Function Approximation

VFA: Monte Carlo

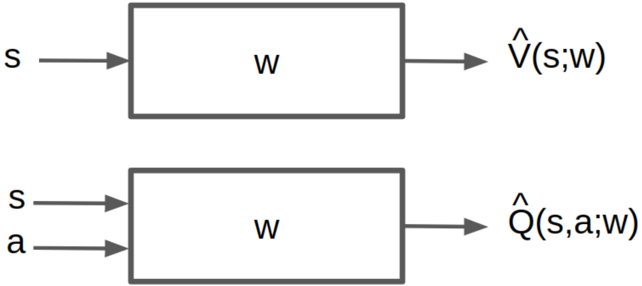
Marius Lindauer

Leibniz Universität Hannover

Winter Term 2021

Overview

- Represent a (state-action/state) value function with a parameterized function instead of a table



- Which function approximator

Monte Carlo Value Function Approximation (VFA)

- Return G_t is an unbiased but noisy sample of the true expected return $V^\pi(s_t)$
- Therefore, we can reduce MC VFA to doing supervised learning on a set of (state, return) pairs; $\langle s_1, G_1 \rangle, \langle s_2, G_2 \rangle, \dots, \langle s_T, G_T \rangle$
  - Substitute G_t for the true $V^\pi(s)$ when fit function approximator
- Concretely when using linear VFA for policy evaluation

$$
\begin{aligned}
\Delta\vec{w} &= \alpha(G_t - \hat{V}(s_t;\vec{w}))\nabla_{\vec{w}\hat{V}(s_t;\vec{w})} \\
&= \alpha(G_t - \hat{V}(s_t;\vec{w}))\vec{x}(s_t) \\
&= \alpha(G_t - \vec{x}(s_t)^T\vec{w})\vec{x}(s_t)
\end{aligned}
$$

- Note: G_t may be a very noisy estimate of true return
- Note(2): We dropped the factor 2 and see it as part of α

MC Linear Value Function Approximation for Policy Evaluation

Initialize $\vec{w} = \mathbf{0}$, $k = 1$

Loop

- Sample k-th episode $s_{k,1}, a_{k,1}, r_{k,1}, s_{k,2}, a_{k,2}, r_{k,2}, \dots$
- for $t = 1, \dots, L_k$ do
  - If First visit to $s_{k,t}$ in episode k then
    - $G_t(s) = \sum_j^{L_k} r_{k,j}$
    - Update weights by $\alpha(G_t - \vec{x}(s_{k,t})^T \vec{w})\vec{x}(s_{k,t})$
- $k = k + 1$

Convergence Guarantees for Linear Value Function Approximation for Policy Evaluation: Preliminaries

- For infinite horizon, the Markov Chain defined by an MDP with a particular policy will eventually converge to a probability distribution over states $d(s)$
- $d(s)$ is called the stationary distribution over states of π
- $\sum_s d(s) = 1$
- $d(s)$ satisfies the following balance equation:

$$d(s') = \sum_s \sum_a \pi(a \mid s) p(s' \mid s, a) d(s)$$

Convergence Guarantees for Linear Value Function Approximation for Policy Evaluation [Tsitsiklis and Van Roy. 1997]

- Define the mean squared error of a linear value function approximation for a particular policy π relative to the true value as

$$\text{MSVE}(\vec{w}) = \sum_{s \in S} d(s)(V^\pi(s) - \hat{V}^\pi(s; \vec{w}))^2$$

- where
  - $d(s)$: stationary distribution of π in the true decision process
  - $\hat{V}^\pi(s; \vec{w}) = \vec{x}(s)^T \vec{w}$, a linear value function approximation
- Monte Carlo policy evaluation with VFA converges to the weights $\vec{w}_{MC}$ which has the minimum mean squared error possible:

$$\text{MSVE}(\vec{w}_{MC}) = \min_{\vec{w}} \sum_{s \in S} d(s)(V^\pi(s) - \hat{V}^\pi(s; \vec{w}))^2$$