Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

October 4, 2021

Temporal difference learning

A stochastic approximation algorithm for **policy evaluation**

Tabular TD(0):

$$v^{(t+1)}(s_t) = v^{(t)}(s_t) + \eta_t \left(r_t + \gamma v^{(t)}(s_{t+1}) - v^{(t)}(s_t) \right)$$

What does this converge to? We're going to study a more general form an introduce function approximation. More specifically, we consider a linear model of the form:

$$v(s; w) = \phi(s)^\top w \quad ,$$

where $\phi : \mathcal{S} \to \mathbb{R}^k$ is a given feature mapping and $w \in \mathbb{R}^k$ is a weight vector.

TD(0) with linear function approximation

We've entered the realm of approximate DP last week via Stochastic Approximation, which gave us randomized algorithms with the important property of being *model-free**: which do not require knowledge of P, r directly, but only samples of the induced process.
Now, we are adding one more layer of approximation: that of approximation of the values across states. This is crucial in large or infinite problems.

TD(0) with linear function approximation

$$w^{(t+1)} = w^{(t)} + \eta_t \left(r_t + \gamma v(s_{t+1}, w^{(t)}) - v(s_t; w^{(t)}) \right) \phi_t \quad .$$

where $\phi_t = \phi(s_t)$. This notational detail is important because it means that also don't have to observe the underlying states directly: only observations of it through the mapping ϕ (most likely nonlinear), which also need not be known.

Tabular case

The *tabular* case can be obtained for $\phi(s) \triangleq e_s$: a *one-hot* encoding.

Analysis: the ODE approach

Remember the key idea in the ODE approach for the analysis of stochastic approximation algorithms: under the conditions, we can approximate the behavior of algorithm by a continuous-time dynamical system. We obtain his deterministic system by averaging out the noise: by studying the mean iterates.

Underlying stochastic process

How are we going to average out this noise? Under which distribution? The natural contender is to take the **stationary distribution** induced by running the given policy insider our MDP.

A Markov chain need not have a stationary distribution!

We write $x_d \in \mathbb{R}^{|\mathcal{S}|}$ to denote the stationary distribution induced by a stationary policy of the decision rule $d \in \mathcal{D}^{MR}$ if:

$$x_d^\top = x_d^\top P_d \quad .$$

A unique stationary distribution x_d exists if the Markov chain is **irreducible and aperiodic**.

TD(0) under the stationary distribution

The ODE approximation of TD(0) is described by the linear system:

$$\Phi^\top X\,(\Phi w - \gamma P_d \Phi w - r_d) = 0 \quad .$$

where $\Phi \in \mathbb{R}^{|\mathcal{S}\times k}$ is a matrix containing the $\phi(s)$ as rows. Furthermore, $X \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is a diagonal matrix containing the stationary distribution corresponding to d on the diagonal.

In order to ensure that w is unique, we often assume that Φ is full rank.

Expectation

$$\Phi^\top X \left(I - \gamma P_d\right) \Phi w = \Phi^\top X r_d \quad .$$

Important terms:

$$\begin{aligned}
\Phi^\top X \Phi &= \sum_{i \in \mathcal{S}} x(i) \phi(i) \phi(i)^\top = \mathbb{E}\left[\phi(S_t)\phi(S_t)^\top\right] \\
\Phi^\top X P \Phi &= \sum_{i \in \mathcal{S}} x(i) [P_d]_{ij} \phi(i) \phi(j)^\top = \mathbb{E}\left[\phi(S_t)\phi(S_{t+1})^\top\right] \quad .
\end{aligned}$$

Therefore:

$$\Phi^\top X \left(I - \gamma P_d\right) \Phi w - \Phi^\top X r_d = \mathbb{E}\left[\phi(S_t)(\phi(S_t)^\top w - \gamma \phi(S_{t+1})^\top w - r(S_t, A_t)\right]$$