Multiple-Step Greedy Policies in Online and Approximate Reinforcement Learning
Neural Information Processing Systems, December ’18

Yonathan Efroni\textsuperscript{1}  Gal Dalal\textsuperscript{1}  Bruno Scherrer\textsuperscript{2}  Shie Mannor\textsuperscript{1}

\textsuperscript{1} Department of Electrical Engineering, Technion, Israel
\textsuperscript{2}INRIA, Villers les Nancy, France
Motivation: Impressive Empirical Success

Multiple-step lookahead policies in RL give state-of-the-art-performance.
Motivation: Impressive Empirical Success

Multiple-step lookahead policies in RL give state-of-the-art-performance.

- **Model Predictive Control (MPC) in RL**
  - Negenborn et al. (2005); Ernst et al. (2009); Zhang et al. (2016);
  - Tamar et al. (2017); Nagabandi et al. (2018), and many more...
Motivation: Impressive Empirical Success

Multiple-step lookahead policies in RL give state-of-the-art performance.

- **Model Predictive Control (MPC) in RL**
  Negenborn et al. (2005); Ernst et al. (2009); Zhang et al. (2016); Tamar et al. (2017); Nagabandi et al. (2018), and many more...

- **Monte Carlo Tree Search (MCTS) in RL**
  Tesauro and Galperin (1997); Baxter et al. (1999); Sheppard (2002); Veness et al. (2009); Lai (2015); Silver et al. (2017); Amos et al. (2018), and many more...
Motivation: Although the Impressive Empirical Success...
Theory on how to combine multiple-step lookahead policies in RL is scarce.
Motivation: Although the Impressive Empirical Success...

Theory on how to combine multiple-step lookahead policies in RL is scarce.

Bertsekas and Tsitsiklis (1995); Efroni et al. (2018):
Multiple-step greedy policies at the improvement stage of Policy Iteration.
Motivation: Although the Impressive Empirical Success...

Theory on how to combine multiple-step lookahead policies in RL is scarce.

Bertsekas and Tsitsiklis (1995); Efroni et al. (2018):
Multiple-step greedy policies at the improvement stage of Policy Iteration.

Here: Extend to online and approximate RL.
Multiple-Step Greedy Policies: $h$-Greedy Policy

$h$-Greedy Policy w.r.t. $\nu^\pi$:
Multiple-Step Greedy Policies: $h$- Greedy Policy

$h$-Greedy Policy w.r.t. $v^\pi$:

Optimal first action in $h$-horizon $\gamma$-discounted Markov Decision Process, total reward $\sum_{t=0}^{h-1} \gamma^t r(s_t, \pi_t(s_t)) + \gamma^h v^\pi(s_h)$. 
**Multiple-Step Greedy Policies: \( h \)-Greedy Policy**

\( h \)-Greedy Policy w.r.t. \( \pi \):

Optimal *first* action in \( h \)-horizon \( \gamma \)-discounted Markov Decision Process, total reward 
\[
\sum_{t=0}^{h-1} \gamma^t r(s_t, \pi_t(s_t)) + \gamma^h v^\pi(s_h).
\]

\[ s_0 \]
\[
\gamma r(s_1, \pi_1(s_1)) \quad \gamma^2 v^\pi(s_2)
\]

\( h = 2 \)-Greedy Policy as a Tree Search
**Multiple-Step Greedy Policies: $h$-Greedy Policy**

$h$-Greedy Policy w.r.t. $v^\pi$:

Optimal *first* action in $h$-horizon $\gamma$-discounted Markov Decision Process, total reward $\sum_{t=0}^{h-1} \gamma^t r(s_t, \pi_t(s_t)) + \gamma^h v^\pi(s_h)$.

Path with max. total reward

$h = 2$-Greedy Policy as a Tree Search
Multiple-Step Greedy Policies: $h$- Greedy Policy

$h$-Greedy Policy w.r.t. $v^\pi$:

Optimal first action in $h$-horizon $\gamma$-discounted Markov Decision Process, total reward $\sum_{t=0}^{h-1} \gamma^t r(s_t, \pi_t(s_t)) + \gamma^h v^\pi(s_h)$.

Path with max. total reward

$h$ = 2-Greedy Policy as a Tree Search
Multiple-Step Greedy Policies: $\kappa$-Greedy Policy

$\kappa$-Greedy Policy w.r.t $v^\pi$:

Optimal action when

$P_r(\text{Solve the } h\text{-horizon MDP}) = (1 - \kappa)\kappa^{h-1}$. 
Multiple-Step Greedy Policies: $\kappa$-Greedy Policy

$\kappa$-Greedy Policy w.r.t $\nu^\pi$:

Optimal action when

$Pr(\text{Solve the } h\text{-horizon MDP}) = (1 - \kappa) \kappa^{h-1}$.

\[
\begin{align*}
Pr(h = 1) &= (1 - \kappa) \\
Pr(h = 2) &= (1 - \kappa) \kappa \\
Pr(h = 3) &= (1 - \kappa) \kappa^2
\end{align*}
\]
1-Step Greedy Policies and Soft Updates

Soft update using a 1-step greedy policy *improves* policy.
1-Step Greedy Policies and Soft Updates

Soft update using a 1-step greedy policy *improves* policy.

