Off-Policy Learning Combined with Automatic Feature Expansion for Solving Large MDPs

Alborz Geramifard, Christoph Dann, Jonathan P. How

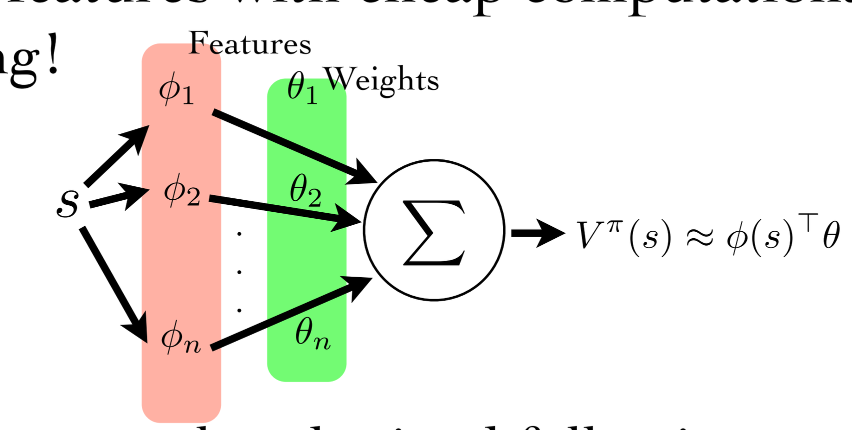
Abstract

Reinforcement learning (RL) techniques with **cheap computational complexity** and **minimal hand-tuning** that **scale to large problems** are highly desired among RL practitioners. Linear function approximation has scaled existing RL techniques to large problems, however this technique has two major drawbacks: 1) conventional **off-policy techniques** can be unstable with linear function approximation and 2) finding the **right set of features** for approximation can be challenging. This paper connects **Greedy-GQ** learning, a convergent off-policy technique with **iFDD⁺** algorithm, a novel feature expansion technique with cheap computational complexity. Empirical results across 3 domains with sizes up to ***77 billion* state-action pairs** verify the scalability of our new approach.

Problem

1 Real-world sequential decision making problems have **large** state spaces ➞ **Linear Function Approximation**

Finding the **right** set of features with cheap computational complexity is challenging!



2 Learning should consider samples obtained following arbitrary policies ➞ **Off-policy Learning**

How to incorporate off-policy learning without the fear of **divergence**?

Literature Review

1

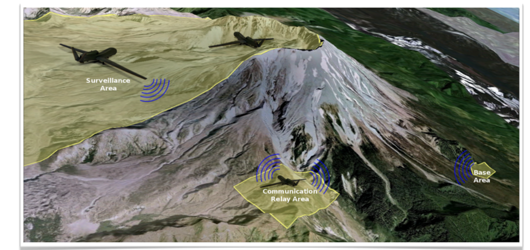
- **Hand Coding** [1]:
  Domain Specific, Time Consuming
- **Online methods** such as iFDD[2]:
  Shown successful with online learning,
  Sample complexity can be improved
- **Batch Techniques** such as BEBF[3] and OMP-TD[4], Batch-iFDD⁺[5]:
  Scalability, Tuning, Computational Complexity

2

- **Greedy-GQ** [6]:
  Convergent off-policy learning with linear function approximation,
  Fixed features

Contributions

- Introduced Greedy-GQ-iFDD⁺ as a new technique that is **off-policy**, **scalable** and **sample efficient**.
- Empirically tested the advantage of Greedy-GQ-iFDD⁺ against online learning techniques in 3 domains with sizes over ***77 Billion* state-action pairs**.



Greedy-GQ[6]

