max}) \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]
  
- \( \text{SCAL}^+ \) [Qian, Fruit, P., Lazaric, 2018c] plans on a single MDP
  \[ \pi_k \in \arg \max_{\pi \in \Pi_c(\hat{M}_k)} g_{\pi}^{\hat{M}_k} \]
Exploration Bonus: the optimistic empirical MDP

\[ \widehat{M}_k = \{ S, A, \bar{p}_k, \bar{r}_k + b_k \} \]

- Consider MLE of transitions \( \bar{p}_k \) and rewards \( \bar{r}_k \)
- Optimism is obtained by an exploration bonus

\[ b_k(s, a) \approx (c + r_{\max}) \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

- SCAL\(^+\) [Qian, Fruit, P., Lazaric, 2018c] plans on a single MDP

\[ \pi_k \in \arg \max_{\pi \in \Pi_c(\widehat{M}_k)} g_{\widehat{M}_k}^\pi \]

Still a Span-Constrained Optimization

\[ \Pi_c(M) := \{ \pi : S \to \mathcal{P}(A) : sp \{ h_M^\pi \leq c \} \]
Exploration bonus

\[ |r(s, a) - \bar{r}_k(s, a)| \lesssim r_{\text{max}} \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

\[ |(p(\cdot|s, a) - \bar{p}_k(\cdot|s, a))^T h^*| \lesssim c \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

Bellman Operator of $\widehat{M}_k$

\[ \widehat{L}h^* = \max_{a \in A} \{ \bar{r}_k(s, a) + \bar{p}_k(\cdot|s, a)^T h^* \} \]

\[ = \max_{a \in A} \left\{ \bar{r}_k(s, a) + r_{\text{max}} \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} + \bar{p}_k(\cdot|s, a)^T h^* + c \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \right\} \]

\[ \geq r(s, a) \]

\[ \geq p(\cdot|s, a)^T h^* \]

\[ \geq Lh^* \]
Exploration bonus

\[ | r(s, a) - \bar{r}_k(s, a) | \lesssim r_{\max} \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

\[ | (p(\cdot|s, a) - \bar{p}_k(\cdot|s, a))^\top h^* | \lesssim c \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

Bellman Operator of $\hat{M}_k$

\[
\hat{L}h^* = \max_{a \in A} \{ \bar{r}_k(s, a) + b_k(s, a) + \bar{p}_k(\cdot|s, a)^\top h^* \} 
\]

\[
\geq \max_{a \in A} \left\{ \bar{r}_k(s, a) + r_{\max} \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} + \bar{p}_k(\cdot|s, a)^\top h^* + c \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \right\} 
\]

\[
\geq r(s, a) + \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} + p(\cdot|s, a)^\top h^* 
\]

\[
\geq Lh^* 
\]
Exploration bonus

\[ | r(s, a) - \bar{r}_k(s, a) | \approx r_{\text{max}} \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

\[ | (p(\cdot|s, a) - \bar{p}_k(\cdot|s, a))^T h^* | \approx c \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

Bellman Operator of \( \hat{M}_k \)

\[ \hat{L} h^* = \max_{a \in \mathcal{A}} \{ \bar{r}_k(s, a) + b_k(s, a) + \bar{p}_k(\cdot|s, a)^T h^* \} \]

\[ = \max_{a \in \mathcal{A}} \left\{ \bar{r}_k(s, a) + r_{\text{max}} \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} + \bar{p}_k(\cdot|s, a)^T h^* + c \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \right\} \]

\[ \geq r(s, a) \quad \geq p(\cdot|s, a)^T h^* \]

\[ \geq L h^* \]
Exploration bonus

\[ | r(s, a) - \bar{r}_k(s, a) | \lesssim r_{\text{max}} \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

\[ | (p(\cdot|s, a) - \bar{p}_k(\cdot|s, a))^\top h^* | \lesssim c \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

Bellman Operator of \( \hat{M}_k \)

\[ \hat{L}h^* = \max_{a \in A} \left\{ \bar{r}_k(s, a) + b_k(s, a) + \bar{p}_k(\cdot|s, a)^\top h^* \right\} \]

\[ \geq r(s, a) + p_k(\cdot|s, a)^\top h^* + \sqrt{c \frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

\[ \geq Lh^* \]
Exploration bonus

\[ | r(s, a) - \bar{r}_k(s, a) | \lesssim r_{\text{max}} \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

\[ | (p(\cdot|s, a) - \bar{p}_k(\cdot|s, a))^T h^* | \lesssim c \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \]

