Tight Regret Bounds for Model-Based Reinforcement Learning with Greedy Policies

Yonathan Efroni*[1] Nadav Merlis*[1]

Mohammad Ghavamzadeh[2] Shie Mannor[1]

[1]Technion

[2]Facebook AI Research

NeurIPS, December 2019

*Equal Contribution

TL;DR

Same minimax bounds to model-based RL with short-term and long-term planning

TL;DR

Same minimax bounds to model-based RL with short-term and long-term planning

- Factor of S less computations, same performance

Model Based or Model Free Reinforcement Learning?

Provably efficient RL in Finite-Horizon MDPs:

- *Model-Based RL*: Minimax regret $O(\sqrt{HSAT})$.
- *Model-Free RL*: Q-learning regret $O(\sqrt{H^3SAT})$.

  - S, A state and action space cardinality
  - H horizon of the MDP
  - T total number of samples

Model Based or Model Free Reinforcement Learning?

Provably efficient RL in Finite-Horizon MDPs:

- *Model-Based RL*: Minimax regret $O(\sqrt{HSAT})$.
- *Model-Free RL*: Q-learning regret $O(\sqrt{H^3SAT})$.

- S, A state and action space cardinality
- H horizon of the MDP
- T total number of samples

Model-Based RL has better sample complexity

Motivation: Why not using Model-Based RL?

Model-Based RL:

- Space Complexity $O(S^2A)$,
- Computational Complexity per-episode $O(S^2AH)$,

Motivation: Why not using Model-Based RL?

Model-Based RL:

- Space Complexity $O(S^2A)$,
- Computational Complexity per-episode $O(S^2AH)$,

Model-Free RL:

- Space Complexity $O(SAH)$,
- Computational Complexity per-episode $O(AH)$,

Motivation: Why the High Computational Complexity?

Motivation: Why the High Computational Complexity?

Long term planning, solve an MDP, in each episode.

Algorithm 2: Generic Model-Based RL

for episode $k = 1, 2, \ldots$ **do**
  $\pi_k \leftarrow$ Optimal policy of an optimistic / sampled MDP
  Act and gather experience by π_k
end for

Motivation: Why the High Computational Complexity?

Long term planning, solve an MDP, in each episode.

Algorithm 3: Generic Model-Based RL

for episode $k = 1, 2, \ldots$ **do**
  $\pi_k \leftarrow$ Optimal policy of an optimistic / sampled MDP
  Act and gather experience by π_k
end for

e.g., Azar et al. [2017], Brafman and Tennenholtz [2002], Dann et al. [2018], Jaksch et al. [2010], Kearns and Singh [2002], Osband et al. [2013], Russo [2019], Simchowitz and Jamieson [2019], Zanette and Brunskill [2019] and more...

Motivation: Why the High Computational Complexity?

Long term planning, solve an MDP, in each episode.

Algorithm 4: Generic Model-Based RL

for episode $k = 1, 2, \ldots$ **do**
  $\pi_k \leftarrow$ Optimal policy of an optimistic / sampled MDP
  Act and gather experience by π_k
end for

e.g., Azar et al. [2017], Brafman and Tennenholtz [2002], Dann et al. [2018], Jaksch et al. [2010], Kearns and Singh [2002], Osband et al. [2013], Russo [2019], Simchowitz and Jamieson [2019], Zanette and Brunskill [2019] and more...

In practice, only short-term planning is used.
Does it perform worse?

This Work: Model-Based RL with Short-Term Planning

Model-Based RL with **short-term planning** is minimax optimal for finite-horizon MDPs.

This Work: Model-Based RL with Short-Term Planning

Model-Based RL with **short-term planning**
is minimax optimal for finite-horizon MDPs.

Algorithm 6: Generic (Optimistic) Model-Based RL with Greedy Policies

---

```
for episode k = 1, 2, ... do
    for time step t = 1, ..., H do
        a_t^k ← Greedy policy current state s_t^k w.r.t. an optimistic value V_k
        Act with a_t^k and observe r_t, s_{t+1}^k
    end for
    Update model with gathered experience
end for
```

This Work: Model-Based RL with Short-Term Planning

Model-Based RL with **short-term planning**
is minimax optimal for finite-horizon MDPs.

Algorithm 7: Generic (Optimistic) Model-Based RL with Greedy Policies

for episode $k = 1, 2, \dots$ **do**
  for time step $t = 1, \dots, H$ **do**
    $a_t^k \leftarrow$ **Greedy policy** current state s_t^k w.r.t. an optimistic value V_k
    Act with a_t^k and observe $r_t,\ s_{t+1}^k$
  end for
  Update model with gathered experience
end for

- **Greedy policy** from s_t w.r.t. V is a **1-step planning** operation:

$$a \in \arg\max_a \mathbb{E}[r(s_t, a) + V(s_{t+1}) \mid s_t].$$

This Work: Model-Based RL with Short-Term Planning

Model-Based RL with **short-term planning**
is minimax optimal for finite-horizon MDPs.

- **Free lunch**: Factor of S less computations, same performance.

This Work: Model-Based RL with Short-Term Planning

Model-Based RL with **short-term planning**
is minimax optimal for finite-horizon MDPs.

- **Free lunch**: Factor of S less computations, same performance.
- **Open Question:** *Why using* **lookahead policies** *in RL if 1-step planning is minimax optimal?*

Meet us at the poster session!

Poster's at Hall B + C #191

Thank you!

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org.

Brafman, R. I. and Tennenholtz, M. (2002). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231.

Dann, C., Li, L., Wei, W., and Brunskill, E. (2018). Policy certificates: Towards accountable reinforcement learning. *arXiv preprint arXiv:1811.03056*.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600.

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232.

Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011.

Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. *arXiv preprint arXiv:1906.02870*.

Simchowitz, M. and Jamieson, K. (2019). Non-asymptotic gap-dependent regret bounds for tabular mdps. *arXiv preprint arXiv:1905.03814*.

Zanette, A. and Brunskill, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*.