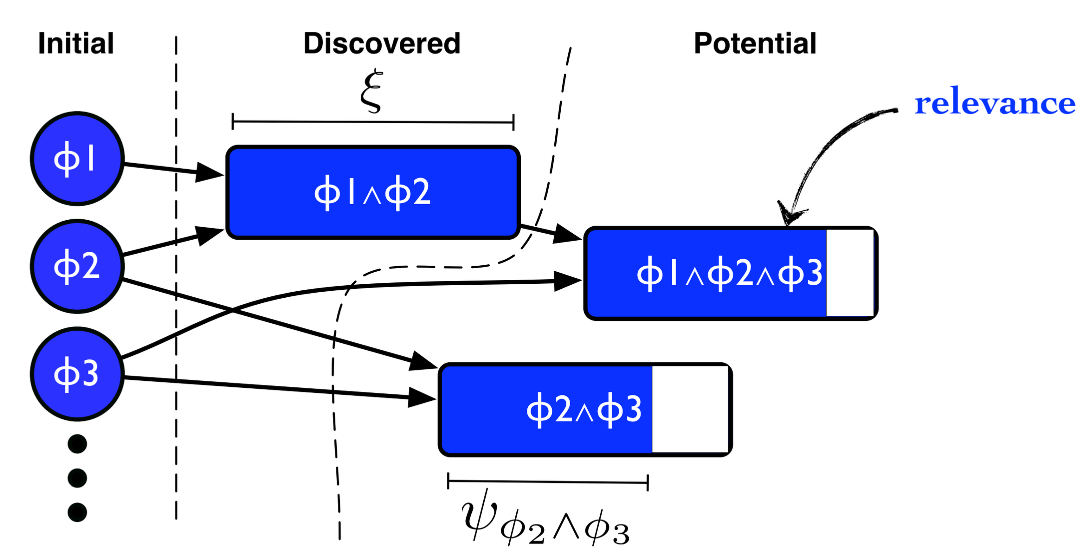
$$s_t \xrightarrow{a_t, r_t} s_{t+1}$$

$$\begin{aligned}
a' &= \mathrm{argmax}_a Q_\theta(s_{t+1},a)\\
\delta_t &= r_t + \gamma Q_{\theta_t}(s_{t+1},a') - Q_{\theta_t}(s_t,a_t)\\
\theta_{t+1} &= \theta_t + \alpha_t\left[\delta_t\phi(s_t,a_t) - \gamma(\omega_t^\top\phi(s_t,a_t))\phi(s_{t+1},a')\right]\\
\omega_{t+1} &= \omega_t + \beta_t\left[\delta_t - (\omega_t^\top\phi(s_t,a_t))\right]\phi(s_t,a_t)
\end{aligned}$$

Greedy-GQ is **convergent** using linear function approximation[6].

iFDD⁺[2,5]

Incremental Feature Dependency Discovery (iFDD)[2]



Given an initial set of binary features, add feature conjunctions where the TD-error **persists** (*i.e.* the relevance grows). Equation 2 (iFDD) was previously used in an online setting to add new features[2]. Equation 1 (iFDD⁺) was then introduced with a better convergence rate and used in a batch setting[5]. In this work we use Equation 1 in the online setting.

$$\text{(1)}\quad \mathrm{relevance}(f) = \frac{\left|\sum_{i\in\{1,\cdots,m\},\phi_f(s_i)=1}\delta_i\right|}{\sqrt{\sum_{i\in\{1,\cdots,m\},\phi_f(s_i)=1} 1}} \qquad \text{iFDD}^{+}\text{[5]}$$

$$\text{(2)}\quad \mathrm{relevance}(f) = \sum_{i\in\{1,\ldots,m\},\phi_f(s_i)=1} |\delta_i| \qquad \text{iFDD[2]}$$

Algorithm

Algorithm 1: Greedy GQ-iFDD⁺

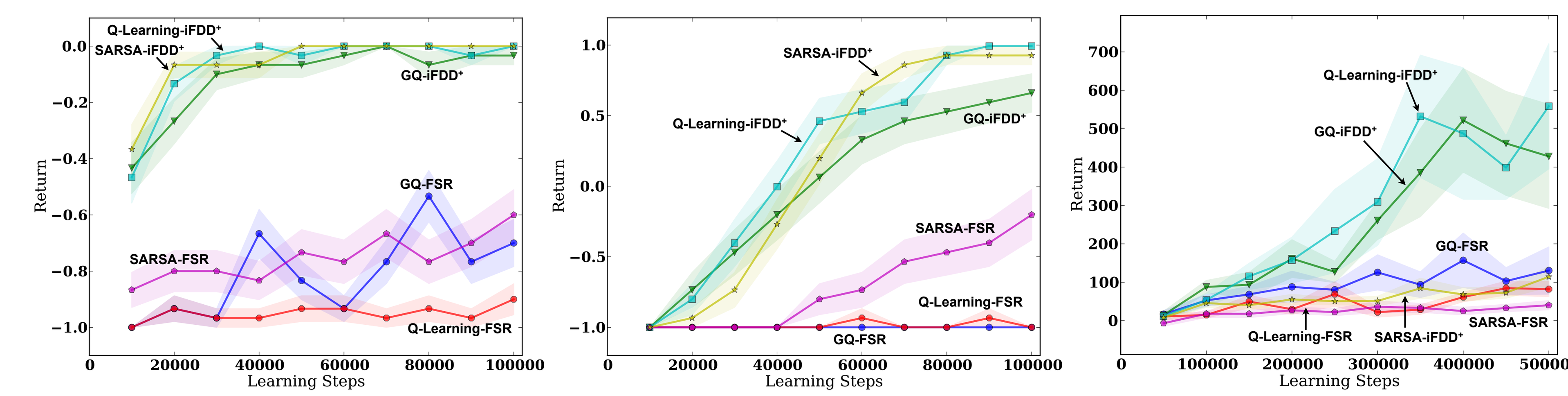
Input: $\alpha_t,\beta_t,\epsilon,\xi,\boldsymbol{F}$
Output: Q_θ

