facebook
Artificial Intelligence Research

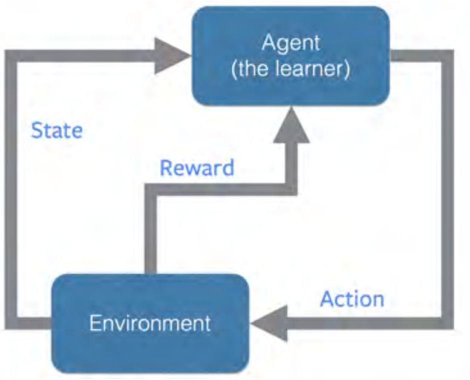
Regret Minimization in Reinforcement Learning under Bias Span Constraint

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

[Bertsekas, 1995, Puterman, 1994]

[Sutton and Barto, 1998]

What is the difference with optimal control?

Reinforcement Learning is optimal control in **unknown** MDPs

⚠ exploration-exploitation trade-off

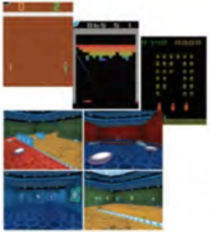
Abbeel and Schulman. Deep Reinforcement Learning Through Policy Optimization. Tutorial at NIPS 2016

Kohl and Stone, 2004

Ng et al, 2004

Tedrake et al, 2005

Kober and Peters, 2009



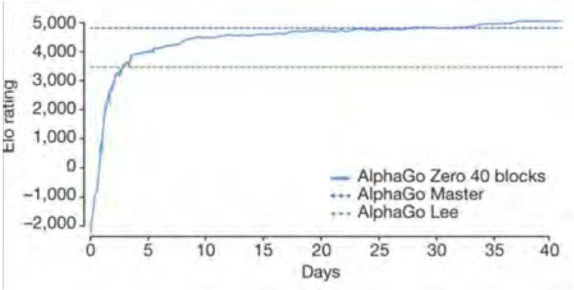
Mnih et al, 2015 (A3C)

Silver et al, 2014 (DPG)
Lillicrap et al, 2015 (DDPG)

Schulman et al, 2016 (TRPO + GAE)

Levine*, Finn*, et al, 2016 (GPS)

Silver*, Huang*, et al, 2016 (AlphaGo**)



GO game

[Mnih et al., 2015]

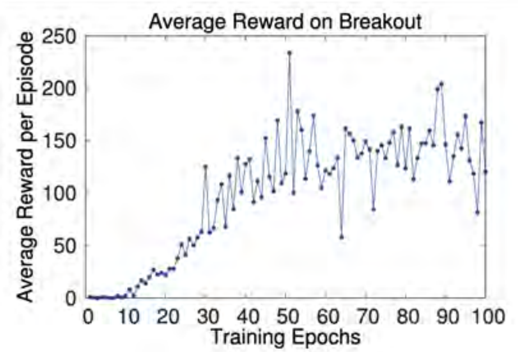
4.9 million games of self-play

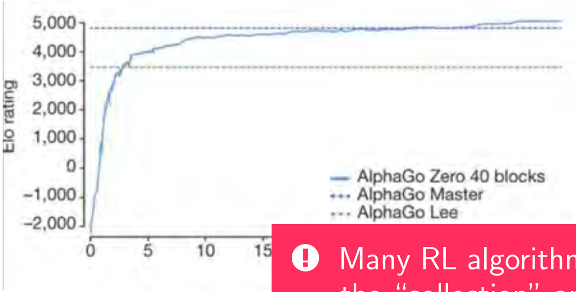
ATARI Games

[Mnih et al., 2013]

train data = 10 million frames

1 epoch = 500000 minibatch updates (≈30 minutes of games)





GO game

[Mnih et al., 2015]

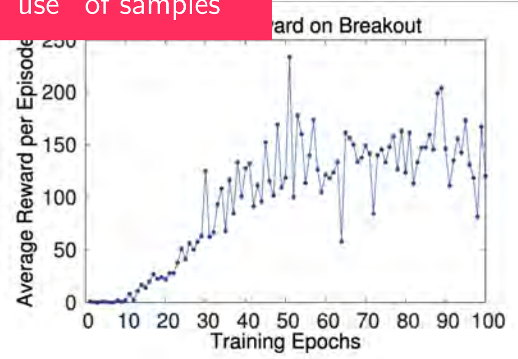
4.9 million games of self-play

> ❗ Many RL algorithms are inefficient in the "collection" and "use" of samples

ATARI Games

[Mnih et al., 2013]

train data = 10 million frames
1 epoch = 500000 minibatch updates (≈30 minutes of games)



Limitations

Model-free

No explicit representation of the system

Poor Exploration

Non effective action selection

ϵ-greedy

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

Softmax

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

Limitations

Model-free

No explicit representation of the system

Poor Exploration

Non effective action selection

ϵ-greedy

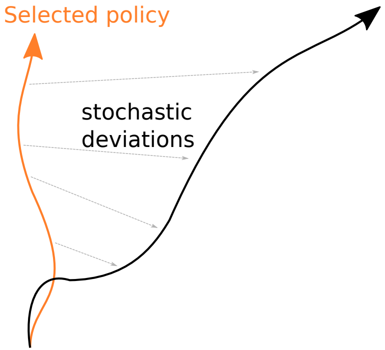
$$a = \begin{cases} \arg\max_a Q^\pi(s,a) & \text{w.p. } 1-\epsilon \\ \textit{random action} & \text{w.p. } \epsilon \end{cases}$$

Softmax

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

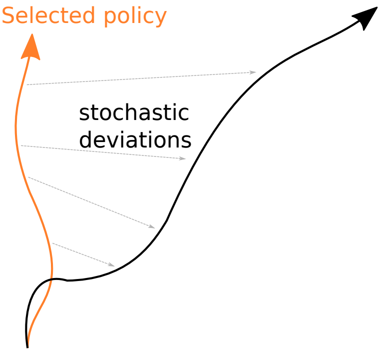
Limitations (cont'd)

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



Limitations (cont'd)

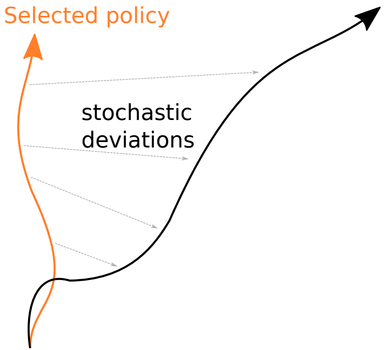
- Dithering effect: stochastic exploration
- Policy shift: policy is changed at every step, no time-consistency (e.g., Q-learning)



We need *directed* and *consistent* exploration!