Bellman Operator of $\hat{M}_k$

\[ \hat{L}h^* = \max_{a \in A} \{ \bar{r}_k(s, a) + b_k(s, a) + \bar{p}_k(\cdot|s, a)^T h^* \} \]

\[ = \max_{a \in A} \left\{ \bar{r}_k(s, a) + r_{\text{max}} \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} + \bar{p}_k(\cdot|s, a)^T h^* + c \sqrt{\frac{\ln(t_k/\delta)}{N_k(s, a)}} \right\} \]

\[ \geq r(s, a) \]

\[ \geq p(\cdot|s, a)^T h^* \]

\[ \geq Lh^* \]

[Puterman, 1994] [Fruit, P., Lazaric, Ortner; 2018b] \[ \Rightarrow \]

\[ g_k = g_c^*(\hat{M}_k) \gtrsim g^* \]
Performance of a learning agent

\[ \Delta(\mathcal{A}, T) = \sum_{t=1}^{T} \left( g^* - r_t(s_t, a_t) \right) \]

Per step reward

\( g^* \) (optimal policy)

(cumulative) regret

learning curve

* different definition for finite-horizon problems
Performance of a learning agent

Regret

\[ \Delta(A, T) = \sum_{t=1}^{T} (g^* - r_t(s_t, a_t)) \]

Per step reward

Learning means **sublinear** regret

(cumulative) regret

learning curve

*different definition for finite-horizon problems
Regret of SCAL$^+$

**Theorem.** For any MDP $M$ such that $sp\{h^*\} \leq c$, with probability at least $1 - \delta$, the regret of SCAL$^+$ is bounded as

$$\Delta(\text{SCAL}^+, T) = O \left( S \sqrt{AT \ln \left( \frac{T}{\delta} \right)} \cdot c \right)$$
Regret of SCAL$^+$

**Theorem.** For any MDP $M$ such that $sp\{h^*\} \leq c$, with probability at least $1 - \delta$, the regret of SCAL$^+$ is bounded as

$$
\Delta(\text{SCAL}^+, T) = O \left( S \sqrt{AT \ln \left( \frac{T}{\delta} \right)} \cdot c \right)
$$

$D$ in UCRL

$\min\{c, D\}$ in SCAL
Regret of SCAL$^+$

**Theorem.** For any MDP $M$ such that $sp\{h^*\} \leq c$, with probability at least $1 - \delta$, the regret of SCAL$^+$ is bounded as

$$\Delta(\text{SCAL}^+, T) = O\left(S \sqrt{AT \ln \left(\frac{T}{\delta}\right)} \cdot c\right)$$

$D$ in UCRL

$\min\{c, D\}$ in SCAL

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

Mean arrival time in $s'$ starting in $s$
Theorem. For any MDP $M$ such that $\text{sp}\{h^*\} \leq c$, with probability at least $1 - \delta$, the regret of SCAL$^+$ is bounded as

$$\Delta(\text{SCAL}^+, T) = O \left( S \sqrt{AT \ln \left( \frac{T}{\delta} \right)} \cdot c \right)$$

- $\text{sp}\{h^*\} \leq D$ [Bartlett and Tewari, 2009]
- The gap can be arbitrarily big, e.g., $D = +\infty$ but $\text{sp}\{h^*\} < +\infty$
Why Exploration Bonus?

- Regret Minimization in continuous state MDPs: C-SCAL+
  - MDP (reward and transitions) is Hölder continuous (parameters $L$ and $\alpha$)
  - C-SCAL+ combines the idea of SCAL+ with state aggregation

\[
b(I, a) \approx \max\{c, r_{\text{max}}\}(1/\sqrt{N_k(s, a)} + LS^{-\alpha})
\]

- Regret bound: \( \Delta(C-\text{SCAL}^+, T) = \tilde{O}\left(\max\{c, r_{\text{max}}\}L\sqrt{AT^{(\alpha+2)/(2\alpha+2)}}\right) \)

For solutions based on plausible MDPs refer to [Ortner and Ryabko, 2012, Lakshmanan et al., 2015]. Not implementable in the current form. Hint: mix with SCAL.
Why Exploration Bonus?

- Exploration-exploitation at scale: deep reinforcement learning
  [Bellemare et al., 2016, Tang et al., 2017, Ostrovski et al., 2017, Martin et al., 2017]
  - Simple additive term to the reward, can be incorporated in any algorithm

\[
\tilde{r}(s, a) = r(s, a) + \sqrt{\frac{\beta}{N_k(\phi(s, a))}}
\]

- Use advanced discretization techniques \(\phi(s, a)\), e.g., hashing
Span-Constrained Planning

\[
\sup_{\pi \in \Pi_c(M)} \{ g^\pi \}
\]

\[
\Pi_c(M) := \{ \pi : S \rightarrow \mathcal{P}(A) : sp\{h_M^\pi\} \leq c \land sp\{g_M^\pi\} = 0 \}
\]

