icml2008

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)

Carnegie Mellon

FCT Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

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?

RL with function approximation

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(λ) 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

$$w_{t+1} = w_t + \alpha_t \phi(x_t) d_t$$

$$w_{t+1} = w_t + \alpha_t \phi(x_t) \underbrace{\left(r_t + \gamma V(x_{t+1}) - V(x_t)\right)}_{d_t}$$

TD(λ) with FA (cont.)

- Analysis in terms of mean ODE:

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

$$\dot{w}_t = \mathbb{E}\Big[\underbrace{\phi(x)\big(r}_{\mathbf{b}} + \gamma\underbrace{\phi^\top(y)w_t - \phi^\top(x)w_t\big)}_{\mathbf{A}w_t}\Big]$$

$$\dot{w}_t = \mathbf{b} + \mathbf{A}w_t$$

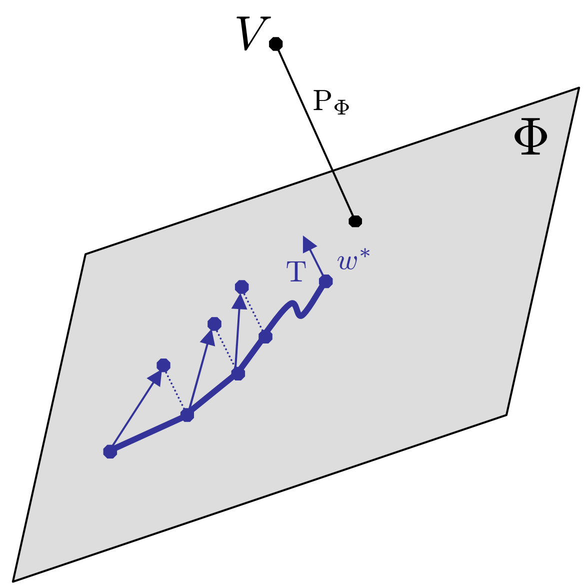
- Algorithm converges to

$$w^* = \mathbf{A}^{-\mathbf{1}}\mathbf{b}$$

TD(λ) with FA (conc.)

- What does this amount to?

$$V_{w^*}(x) = \big(\mathcal{P}_\Phi \mathbf{T} V_{w^*}\big)(x)$$



What about control?

- Does the same result apply to Q-learning?

NO

- For Q-learning, $\mathcal{P}_\Phi$ and “$\mathbf{T}$” 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

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

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

Convergence of Q-learning

- 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:

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

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

$$\frac{d}{dt}\|w_1 - w_2\|_2^2 \rightarrow 0$$

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

$\phi(x,\, a) \approx \phi(x,\, a^*)$ or $\gamma \ll 1$

If future important ($\gamma \approx 1$)... generalization unreliable

On-policy vs. off-policy

- Q-learning is off-policy

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

- On-policy methods: SARSA

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

... 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:

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

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

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

where $\mathbf{A}$ 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 $\mathcal{P}_\Phi$ and "$\mathbf{T}$" solved by using other "projections" (Szepesvari & Smart, 2004)