Limitations (cont'd)

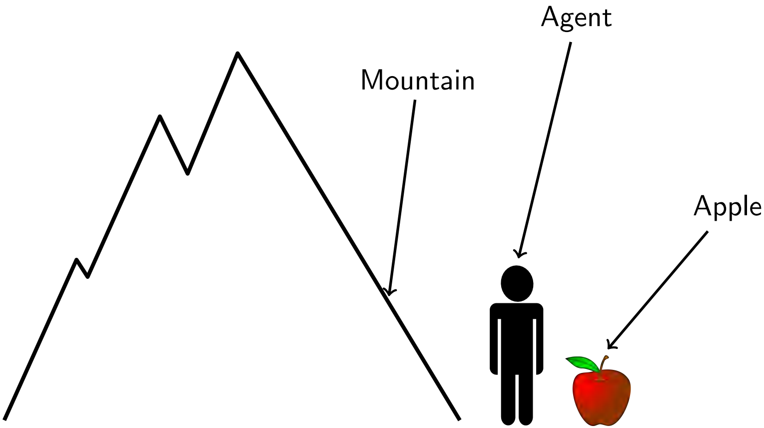
- Dithering effect: stochastic exploration
- Policy shift: policy is changed at every step, no time-consistency (e.g., Q-learning)



We need *directed* and *consistent* exploration!

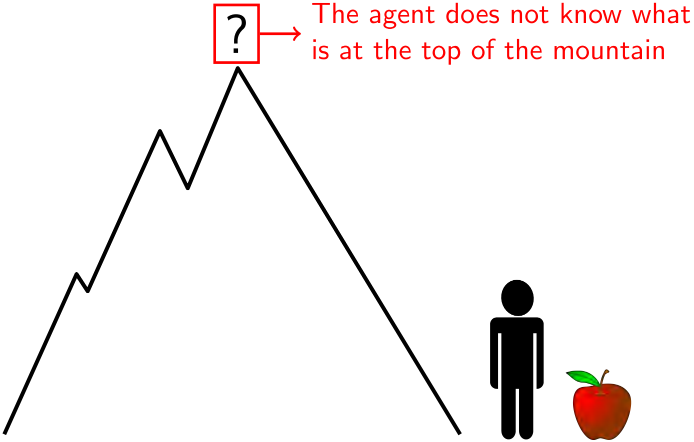
SOLUTION:
Optimism in face of uncertainty principle

OFU Example



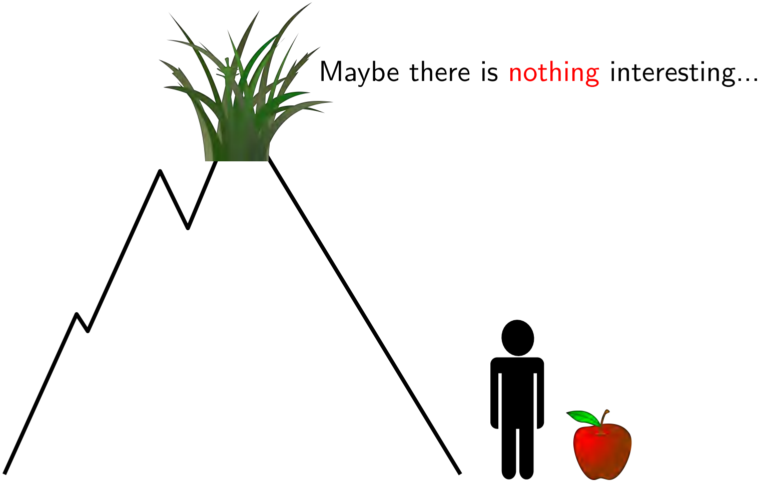
Thanks Ronan Fruit for the example

OFU Example



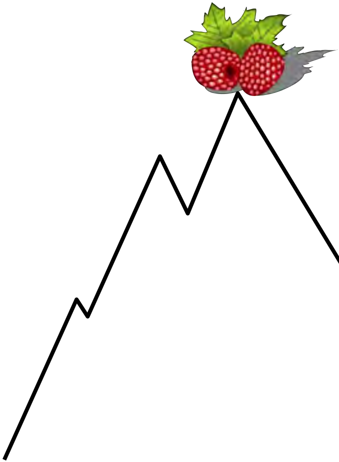
Thanks Ronan Fruit for the example

OFU Example

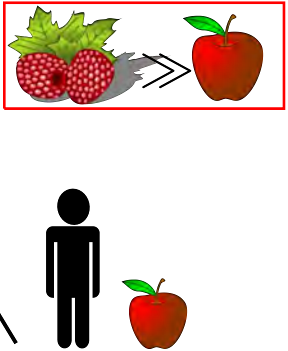


Thanks Ronan Fruit for the example

OFU Example

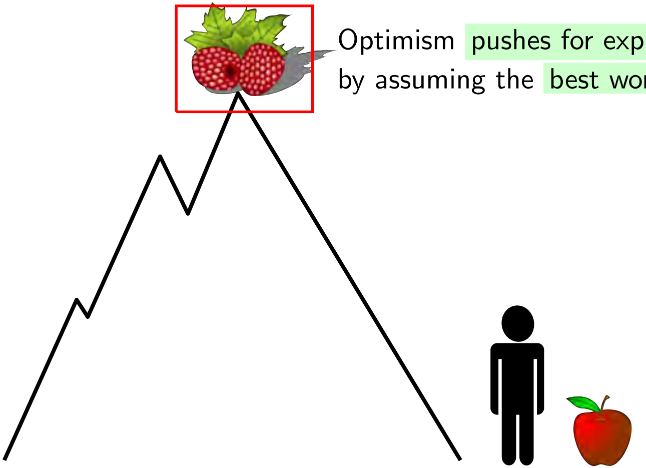


... or maybe there is!



Thanks Ronan Fruit for the example

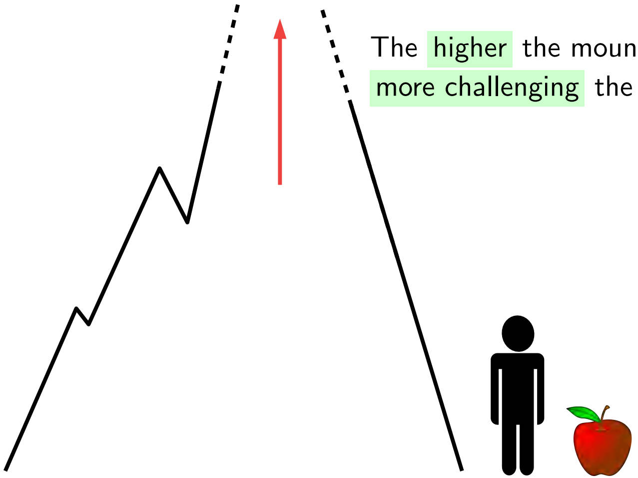
OFU Example



Optimism pushes for exploration
by assuming the best world!

Thanks Ronan Fruit for the example

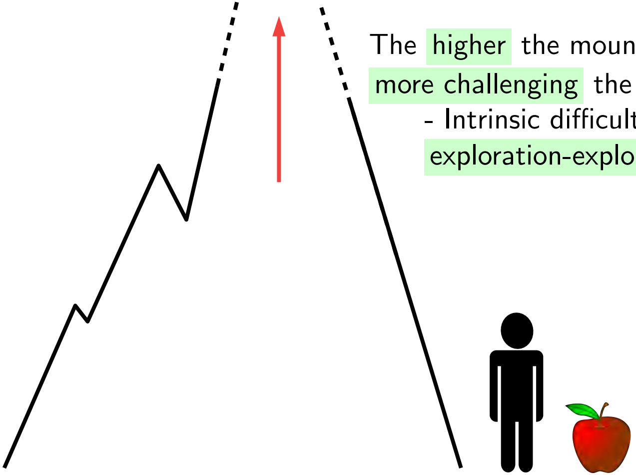
OFU Example



The higher the mountain the more challenging the exploration!