Connection with the exploration-exploitation framework

- **SCAL:** \( M := \tilde{M}_k \), an extended MDP with continuous actions \( \tilde{A}_k \)

\[
(M_k, \pi_k) \in \arg \max_{M \in M_k, \pi \in \Pi_c(M)} g_M^\pi \quad \text{equivalent} \quad \tilde{\pi}_k \in \arg \max_{\pi : S \rightarrow \mathcal{P}(\tilde{A}_k) \land sp\{h^\pi\} \leq c} g_M^\pi
\]

i.e., where the Bellman operator \( \tilde{L} \) is defined in Eq. 1

- **SCAL^+:** \( M := \hat{M}_k \) where \( \hat{L} \) is defined as in Eq. 2
Span-Constrained Planning

\[ \sup_{\pi \in \Pi_c(M)} \{ g^\pi \} \]

\[ \Pi_c(M) := \{ \pi : S \to \mathcal{P}(A) : sp \{ h^\pi_M \} \leq c \land sp \{ g^\pi_M \} = 0 \} \]

- NOT trivial optimization problem
- but apparently simple solution: SCOPT [Fruit, P., Lazaric, Ortner, 2018b]

\[ v_{n+1} = L v_n := \max_{a \in A} \left\{ r(s, a) + \sum_{s' \in S} p(s' | s, a) v_n(s') \right\} \]

\[ v_{n+1} \forall s \begin{cases} c & \text{if } v_{n+1}(s) \geq \min \{ v_{n+1} \} + c \\ v_{n+1}(s) & \text{otherwise} \end{cases} \]
Span-Constrained Planning

\[ \sup_{\pi \in \Pi_c(M)} \{ g^\pi \} \]

\[ \Pi_c(M) := \{ \pi : S \rightarrow P(A) : sp \{ h_M^\pi \} \leq c \wedge sp \{ g_M^\pi \} = 0 \} \]

- NOT trivial optimization problem
- but apparently simple solution: SCOPT \cite{FruitPiazzari2018b} i.e., “truncated” above

\[ v_{n+1} = L v_n := \max_{a \in A} \left\{ r(s, a) + \sum_{s' \in S} p(s'|s, a) v_n(s') \right\} \]

\[ v_{n+1} \forall s = \begin{cases} c & \text{if } v_{n+1}(s) \geq \min \{ v_{n+1} \} + c \\ v_{n+1}(s) & \text{otherwise} \end{cases} \]
Span-Constrained Planning

⚠️ Issues

- The associated one-step policy can be stochastic ... and may not exist
- Truncated value iteration (i.e., ScOpt) may not converge

Theorem. If

1. $L$ is a $(\gamma < 1)$-span contraction
2. All policies are unichain
3. $\forall v : sp\{v\} \leq c, \min_a \left\{ r(s, a) + p(s' | s, a)^T v \right\} \leq \min_{s'} \{ L v(s') \} + c$

then

- optimality equation: $T_c h^+ = h^+ + g^+ e$ and $g^+ = g^*_c$
- convergence: $\lim_{n \to \infty} T_c^{n+1} v_0 - T_c^n v_0 = g^+ e$
Consider a biased (but asymptotically consistent) estimator of the transition probabilities

\[
\hat{p}_k(s'|s, a) = \frac{N_k(s, a)p_k(s'|s, a)}{N_k(s, a) + 1} + \frac{1(s' = \bar{s})}{N_k(s, a) + 1}
\]

\(\Rightarrow\) SCOPT converges

**Problem:** there might not be any policy associated to \(g_c^*\)!

- Augment the reward: duplicate all the actions

\[\forall s \in S, a \in A_t, \text{ define } b \text{ such that } p(\cdot | s, b) = p(\cdot | s, a) \text{ and } r(s, b) = 0\]
How to force these properties in exp-exp

The estimated MDP

- Consider a biased (but asymptotically consistent) estimator of the transition probabilities

\[ \hat{p}_k(s'|s,a) = \frac{N_k(s,a)\bar{p}_k(s'|s,a)}{N_k(s,a) + 1} + \frac{1(s' = \overline{s})}{N_k(s,a) + 1} \]

\[ \implies \text{SCOPT converges} \]

**Problem:** there might not be any policy associated to \( g^*_c \)!

- Augment the reward: duplicate all the actions

\[ \forall s \in S, a \in A_t, \text{ define } b \text{ such that } p(\cdot|s,b) = p(\cdot|s,a) \text{ and } r(s,b) = 0 \]
Consider a biased (but asymptotically consistent) estimator of the transition probabilities
\[ \hat{p}_k(s'|s,a) = \frac{N_k(s,a)p_k(s'|s,a)}{N_k(s,a) + 1} + \frac{1(s' = \bar{s})}{N_k(s,a) + 1} \]

\[ \implies \text{SCOPT converges} \]

