Multiple-Step Greedy Policies in Online and Approximate Reinforcement Learning

Neural Information Processing Systems, December '18

Yonathan Efroni[1] Gal Dalal[1] Bruno Scherrer[2] Shie Mannor[1]

[1] Department of Electrical Engineering, Technion, Israel

[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.

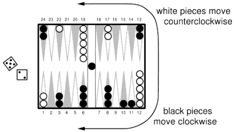
- **Model Predictive Control (MPC) in RL**
  Negenborn et al. (2005); Ernst et al. (2009); Zhang et al. (2016); Tamar et al. (2017); Nagabandi et al. (2018), and many more...

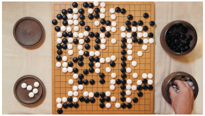
Motivation: Impressive Empirical Success

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...

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. v^π:

Multiple-Step Greedy Policies: h- Greedy Policy

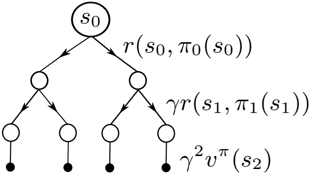
h-Greedy Policy w.r.t. v^π:

Optimal *first* action in h-horizon γ-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. v^π:

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



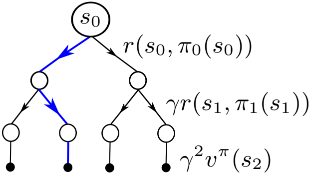
$h = 2$-Greedy Policy as a Tree Search

Multiple-Step Greedy Policies: h- Greedy Policy

h-Greedy Policy w.r.t. v^π:

Optimal *first* action in h-horizon γ-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

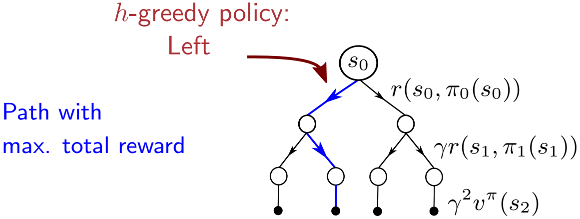


$h = 2$-Greedy Policy as a Tree Search

Multiple-Step Greedy Policies: h- Greedy Policy

h-Greedy Policy w.r.t. v^π:

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



$h = 2$-Greedy Policy as a Tree Search

Multiple-Step Greedy Policies: κ- Greedy Policy

κ-Greedy Policy w.r.t v^π:

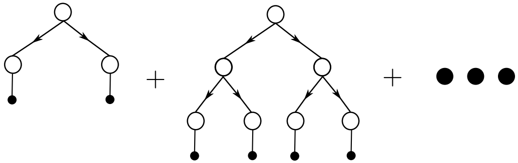
Optimal action when
P_r(Solve the h-horizon MDP) $= (1-\kappa)\kappa^{h-1}$.

Multiple-Step Greedy Policies: κ- Greedy Policy

κ-Greedy Policy w.r.t v^π:

Optimal action when
$P_r(\text{Solve the } h\text{-horizon MDP}) = (1-\kappa)\kappa^{h-1}$.

$P_r(h=1) = (1-\kappa)$ $\quad$ $P_r(h=2) = (1-\kappa)\kappa$ $\quad$ $P_r(h=3) = (1-\kappa)\kappa^2$



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 π be a policy,

1-Step Greedy Policies and Soft Updates

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

A bit formally,

- Let π be a policy,
- $\pi_{\mathcal{G}_1}$ 1-step greedy policy w.r.t. v^π.

1-Step Greedy Policies and Soft Updates

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

A bit formally,

- Let π be a policy,
- $\pi_{\mathcal{G}_1}$ 1-step greedy policy w.r.t. v^π.

Then, $\forall \alpha \in [0,1]$, $(1-\alpha)\pi + \alpha \pi_{\mathcal{G}_1}$, *is always better than* π.

1-Step Greedy Policies and Soft Updates

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

A bit formally,

- Let π be a policy,
- $\pi_{\mathcal{G}_1}$ 1-step greedy policy w.r.t. v^π.