Thanks Ronan Fruit for the example

OFU Example

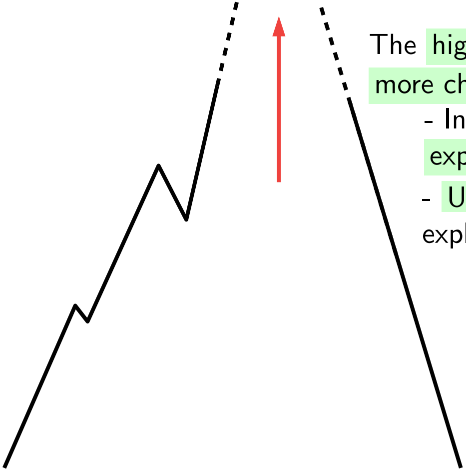


The higher the mountain the more challenging the exploration!

- Intrinsic difficulty of exploration-exploitation in RL!

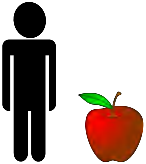
Thanks Ronan Fruit for the example

OFU Example



The higher the mountain the more challenging the exploration!

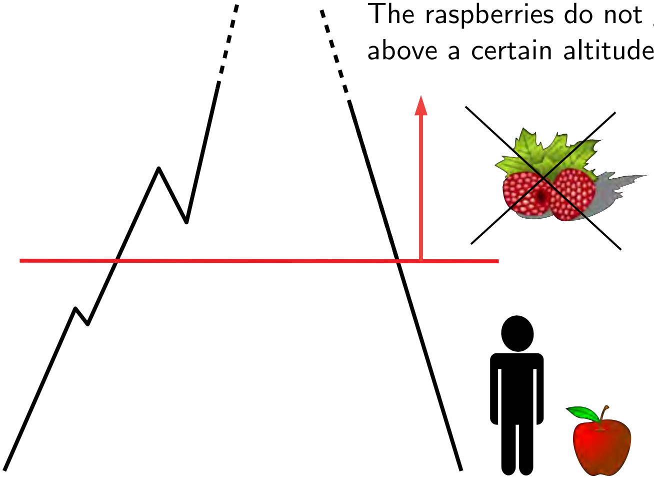
- Intrinsic difficulty of exploration-exploitation in RL!
- Unavoidable except if we can exploit some prior knowledge!



Thanks Ronan Fruit for the example

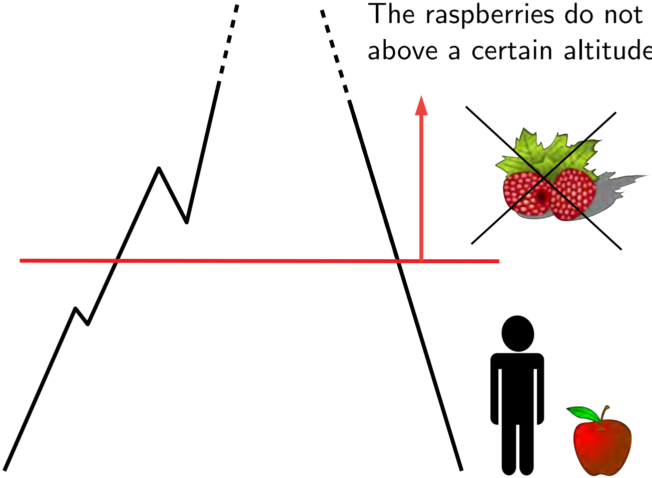
OFU Example

The raspberries do not grow
above a certain altitude!



Thanks Ronan Fruit for the example

OFU 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

Setting

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

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

Setting

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

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

$\longrightarrow$ Unknown!
On-line learning problem

Setting

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

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

$\rightarrow$ Unknown!
On-line learning problem

GOAL: Learn the optimal policy $\pi^* : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$

Setting

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

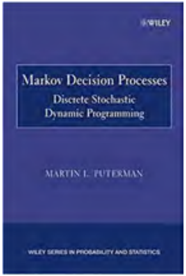
- $\mathcal{S}$ is the *finite* state space ($S = |\mathcal{S}| < +\infty$)
- $\mathcal{A}$ is the *finite* action space ($A = |\mathcal{A}| < +\infty$)
- $p(s'|s, a)$ is the transition kernel
- $r(s, a) \in [0, 1]$ is the reward

$\rightarrow$ Unknown! On-line learning problem

<u>GOAL:</u> Learn the optimal policy $\pi^* : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$

?

Average Reward (the gain)



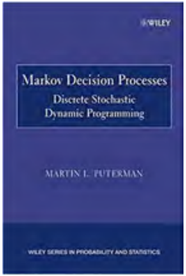
Average expected reward or gain

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

Optimal gain g^* and optimal policy π^*

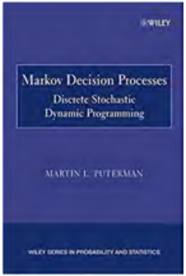
$$\pi^* := \arg\max_\pi \; g_M^\pi(s)$$
$$g^* := g_M^{\pi^*}(s) = \max_\pi \; g_M^\pi(s)$$

Average Reward (the bias)



$$h_M^\pi(s) \;:= \lim_{T\to+\infty} \mathbb{E}\left[\sum_{t=1}^{T}\Big(r(s_t,\pi(s_t)) - \; g_M^\pi(s_t)\;\Big)\right]$$

Average Reward (the bias)



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

"transient" reward
difference between immediate reward and asymptotic reward

"stationary" reward

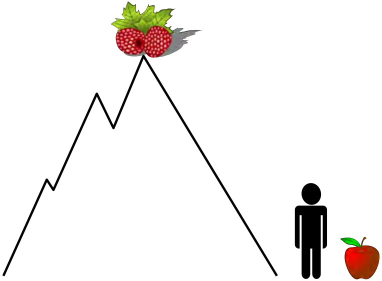
Optimality Equation

$$\begin{aligned} h^* + g^* e &= L h^* \\ &= \max_a \{ r(s,a) + p(\cdot|s,a)^\mathsf{T} h^* \} \end{aligned}$$

Optimal gain and bias span

Thanks Ronan Fruit for the example

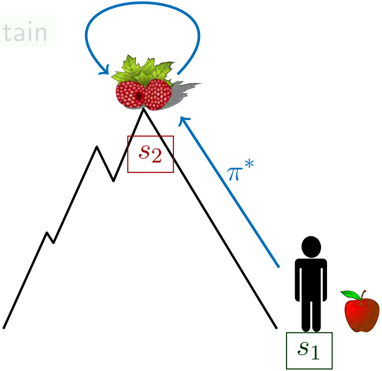
- Remember the "fruity" example!
- Gain g^* $\Longleftrightarrow$ preferred fruit (raspberry $\gg$ apple)
- Bias span $sp\{h^*\}$ $\Longleftrightarrow$ altitude of the mountain