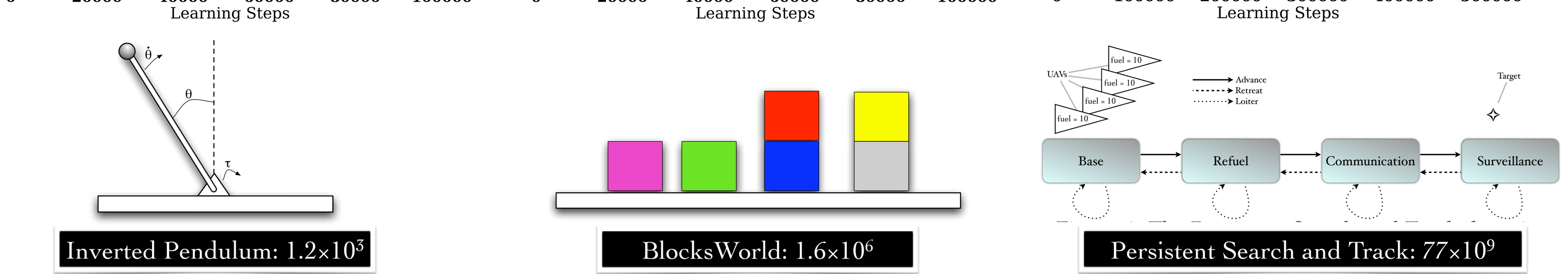
```
Initialize ψ,N to empty maps and ω and θ to 0̄
while time permits do
    initialize s from S_0
    repeat
        a ← ε-greedy w.r.t Q_θ
        s′,r ← execute a
        a′ ← argmax_a′ Q_θ(s′,a′)
        δ ← r + Q_θ(s′,a′) − Q_θ(s,a)
        θ ← θ + α_t [δφ(s,a) − γ(ω^⊤φ(s,a)) φ(s′,a′)]
        ω ← ω + β_t [δ − φ(s,a)^⊤ω] φ(s,a)
        φ ← Discover(φ(s),δ,ξ,F,ψ,N)
        Pad ω and θ if new features are added.
        s ← s′
    until s is terminal;
```

Greedy-GQ (lines 5–10)
iFDD⁺ (lines 11–12)

Empirical Results

- Learning Algorithms:
  Greedy-GQ (GQ), Q-Learning (i.e., GQ with β_t=0), and SARSA
- Representations:
  iFDD⁺, fixed sparse representation (FSR)
- Averaged 30 runs using RLPy framework[7].





Conclusion

- We introduced **Greedy-GQ-iFDD⁺** as an off-policy online technique with adaptive representation that has **cheap computational complexity**. Empirical results across three domains with sizes up to ***77* Billion** state-action pairs verified the effectiveness of iFDD⁺ for representation expansion and the advantage of using off-policy techniques compared to on-policy techniques when combined with iFDD⁺ for tackling **large scale domains**.

Sponsored by:

[1] P. Stone, R. S. Sutton, and G. Kuhlmann, **Reinforcement learning for RoboCup-soccer keepaway**. International Society for Adaptive Behavior, 13(3):165–188, 2005.
[2] A. Geramifard, F. Doshi, J. Redding, N. Roy, and J. P. How, **Online discovery of feature dependencies**. In Getoor, Lise and Scheffer, Tobias (eds.), International Conference on Machine Learning (ICML), pp. 881–888.ACM, June 2011.
[3] R. Parr, C. Painter-Wakefield, L.Li, and M. Littman, **Analyzing feature generation for value-function approximation**. In International Conference on Machine Learning (ICML), pp. 737–744, New York, NY, USA, 2007.
[4] C. Painter-Wakefield, and R. Parr, **Greedy algorithms for sparse reinforcement learning**. In International Conference on Machine Learning (ICML), pp. 968–975. ACM, 2012.
[5] A. Geramifard, T. Walsh, N. Roy, and J. How, **Batch iFDD: A Scalable Matching Pursuit Algorithm for Solving MDPs**, In Conference on Uncertainty in Artificial Intelligence (UAI), 2013
[6] H. Maei, C. Szepesvári, S. Bhatnagar, and R. S. Sutton. **Toward off-policy learning control with function approximation**. In International Conference on Machine Learning (ICML), pages 719–726, 2010.
[7] A. Geramifard, R. H. Klein, and J. P. How, **RLPy: The Reinforcement Learning Library for Education and Research**. http://acl.mit.edu/RLPy, April 2013.