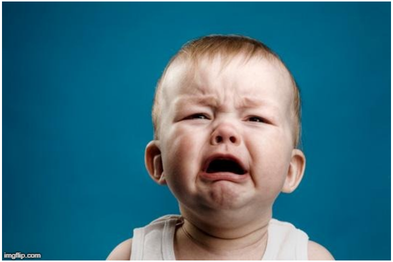
Then, $\forall \alpha \in [0,1]$, $(1-\alpha)\pi + \alpha\pi_{\mathcal{G}_1}$, *is always better than* π.

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: α 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: α is large enough.

Theorem 1

Let $\pi_{\mathcal{G}_h}$ and $\pi_{\mathcal{G}_\kappa}$ be the h-greedy and κ-greedy policies w.r.t. v^π. 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: α is large enough.

Theorem 1

Let $\pi_{\mathcal{G}_h}$ and $\pi_{\mathcal{G}_\kappa}$ be the h-greedy and κ-greedy policies w.r.t. v^π. Then.

- $(1-\alpha)\pi + \alpha\pi_{\mathcal{G}_h}$ *is always better than* π *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: α is large enough.

Theorem 1

Let $\pi_{\mathcal{G}_h}$ and $\pi_{\mathcal{G}_\kappa}$ be the h-greedy and κ-greedy policies w.r.t. v^π. Then.

- $(1-\alpha)\pi + \alpha\pi_{\mathcal{G}_h}$ *is always better than* π *for* $h > 1$ *iff* $\alpha = 1$.
- $(1-\alpha)\pi + \alpha\pi_{\mathcal{G}_\kappa}$ *is always better than* π *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.

Amos, B., Dario Jimenez Rodriguez, I., Sacks, J., Boots J., B., and Kolter, Z. (2018). Differentiable mpc for end-to-end planning and control. *Advances in Neural Information Processing Systems*.

Baxter, J., Tridgell, A., and Weaver, L. (1999). Tdleaf (lambda): Combining temporal difference learning with game-tree search. *arXiv preprint cs/9901001*.

Bertsekas, D. P. and Tsitsiklis, J. N. (1995). Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1. IEEE.

Efroni, Y., Dalal, G., Scherrer, B., and Mannor, S. (2018). Beyond the one-step greedy approach in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1386–1395.

Ernst, D., Glavic, M., Capitanescu, F., and Wehenkel, L. (2009). Reinforcement learning versus model predictive control: a comparison on a power system problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):517–529.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pages 267–274.

Konda, V. R. and Borkar, V. S. (1999). Actor-critic–type learning algorithms for markov decision processes. *SIAM Journal on control and Optimization*, 38(1):94–123.

Lai, M. (2015). Giraffe: Using deep reinforcement learning to play chess. *arXiv preprint arXiv:1509.01549*.

Nagabandi, A., Kahn, G., Fearing, R. S., and Levine, S. (2018). Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7559–7566. IEEE.

Negenborn, R. R., De Schutter, B., Wiering, M. A., and Hellendoorn, H. (2005). Learning-based model predictive control for markov decision processes. *IFAC Proceedings Volumes*, 38(1):354–359.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897.

Sheppard, B. (2002). World-championship-caliber scrabble. *Artificial Intelligence*, 134(1-2):241–275.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354.

Tamar, A., Thomas, G., Zhang, T., Levine, S., and Abbeel, P. (2017). Learning from the hindsight planepisodic mpc improvement. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 336–343. IEEE.

Tesauro, G. and Galperin, G. R. (1997). On-line policy improvement using monte-carlo search. In *Advances in Neural Information Processing Systems*, pages 1068–1074.

Veness, J., Silver, D., Blair, A., and Uther, W. (2009). Bootstrapping from game tree search. In *Advances in neural information processing systems*, pages 1937–1945.

Zhang, T., Kahn, G., Levine, S., and Abbeel, P. (2016). Learning deep control policies for autonomous aerial vehicles with mpc-guided policy search. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 528–535. IEEE.