Optimal gain and bias span

Thanks Ronan Fruit for the example

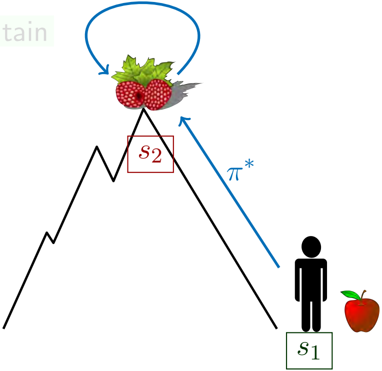
- Remember the "fruity" example!
- Gain g^* $\iff$ preferred fruit (raspberry $\gg$ apple)
- Bias span $sp\{h^*\}$ $\iff$ altitude of the mountain



Optimal gain and bias span

Thanks Ronan Fruit for the example

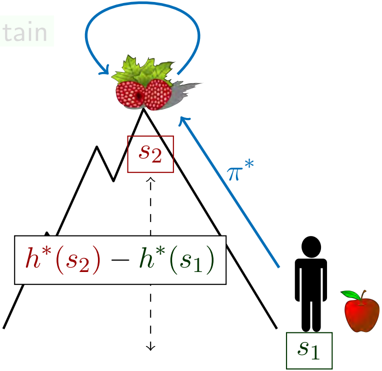
- Remember the "fruity" example!
- Gain g^* $\Longleftrightarrow$ preferred fruit (raspberry $\gg$ apple)
- Bias span $sp\{h^*\}$ $\Longleftrightarrow$ altitude of the mountain



Optimal gain and bias span

Thanks Ronan Fruit for the example

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



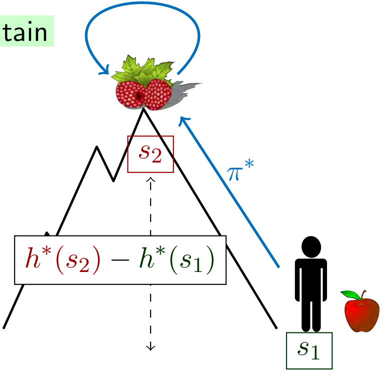
Optimal gain and bias span

Thanks Ronan Fruit for the example

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

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

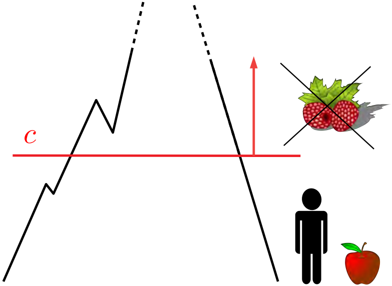
$sp\left\{h^*\right\}$ characterizes the complexity of the problem!



Optimal gain and bias span

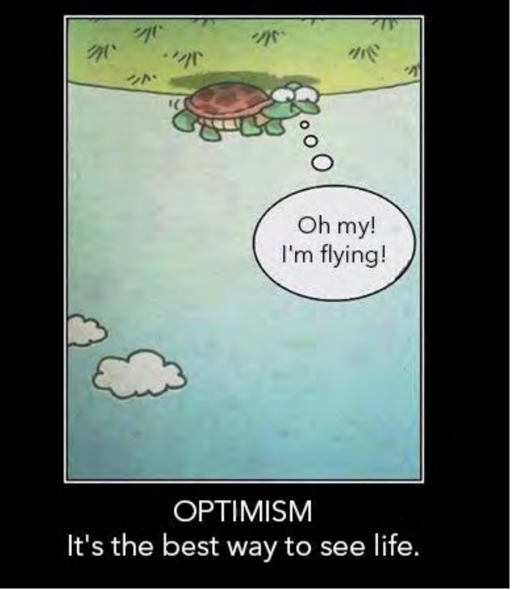
Thanks Ronan Fruit for the example

- 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



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

$sp\{h^*\}$ characterizes the complexity of the problem!



Optimism in Face of Uncertainty (OFU)

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

[Brafman and Tennenholtz, 2003, Strehl and Littman, 2008, Ortner, 2008, Jaksch et al., 2010, Bartlett and Tewari, 2009, Ortner and Ryabko, 2012, Osband et al., 2013, Abbasi-Yadkori and Szepesvári, 2015, Maillard et al., 2013, Gopalan and Mannor, 2015, Lakshmanan et al., 2015, Ouyang et al., 2017, Azar et al., 2017, Jin et al., 2018, Kakade et al., 2018, Agrawal and Jia, 2017], [Fruit et al., 2017, 2018a,b] and many more

Formally: $\boxed{g_k \gtrsim g^*}$

OFU in RL

$t = 0$
for *episode* $k = 1, 2, \ldots$ **do**
  Optimistic Planning $\to \pi_k$
  $\mathcal{H}_{k+1} = \mathcal{H}_k$
  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

Execute policy

OFU in RL

$t = 0$

for *episode* $k = 1, 2, \dots$ **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

Execute policy

- provides consistency
- avoids policy shift

OFU in RL

$t = 0$
for *episode* $k = 1, 2, \dots$ **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

Execute policy

- provides consistency
- avoids policy shift

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]

OFU in RL

$t = 0$
for *episode* $k = 1, 2, \dots$ **do**
Optimistic Planning $\to \pi_k$
$\mathcal{H}_{k+1} = \mathcal{H}_k$
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

Execute policy

- provides consistency
- avoids policy shift

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]

Plausible MDPs: Confidence intervals

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

$$\Big\| \underbrace{\widetilde{p}_k(\cdot|s,a)}_{\text{Admissible transitions}} - \overline{p}_k(\cdot|s,a) \Big\|_1 \leq \beta_{p,k}(s,a) \approx \sqrt{S \frac{\ln(1/\delta)}{\underbrace{N_k(s,a)}_{\text{number of visits in } (s,a)}}}$$

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

Plausible MDPs: Confidence intervals

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

$$\Big\| \underbrace{\widetilde{p}_k(\cdot|s,a)}_{\text{Admissible transitions}} - \overline{p}_k(\cdot|s,a) \Big\|_1 \leq \beta_{p,k}(s,a) \approx \sqrt{S\frac{\ln(1/\delta)}{\underbrace{N_k(s,a)}_{\text{number of visits in } (s,a)}}}$$

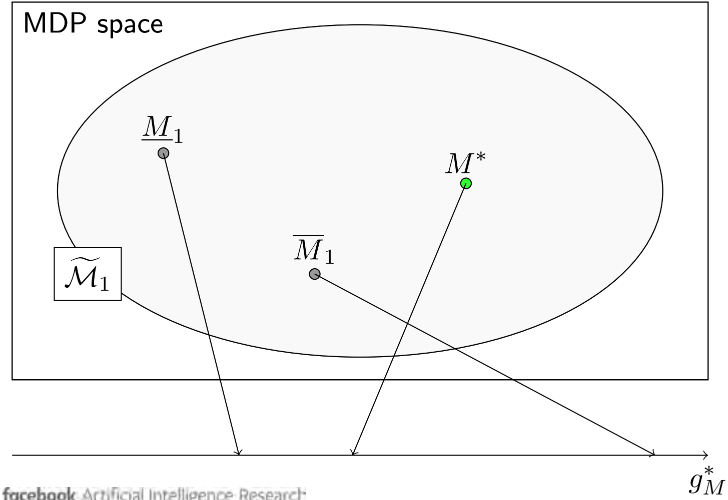
$$\Big| \widetilde{r}_k(s,a) - \overline{r}_k(s,a) \Big| \leq \beta_{r,k}(s,a) \approx r_{\max}\sqrt{\frac{\ln(1/\delta)}{N_k(s,a)}}$$

