An analysis of RL with function approximation

Francisco S. Melo (Carnegie Mellon University, USA)
Sean P. Meyn (Univ. Illinois at Urbana-Champaign, USA)
M. Isabel Ribeiro (Inst. Systems and Robotics, Portugal)
Our problem:

Convergence of reinforcement learning with function approximation

- Useful for large problems
- Useful for problems with state uncertainty
- Established for policy evaluation (TD)

Why so hard?
Some “historical” notes:

- Samuel’s checkers (Samuel, 1959)
- Tesauro’s TD-Gammon (Tesauro, 1994)
- Soft-state aggregation approaches (Singh et al., 1994; Gordon, 1995; Tsitsiklis and Van Roy, 1996)
- TD with function approximation (Tsitsiklis and Van Roy, 1996)

...  
- “Sampling-based” approaches, policy-gradient, etc...
TD(\(\lambda\)) with FA
(Tsitsiklis and Van Roy, 1996)

- **Represent value function as**
  \[
  V(x) = \sum_i \phi_i(x)w_i = \phi^\top(x)w
  \]

- **TD(0) update**
  \[
  w_{t+1} = w_t + \alpha_t \phi(x_t)d_t
  \]
  \[
  w_{t+1} = w_t + \alpha_t \phi(x_t)(r_t + \gamma V(x_{t+1}) - V(x_t))
  \]
Analysis in terms of mean ODE:

\[ \dot{w}_t = \mathbb{E} \left[ \phi(x) \left( r + \gamma V(y) - V(x) \right) \right] \]

\[ \dot{w}_t = \mathbb{E} \left[ \phi(x) \left( r + \gamma \phi^\top(y) w_t - \phi^\top(x) w_t \right) \right] \]

\[ \dot{w}_t = \mathbb{E} \left[ \phi(x) \left( r + \gamma \phi^\top(y) w_t - \phi^\top(x) w_t \right) \right] \]

\[ \dot{w}_t = \mathbb{b} + A w_t \]

Algorithm converges to

\[ w^* = A^{-1} \mathbb{b} \]
TD(\(\lambda\)) with FA (conc.)

- What does this amount to?

\[ V_{w^*}(x) = (P_{\Phi} TV_{w^*})(x) \]
What about control?

- Does the same result apply to Q-learning?

Motivation

RL with FA

NO

- For Q-learning, $\mathcal{P}_\Phi$ and "$T$" are "incompatible"
Q-learning

- Represent value function as

\[ Q(x, a) = \sum_i \phi_i(x, a)w_i = \phi^\top(x, a)w \]

- Q-learning update

\[ w_{t+1} = w_t + \alpha_t \phi(x_t, a_t)d_t \]

\[ w_{t+1} = w_t + \alpha_t \phi(x_t, a_t)(r_t + \gamma \max_b Q(x_{t+1}, b) - Q(x_t, a_t)) \]
Motivation

Convergence

We define

\[ \Sigma = \mathbb{E} \left[ \phi(x, a)\phi^\top(x, a) \right] \]

\[ \Sigma^*(w) = \mathbb{E} \left[ \phi(x, a^*)\phi^\top(x, a^*) \right] \]

**Result:** Under “mild” conditions on the MDP, Q-learning with FA converges w.p.1 as long as

\[ \Sigma > \gamma^2 \Sigma^*(w) \]

for all \( w \).
Sketch of the proof

- We write the associated ODE:

\[ \dot{w}_t = \mathbb{E} \left[ \phi(x, a) \left( r + \gamma \max_b \phi^T (y, b) w_t - \phi^T (x, a) w_t \right) \right] \]

- For any two initial conditions \( w_1 \) and \( w_2 \), we show that

\[ \frac{d}{dt} \| w_1 - w_2 \|_2^2 \to 0 \]
What does this mean?

- Writing down the previous condition:
  \[
  \mathbb{E} \left[ \phi(x, a) \phi^\top(x, a) \right] > \gamma^2 \mathbb{E} \left[ \phi(x, a^*) \phi^\top(x, a^*) \right]
  \]

- This happens if
  \[
  \phi(x, a) \approx \phi(x, a^*) \text{ or } \gamma \ll 1
  \]

If future important (\(\gamma \approx 1\))... generalization unreliable
On-policy vs. off-policy

- **Q-learning is off-policy**

  \[ w_{t+1} = w_t + \alpha_t \phi(x_t, a_t) (r_t + \gamma \max_b Q(x_{t+1}, b) - Q(x_t, a_t)) \]

- **On-policy methods: SARSA**

  \[ w_{t+1} = w_t + \alpha_t \phi(x_t, a_t) (r_t + \gamma Q(x_{t+1}, a_{t+1}) - Q(x_t, a_t)) \]

... must have some form of *policy adjustment.*
Convergence of SARSA

- Require the policy to be Lipschitz w.r.t. $w$ with constant $C$.

**Result:** Under “mild” conditions on the MDP, there is $C_0 > 0$ such that SARSA with FA converges w.p.1 as long as $C < C_0$. 
Sketch of the proof

- We write the associated ODE:

\[
\dot{w}_t = \mathbb{E} \left[ \phi(x, a) \left( r + \gamma \phi^\top(y, b)w_t - \phi^\top(x, a)w_t \right) \right]
\]

- For any two initial conditions \(w_1\) and \(w_2\), we show that

\[
\frac{d}{dt} \|\tilde{w}\|_2^2 \leq \tilde{w}^\top (A + \lambda I)\tilde{w}
\]

where \(A\) is negative definite and \(\lambda \to 0\) with \(C\).
Discussion

- Second result recovers result from (Perkins & Precup, 2003)
- Sampling policy cannot become completely greedy (not Lipschitz)
- Conditions are sufficient, not necessary
- Incompatibility of $P_\Phi$ and "T" solved by using other "projections" (Szepesvari & Smart, 2004)