A bit formally,

- Let $\pi$ be a policy,
1-Step Greedy Policies and Soft Updates

Soft update using a 1-step greedy policy *improves* policy.

A bit formally,

- Let $\pi$ be a policy,
- $\pi_{G_1}$ 1-step greedy policy w.r.t. $v^\pi$. 
1-Step Greedy Policies and Soft Updates

Soft update using a 1-step greedy policy *improves* policy.

A bit formally,

- Let $\pi$ be a policy,
- $\pi_{G_1}$ 1-step greedy policy w.r.t. $v^\pi$.

Then, $\forall \alpha \in [0, 1]$, $(1 - \alpha)\pi + \alpha\pi_{G_1}$, is always better than $\pi$. 
1-Step Greedy Policies and Soft Updates

Soft update using a 1-step greedy policy improves policy.

A bit formally,

- Let $\pi$ be a policy,
- $\pi_{G_1}$ 1-step greedy policy w.r.t. $v^\pi$.

Then, $\forall \alpha \in [0, 1]$, $(1 - \alpha)\pi + \alpha\pi_{G_1}$, is always better than $\pi$.

**Important fact in:**

Two-timescale online PI (Konda and Borkar (1999)), Conservative PI (Kakade and Langford (2002)), TRPO (Schulman et al. (2015)), and many more...
Negative Result on Multiple-Step Greedy Policies

Soft update using a multiple-step greedy policy does not necessarily improves policy.
Negative Result on Multiple-Step Greedy Policies

Soft update using a multiple-step-greedy-policy does not necessarily improves policy.

Necessary and sufficient condition: $\alpha$ is large enough.
Negative Result on Multiple-Step Greedy Policies

Soft update using a multiple-step-greedy-policy does not necessarily improves policy.

Necessary and sufficient condition: $\alpha$ is large enough.

**Theorem 1**

Let $\pi_{G_h}$ and $\pi_{G_\kappa}$ be the $h$-greedy and $\kappa$-greedy policies w.r.t. $v^\pi$. Then.
Negative Result on Multiple-Step Greedy Policies

Soft update using a multiple-step-greedy-policy does not necessarily improves policy.

Necessary and sufficient condition: $\alpha$ is large enough.

**Theorem 1**

Let $\pi_{G_h}$ and $\pi_{G_\kappa}$ be the $h$-greedy and $\kappa$-greedy policies w.r.t. $v^\pi$. Then.

- $(1 - \alpha)\pi + \alpha\pi_{G_h}$ is always better than $\pi$ for $h > 1$ iff $\alpha = 1$. 


Negative Result on Multiple-Step Greedy Policies

Soft update using a multiple-step-greedy-policy does not necessarily improves policy.

Necessary and sufficient condition: \( \alpha \) is large enough.

**Theorem 1**

Let \( \pi_{G_h} \) and \( \pi_{G_\kappa} \) be the \( h \)-greedy and \( \kappa \)-greedy policies w.r.t. \( v^\pi \). Then.

- \( (1 - \alpha)\pi + \alpha \pi_{G_h} \) is always better than \( \pi \) for \( h > 1 \) iff \( \alpha = 1 \).
- \( (1 - \alpha)\pi + \alpha \pi_{G_\kappa} \) is always better than \( \pi \) iff \( \alpha \geq \kappa \).
How to Circumvent the Problem? (and have Theoretical Guarantees)
How to Circumvent the Problem? (and have Theoretical Guarantees)

Give ‘natural’ solutions to the problem with theoretical guarantees:
How to Circumvent the Problem? (and have Theoretical Guarantees)

Give ‘natural’ solutions to the problem with theoretical guarantees:

- Two-timescale, online, multiple-step PI.
How to Circumvent the Problem? (and have Theoretical Guarantees)

Give ‘natural’ solutions to the problem with theoretical guarantees:

▶ Two-timescale, online, multiple-step PI.
▶ Approximate multiple-step PI methods.
How to Circumvent the Problem? (and have Theoretical Guarantees)

Give ‘natural’ solutions to the problem with theoretical guarantees:

▶ Two-timescale, online, multiple-step PI.
▶ Approximate multiple-step PI methods.

Open Problem:

More techniques to circumvent the problem.
Take Home Messages

▶ Important difference between multiple- and 1-step greedy methods.
Take Home Messages

▶ Important difference between multiple- and 1-step greedy methods.

▶ Multiple-step PI has *theoretical* benefits (more discussion at the poster session).
Take Home Messages

▶ Important difference between multiple- and 1-step greedy methods.

▶ Multiple-step PI has *theoretical* benefits (more discussion at the poster session).

▶ Further study should be devoted.