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

Plausible MDPs: Optimistic Planning

- UCRL [Jaksch et al., 2010]

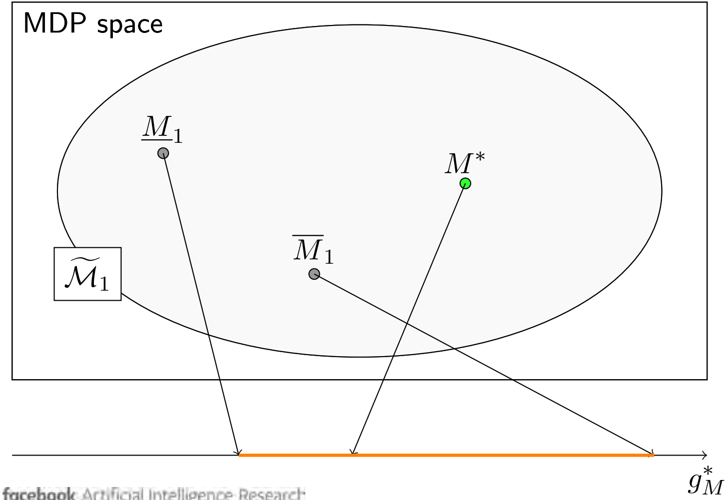
$$(M_k , \pi_k) \in \arg\max_{M \in \mathcal{M}_t,\ \pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})} g_M^\pi$$



Plausible MDPs: Optimistic Planning

- UCRL [Jaksch et al., 2010]

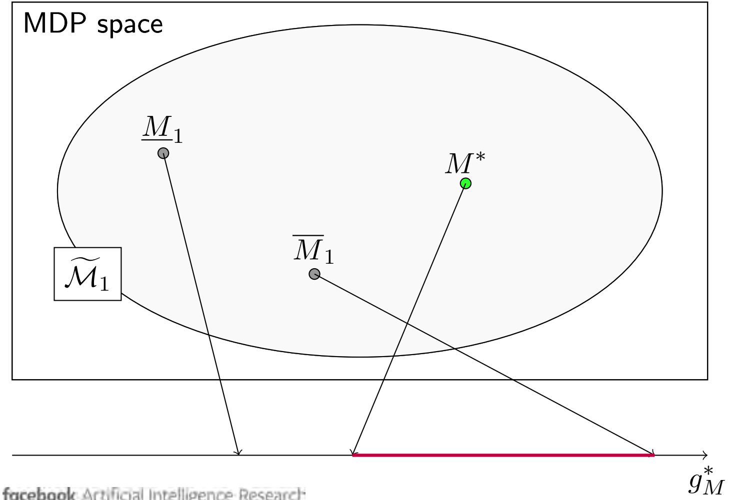
$$(M_k, \pi_k) \in \arg\max_{M \in \mathcal{M}_t,\ \pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})} g_M^{\pi}$$



Plausible MDPs: Optimistic Planning

- UCRL [Jaksch et al., 2010]

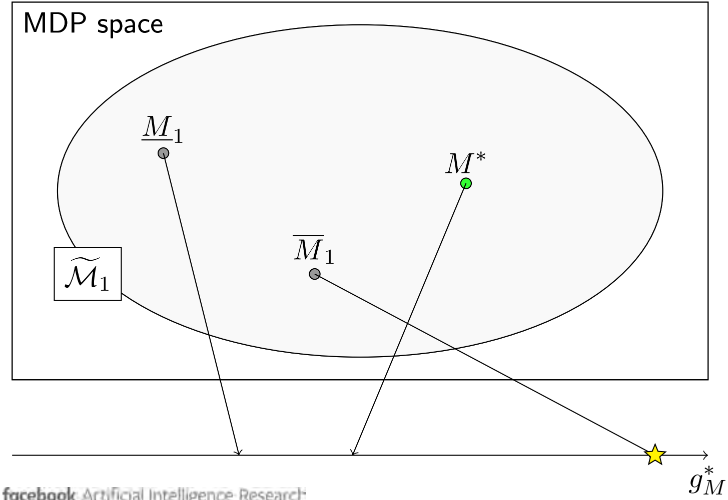
$$(M_k \;,\; \pi_k) \in \operatorname*{arg\,max}_{M \in \mathcal{M}_t,\; \pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})} g_M^\pi$$



Plausible MDPs: Optimistic Planning

- UCRL [Jaksch et al., 2010]

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



MDP with highest gain

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

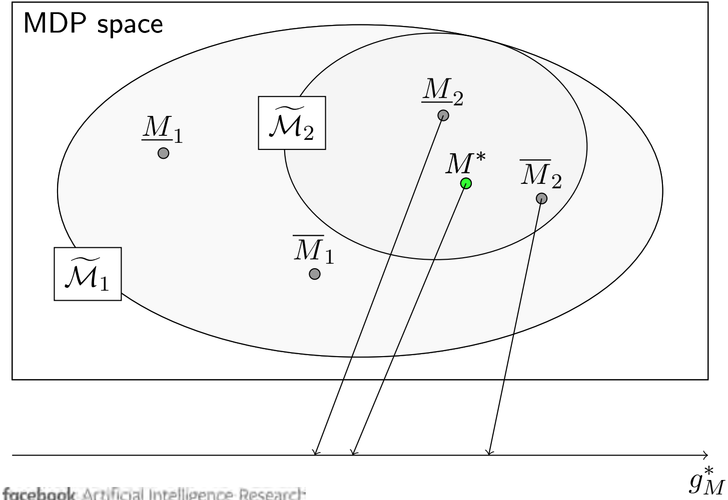
$$\pi_k \in \arg\max_{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})} \{ g_{M_k}^{\pi} \}$$

Optimal policy of M_k

Plausible MDPs: Optimistic Planning

- UCRL [Jaksch et al., 2010]

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



MDP with highest gain

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

$$\pi_k \in \arg\max_{\pi:\mathcal{S}\to\mathcal{P}(\mathcal{A})}\{g_{M_k}^{\pi}\}$$

Optimal policy of M_k

Plausible MDPs: Optimistic Planning

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

$$(M_k, \pi_k) \in \underset{M \in \mathcal{M}_k \;,\; \pi \in \Pi_C(M)}{\arg\max} \{g_M^\pi\}$$

$$\Pi_C(M) := \left\{ \pi : \mathcal{S} \to \mathcal{P}(\mathcal{A}) \;:\; sp\left\{h_M^\pi\right\} \leq c \right\}$$

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

❗ this is a *constrained* optimization problem

Plausible MDPs: Optimistic Planning

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

$$(M_k, \pi_k) \in \arg\max_{M \in \mathcal{M}_k \;,\; \pi \in \Pi_C(M)} \{g_M^\pi\}$$

$$\Pi_C(M) := \left\{ \pi : \mathcal{S} \to \mathcal{P}(\mathcal{A}) \;:\; sp\{h_M^\pi\} \leq c \right\}$$

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

! this is a *constrained* optimization problem

NOT trivial optimization

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

$$v_{n+1} = \widetilde{L} v_n := \max_{a \in \mathcal{A}} \left\{ \max_{r \in \beta_{r,k}(s,a)} r + \max_{p \in \beta_{p,k}(s,a)} p(\cdot|s,a)^\intercal v_n \right\} \tag{1}$$