**Problem:** there might not be any policy associated to \( g_c^* \)!

- Augment the reward: duplicate all the actions

\[ \forall s \in S, a \in A_t, \text{ define } b \text{ such that } p(\cdot|s,b) = p(\cdot|s,a) \text{ and } r(s,b) = 0 \]
How to force these properties in exp-exp

The estimated MDP

- Consider a biased (but asymptotically consistent) estimator of the transition probabilities

\[
\hat{p}_k(s'|s, a) = \frac{N_k(s, a)\bar{p}_k(s'|s, a)}{N_k(s, a) + 1} + \frac{1(s' = \bar{s})}{N_k(s, a) + 1}
\]

\[\implies \text{SCOPT converges}\]

**Problem:** there might not be any policy associated to \(g^*_c\)!

- Augment the reward: duplicate all the actions

\[\forall s \in S, a \in A_t, \text{ define } b \text{ such that } p(\cdot|s, b) = p(\cdot|s, a) \text{ and } r(s, b) = 0\]

👍 When \(\hat{M}_k\) is perturbed and augmented, SCOPT converges to

\[g^+ \gtrsim g^*\]
The role of prior knowledge

- provides a sense of what it is realizable in the true MDP
- avoids over-optimism

This information is mandatory to define the exploration bonus

\[ |(p(\cdot|s, a) - \bar{p}_k(\cdot|s, a))^\top h^*| \leq \|p(\cdot|s, a) - \bar{p}_k(\cdot|s, a)\|_1 \|h^*\|_\infty \]

Intrinsic in other settings (infinite-horizon undiscounted, finite-horizon)
The role of prior knowledge

Intrinsic Horizon

<table>
<thead>
<tr>
<th>Setting</th>
<th>MDP parameter</th>
<th>Horizon</th>
<th>Knowledge</th>
<th>Exploration Bonus</th>
</tr>
</thead>
<tbody>
<tr>
<td>infinite-horizon discounted</td>
<td>$\gamma$</td>
<td>$\frac{1}{1-\gamma}$</td>
<td>$</td>
<td>Q(s, a)</td>
</tr>
<tr>
<td>finite-horizon</td>
<td>$H$</td>
<td>$H$</td>
<td>$</td>
<td>Q(s, a)</td>
</tr>
<tr>
<td>others</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

average reward

+\infty

??
## The role of prior knowledge

### Intrinsic Horizon

<table>
<thead>
<tr>
<th>Setting</th>
<th>MDP parameter</th>
<th>Horizon</th>
<th>Knowledge</th>
<th>Exploration Bonus</th>
</tr>
</thead>
<tbody>
<tr>
<td>infinite-horizon</td>
<td>$\gamma$</td>
<td>$\frac{1}{1-\gamma}$</td>
<td>$</td>
<td>Q(s,a)</td>
</tr>
<tr>
<td>discounted</td>
<td></td>
<td></td>
<td></td>
<td><strong>MBIE-EB</strong> [Strehl and Littman, 2008]</td>
</tr>
<tr>
<td>finite-horizon</td>
<td>$H$</td>
<td>$H$</td>
<td>$</td>
<td>Q(s,a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>UCBVI-1</strong> [Azar et al., 2017]</td>
</tr>
<tr>
<td>others</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average reward</td>
<td>?</td>
<td>$+\infty$</td>
<td></td>
<td>$sp {h^*} \leq c$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>assumption</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\tilde{\Theta} \left( c \sqrt{\frac{1}{N_k(s,a)}} \right)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>SCAL</strong> [Qian, Fruit, P., Lazaric, 2018]</td>
</tr>
</tbody>
</table>
Almost all the algorithms require prior knowledge.

<table>
<thead>
<tr>
<th>MDP</th>
<th>Algorithm</th>
<th>Properties/Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ergodic</td>
<td>KL-UCRL</td>
<td>$D &lt; +\infty$</td>
</tr>
<tr>
<td>Communicating</td>
<td>UCRL</td>
<td>$D = +\infty$ but we need $sp {h^*} \leq c$</td>
</tr>
<tr>
<td>Weakly Comm.</td>
<td>REGAL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCAL</td>
<td>$D = +\infty$ but we need $sp {h^*} \leq c$</td>
</tr>
<tr>
<td></td>
<td>SCAL+</td>
<td></td>
</tr>
<tr>
<td>Non Comm.</td>
<td>TUCRL</td>
<td>No assumptions but impossible to have logarithmic regret</td>
</tr>
</tbody>
</table>

[Puterman, 1994] Sec. 8.3
Outlook

span-constrained exp-exp ⇔ regularization

Open Questions?

- in practice
  - Constrained planning
  - Model-based planning

- in theory
  - Closing the gap between lower and upper bound
  - Exploration bonus with different algorithm structure
  - Model-free approaches
Thank you for the attention

facebook
Artificial Intelligence Research