   - 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
   $$v_{n+1} = \widetilde{L} v_n := \max_{a \in \mathcal{A}} \left\{ \max_{r \in \beta_{r,k}(s,a)} r + \max_{p \in \beta_{p,k}(s,a)} p(\cdot|s,a)^\intercal v_n \right\} \tag{1}$$
   - SCOPT for SCAL
3. Complicated to generalize outside finite MDPs

SOLUTION
exploration bonus

Exploration Bonus: the optimistic empirical MDP

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

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

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

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

$$\pi_k \in \arg\max_{\pi \in} \ g^{\pi}_{\widehat{M}_k}$$

Exploration Bonus: the optimistic empirical MDP

Optimistic Empirical MDP:

$$\widehat{M}_k = \{\mathcal{S}, \mathcal{A}, \overline{p}_k, \overline{r}_k + b_k\}$$

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

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

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

$$\pi_k \in \arg\max_{\pi \in} g^{\pi}_{\widehat{M}_k}$$

Exploration Bonus: the optimistic empirical MDP

$$\widehat{M}_k = \{\mathcal{S}, \mathcal{A}, \; \overline{p}_k \; , \; \overline{r}_k \; + b_k \; \}$$

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

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

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

$$\pi_k \in \operatorname*{arg\,max}_{\pi \in \Pi_c(\widehat{M}_k)} \; g^{\pi}_{\widehat{M}_k}$$

Exploration Bonus: the optimistic empirical MDP

$$\widehat{M}_k = \{\mathcal{S}, \mathcal{A}, \; \overline{p}_k \;, \; \overline{r}_k \; + b_k \; \}$$

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

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

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

$$\pi_k \in \operatorname*{arg\,max}_{\pi \in \boxed{\Pi_c(\widehat{M}_k)}} \; g^{\pi}_{\widehat{M}_k}$$

Still a Span-Constrained Optimization

$$\Pi_c(M) := \{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A}) : sp\{h^{\pi}_{M}\} \le c\}$$

Exploration bonus

$$\left|\, r(s,a) \;-\; \overline{r}_k(s,a) \,\right| \lesssim\; r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}$$

$$\left|\, (p(\cdot|s,a) - \overline{p}_k(\cdot|s,a))^\mathsf{T}\, h^* \right| \lesssim\; c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}$$

Bellman Operator of $\widehat{M}_k$

$$\widehat{L}h^* = \max_{a\in\mathcal{A}} \left\{ \overline{r}_k(s,a) + \overline{p}_k(\cdot|s,a)^\mathsf{T} h^* \right\} \tag{2}$$

$$= \max_{a\in\mathcal{A}} \left\{ \underbrace{\overline{r}_k(s,a) + r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq r(s,a)} + \underbrace{\overline{p}_k(\cdot|s,a)^\mathsf{T} h^* + c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq p(\cdot|s,a)^\mathsf{T} h^*} \right\}$$

$$\geq Lh^*$$

Exploration bonus

$$\left| r(s,a) - \overline{r}_k(s,a) \right| \lesssim \boxed{r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}$$

$$\left| (p(\cdot|s,a) - \overline{p}_k(\cdot|s,a))^\intercal h^* \right| \lesssim \boxed{c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}$$

Bellman Operator of $\widehat{M}_k$

$$\widehat{L}h^* = \max_{a\in\mathcal{A}} \left\{ \overline{r}_k(s,a) + b_k(s,a) + \overline{p}_k(\cdot|s,a)^\intercal h^* \right\} \tag{2}$$

$$= \max_{a\in\mathcal{A}} \left\{ \underbrace{\overline{r}_k(s,a) + r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq r(s,a)} + \underbrace{\overline{p}_k(\cdot|s,a)^\intercal h^* + c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq p(\cdot|s,a)^\intercal h^*} \right\}$$

$$\geq Lh^*$$

Exploration bonus

$$\left| r(s,a) - \overline{r}_k(s,a) \right| \lesssim \boxed{r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}$$

$$\left| (p(\cdot|s,a) - \overline{p}_k(\cdot|s,a))^\mathsf{T} h^* \right| \lesssim \boxed{c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}$$

Bellman Operator of $\widehat{M}_k$

$$\widehat{L}h^* = \max_{a\in\mathcal{A}} \left\{ \overline{r}_k(s,a) + b_k(s,a) + \overline{p}_k(\cdot|s,a)^\mathsf{T} h^* \right\} \tag{2}$$

$$= \max_{a\in\mathcal{A}} \left\{ \underbrace{\overline{r}_k(s,a) + r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq r(s,a)} + \underbrace{\overline{p}_k(\cdot|s,a)^\mathsf{T} h^* + c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq p(\cdot|s,a)^\mathsf{T} h^*} \right\}$$

$$\geq Lh^*$$

Exploration bonus

$$\left| r(s,a) - \overline{r}_k(s,a) \right| \lesssim \boxed{r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}$$

$$\left| (p(\cdot|s,a) - \overline{p}_k(\cdot|s,a))^\mathsf{T} h^* \right| \lesssim \boxed{c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}$$

Bellman Operator of $\widehat{M}_k$

$$\widehat{L}h^* = \max_{a\in\mathcal{A}} \{\overline{r}_k(s,a) + b_k(s,a) + \overline{p}_k(\cdot|s,a)^\mathsf{T} h^*\} \tag{2}$$

$$= \max_{a\in\mathcal{A}} \Bigg\{ \underbrace{\overline{r}_k(s,a) + r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq r(s,a)} + \underbrace{\overline{p}_k(\cdot|s,a)^\mathsf{T} h^* + c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq p(\cdot|s,a)^\mathsf{T} h^*} \Bigg\}$$

$$\geq Lh^*$$

Exploration bonus

$$\left| r(s,a) - \overline{r}_k(s,a) \right| \lesssim \boxed{r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}$$

$$\left| (p(\cdot|s,a) - \overline{p}_k(\cdot|s,a))^\mathsf{T} h^* \right| \lesssim \boxed{c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}$$

$\boxplus$

Bellman Operator of $\widehat{M}_k$

$$\widehat{L}h^* = \max_{a\in\mathcal{A}} \left\{ \overline{r}_k(s,a) + b_k(s,a) + \overline{p}_k(\cdot|s,a)^\mathsf{T} h^* \right\} \tag{2}$$

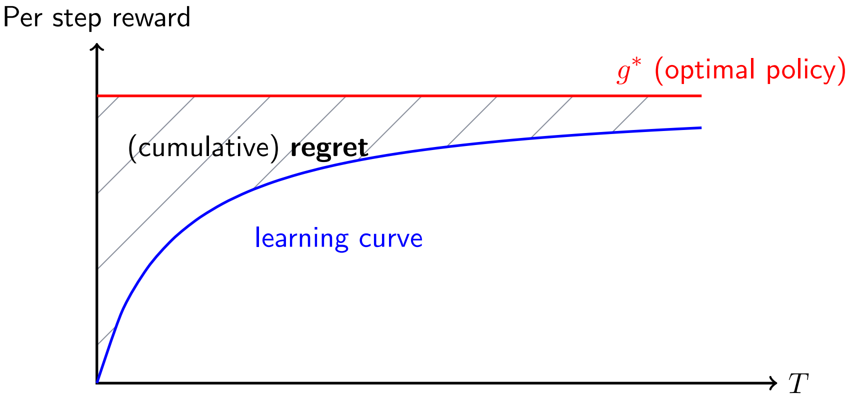
$$= \max_{a\in\mathcal{A}} \Bigg\{ \underbrace{\overline{r}_k(s,a) + r_{\max}\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq r(s,a)} + \underbrace{\overline{p}_k(\cdot|s,a)^\mathsf{T} h^* + c\sqrt{\frac{\ln(t_k/\delta)}{N_k(s,a)}}}_{\geq p(\cdot|s,a)^\mathsf{T} h^*} \Bigg\}$$

$$\geq Lh^*$$

[Puterman, 1994] [Fruit, **P.**, Lazaric, Ortner; 2018b] $\Longrightarrow$ $\boxed{g_k = g_c^*(\widehat{M}_k) \gtrsim g^*}$

Performance of a learning agent

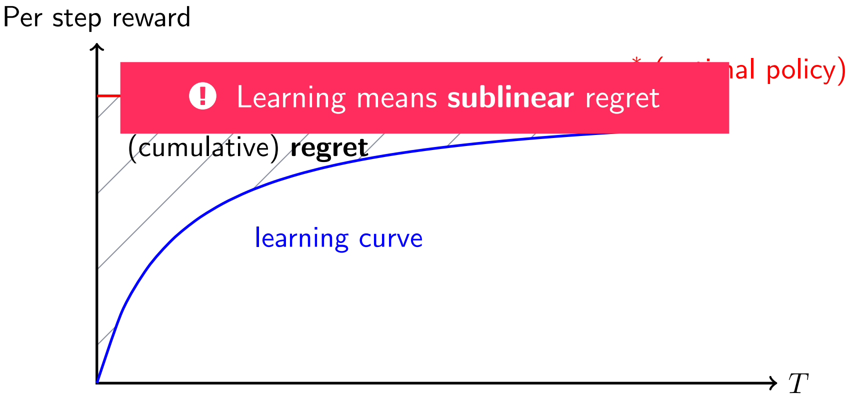
$$\text{Regret} \quad \Delta(\mathfrak{A}, T) = \sum_{t=1}^{T} \Big(g^* - r_t(s_t, a_t) \Big)$$



*different definition for finite-horizon problems

Performance of a learning agent

$$\text{Regret} \quad \Delta(\mathfrak{A}, T) = \sum_{t=1}^{T} \Big(g^* - r_t(s_t, a_t) \Big)$$



! Learning means **sublinear** regret

*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

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

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

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

D in UCRL

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

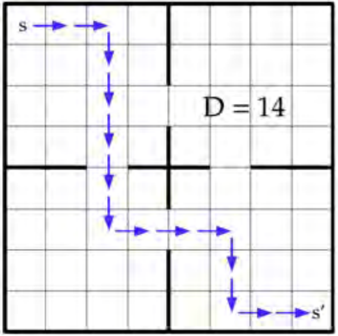
Regret of SCAL$^+$

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

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

D in UCRL

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



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

Mean arrival time in s' starting in s

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

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

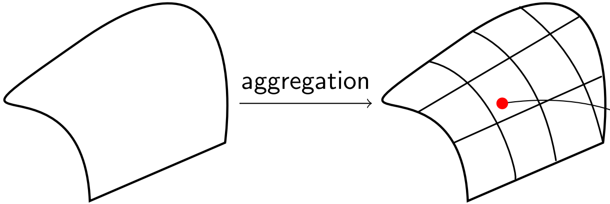
D in UCRL

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

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

Why Exploration Bonus?

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



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

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

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

$$\widetilde{r}(s,a) = r(s,a) + \sqrt{\frac{\beta}{N_k(\phi(s,a))}}$$

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

Span-Constrained Planning

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

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

- Connection with the exploration-exploitation framework
  - SCAL: $M := \widetilde{\mathcal{M}}_k$, an extended MDP with continuous actions $\widetilde{\mathcal{A}}_k$

    $$(M_k, \pi_k) \in \mathop{\arg\max}_{M \in \mathcal{M}_k,\ \pi \in \Pi_c(M)} g^\pi_M \quad \text{equivalent} \quad \widetilde{\pi}_k \in \mathop{\arg\max}_{\pi : \mathcal{S} \to \mathcal{P}(\widetilde{\mathcal{A}}_k) \wedge sp\{h^\pi\} \leq c} g^\pi_{\widetilde{\mathcal{M}}_k}$$

    i.e., where the Bellman operator $\widetilde{L}$ is defined in Eq. 1
  - SCAL$^+$: $M := \widehat{M}_k$ where $\widehat{L}$ is defined as in Eq. 2

Span-Constrained Planning

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

$$\Pi_c(M) := \{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A}) : sp\{h_M^\pi\} \leq c \wedge sp\{g_M^\pi\} = 0\}$$

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

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

$$v_{n+1} \overset{\forall s}{=} \begin{cases} c & \text{if } v_{n+1}(s) \geq \min\{v_{n+1}\} + c \\ v_{n+1}(s) & \text{otherwise} \end{cases}$$

Span-Constrained Planning

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

$$\Pi_c(M) := \{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A}) : sp\{h_M^\pi\} \leq c \wedge sp\{g_M^\pi\} = 0\}$$

- NOT trivial optimization problem
- but apparently simple solution: ScOpt [Fruit, P., Lazari

i.e., "truncated" above

$\forall n,\ sp\{v_n\} \leq c$

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

$$v_{n+1} \overset{\forall s}{=} \begin{cases} c & \text{if } v_{n+1}(s) \geq \min\{v_{n+1}\} + c \\ v_{n+1}(s) & \text{otherwise} \end{cases}$$

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. $\forall v: \; sp\{v\} \leq c, \quad \min_a \left\{ r(s,a) + p(\cdot|s,a)^\mathsf{T} v \right\} \leq \min_{s'}\{Lv(s')\} + c$

then

- *optimality equation*: $T_c h^+ = h^+ + g^+ e$ and $g^+ = g_c^*$
- *convergence*: $\lim_{n\to\infty} T_c^{n+1} v_0 - T_c^n v_0 = g^+ e$

How to force these properties in exp-exp

The estimated MDP

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

$$\widehat{p}_k(s'|s,a) = \frac{N_k(s,a)\overline{p}_k(s'|s,a)}{N_k(s,a)+1} + \frac{1(s'=\overline{s})}{N_k(s,a)+1}$$

$\implies \text{SCOPT}$ converges

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

- Augment the reward: duplicate all the actions

$$\forall s \in \mathcal{S}, a \in \mathcal{A}_t, \quad \text{define } b \text{ such that } p(\cdot|s,b) = p(\cdot|s,a) \text{ and } \boxed{r(s,b) = 0}$$

How to force these properties in exp-exp

The estimated MDP

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

$$\widehat{p}_k(s'|s,a) = \frac{N_k(s,a)\overline{p}_k(s'|s,a)}{N_k(s,a)+1} + \frac{1(s'=\overline{s})}{N_k(s,a)+1}$$

$\implies$ SCOPT converges

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

- Augment the reward: duplicate all the actions

$$\forall s \in \mathcal{S}, a \in \mathcal{A}_t, \quad \text{define } b \text{ such that } p(\cdot|s,b) = p(\cdot|s,a) \text{ and } \boxed{r(s,b) = 0}$$

How to force these properties in exp-exp

The estimated MDP

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

$$\widehat{p}_k(s'|s,a) = \frac{N_k(s,a)\overline{p}_k(s'|s,a)}{N_k(s,a)+1} + \frac{1(s'=\overline{s})}{N_k(s,a)+1}$$

  $\Longrightarrow$ SCOPT converges

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

- Augment the reward: duplicate all the actions

$$\forall s \in \mathcal{S}, a \in \mathcal{A}_t, \quad \text{define } b \text{ such that } p(\cdot|s,b) = p(\cdot|s,a) \text{ and } \boxed{r(s,b)=0}$$

How to force these properties in exp-exp

The estimated MDP

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

$$\widehat{p}_k(s'|s,a) = \frac{N_k(s,a)\overline{p}_k(s'|s,a)}{N_k(s,a)+1} + \frac{1(s'=\overline{s})}{N_k(s,a)+1}$$

$\Longrightarrow \text{SCOPT}$ converges

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

- Augment the reward: duplicate all the actions

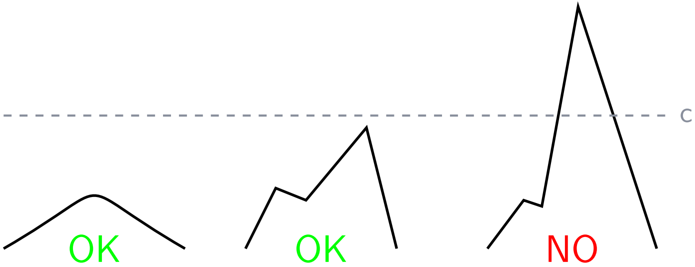
$$\forall s \in \mathcal{S}, a \in \mathcal{A}_t, \quad \text{define } b \text{ such that } p(\cdot|s,b) = p(\cdot|s,a) \text{ and } \boxed{r(s,b) = 0}$$

When $\widehat{M}_k$ 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

$$\left| \left(p(\cdot|s,a) - \overline{p}_k(\cdot|s,a)\right)^\intercal h^* \right| \leq \|p(\cdot|s,a) - \overline{p}_k(\cdot|s,a)\|_1 \|h^*\|_\infty$$

Intrinsic in other settings (infinite-horizon undiscounted, finite-horizon)

The role of prior knowledge

Intrinsic Horizon

Setting	MDP parameter	Horizon	Knowledge	Exploration Bonus
infinite-horizon discounted	γ	$\frac{1}{1-\gamma}$	$\lvert Q(s,a)\rvert \leq \frac{r_{\max}}{1-\gamma}$	$\widetilde{\Theta}\left(\frac{r_{\max}}{1-\gamma}\sqrt{\frac{1}{N_k(s,a)}}\right)$ MBIE-EB [Strehl and Littman, 2008]
finite-horizon	H	H	$\lvert Q(s,a)\rvert \leq r_{\max}H$	$\widetilde{\Theta}\left(r_{\max}H\sqrt{\frac{1}{N_k(s,a)}}\right)$ UCBVI-1 [Azar et al., 2017]
others [Azar et al., 2017, Kakade et al., 2018, Jin et al., 2018]				
average reward	?	$+\infty$	?	?

The role of prior knowledge

Intrinsic Horizon

Setting	MDP parameter	Horizon	Knowledge	Exploration Bonus
infinite-horizon discounted	γ	$\frac{1}{1-\gamma}$	$\|Q(s,a)\| \leq \frac{r_{\max}}{1-\gamma}$	$\widetilde{\Theta}\left(\frac{r_{\max}}{1-\gamma}\sqrt{\frac{1}{N_k(s,a)}}\right)$ MBIE-EB [Strehl and Littman, 2008]
finite-horizon	H	H	$\|Q(s,a)\| \leq r_{\max} H$	$\widetilde{\Theta}\left(r_{\max} H\sqrt{\frac{1}{N_k(s,a)}}\right)$ UCBVI-1 [Azar et al., 2017]
others [Azar et al., 2017, Kakade et al., 2018, Jin et al., 2018]				
average reward	?	$+\infty$	$sp\{h^*\} \leq c$ *assumption*	$\widetilde{\Theta}\left(c\sqrt{\frac{1}{N_k(s,a)}}\right)$ SCAL$^+$ [Qian, Fruit, **P.**, Lazaric, 2018]

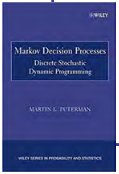
The role of prior knowledge

in Average Reward settings

- Almost all the algorithms requires prior knowledge

+ complexity/generality

MDP	Algorithm	Properties/Assumptions
Ergodic	KL-UCRL [Talebi and Maillard, 2018]	
Communicating	UCRL [Jaksch et al., 2010]	$D < +\infty$
Weakly Comm.	REGAL [Bartlett and Tewari, 2009] SCAL [Fruit, **P.**, Lazaric, Ortner, 2018b] SCAL$^+$ [Qian, Fruit, **P.**, Lazaric, 2018a]	$D = +\infty$ but we need $sp\{h^*\} \leq c$
Non Comm.	TUCRL [Fruit, **P.**, Lazaric, 2018a]	No assumptions but impossible to have logarithmic regret



[Puterman, 1994] Sec. 8.3

Outlook

span-constrained exp-exp $\Longleftrightarrow$ 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

Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *UAI*, pages 1–11. AUAI Press, 2015.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *NIPS*, pages 1184–1194, 2017.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *ALT*, pages 150–165, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

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

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *UAI*, pages 35–42. AUAI Press, 2009.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *NIPS*, pages 1471–1479, 2016.

Dimitri P Bertsekas. *Dynamic programming and optimal control. Vol II*. Number 2. Athena scientific Belmont, MA, 1995.

Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231, March 2003. ISSN 1532-4435.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Emma Brunskill. Regret minimization in mdps with options without prior knowledge. In *NIPS*, pages 3169–3179, 2017.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating markov decision processes. In *NIPS*, 2018a.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *ICML*, Proceedings of Machine Learning Research. PMLR, 2018b.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 861–898. JMLR.org, 2015.

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

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? *CoRR*, abs/1807.03765, 2018.

Sham Kakade, Mengdi Wang, and Lin F. Yang. Variance reduction methods for sublinear reinforcement learning. *CoRR*, abs/1802.09184, 2018.

A. Klenke and M. Loève. *Probability Theory: A Comprehensive Course*. Graduate texts in mathematics. Springer, 2013. ISBN 9781447153627.

K. Lakshmanan, Ronald Ortner, and Daniil Ryabko. Improved regret bounds for undiscounted continuous reinforcement learning. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 524–532. JMLR.org, 2015.

Odalric-Ambrym Maillard, Phuong Nguyen, Ronald Ortner, and Daniil Ryabko. Optimal regret bounds for selecting the state representation in reinforcement learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 543–551, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. Count-based exploration in feature space for reinforcement learning. *CoRR*, abs/1706.08090, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

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 *NIPS*, pages 1772–1780, 2012.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *NIPS*, pages 3003–3011, 2013.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 2721–2730. PMLR, 2017.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *NIPS*, pages 1333–1342, 2017.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *ALT*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR, 2018.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In *NIPS*, pages 2750–2759, 2017.