Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

October 13, 2021

Recap

Over the last few lectures, we studied TD(0) with linear function approximation under the stochastic approximation framework. We saw that TD(0):

$$w^{(t+1)} = w^{(t)} + \eta_t \left(r_t + \gamma \phi_{t+1}^\top w^{(t)} - \phi_t^\top w^{(t)} \right) \phi_t \quad .$$

can be seen as an form root-finding stochastic approximation:

$$x^{(t+1)} = x^{(t)} + \eta_t \left(\bar{c} - y_t \right) \quad ,$$

corresponding to the root-finding problem:

$$\bar{c} - f(x) = 0 \quad ,$$

in which we have only noisy observations of $f(x)$.

Recap: Mean iterates

We then analyzed TD(0) using the ODE method and found that the mean iterates can be written as:

$$\begin{aligned}\bar{w}^{(k)} &= \bar{w}^{(k)} + \eta_k \mathbb{E}\left[\left(R_t + \phi_t^\top \bar{w}^{(k)} - \gamma \phi_{t+1}^\top \bar{w}^{(k)}\right) \phi_t\right] \\ &= \bar{w}^{(k)} + \eta_k \left(\Phi^\top X r_d - \Phi^\top X \left(I - \gamma P_d\right) \Phi \bar{w}^{(k)}\right) \quad .\end{aligned}$$

This lead us to study the corresponding linear ODE:

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

Recap: Asymptotic stability for linear ODEs

Consider an ODE of the form:

$$\dot{x}(t) = Ax(t) \quad .$$

An equilibrium solution in this case is asymptotically stable if the real part of the **eigenvalues** of A are **negative**.

Another equivalent characterization (used by Sutton in the analysis of TD), is that for some positive definite matrix M:

$$A^\top M + MA \quad ,$$

is negative definite.

Recap: Operator-theoretic viewpoint

Instead of going through the above route, we showed instead that the mean iterates coincide with that of a **projected** operator. That is, we have shown that $w^\star$ is the unique fixed point of the composed operator TL_d, ie:

$$\Phi w^\star = TL_d(\Phi w^\star) \triangleq T\left(r_d + \gamma P_d \Phi w^\star\right) \quad ,$$

where T computes the projection of $L_d \Phi w$ for any w onto the representable subspace.

Recap: Convergence

The main ingredient of our analysis was to establish what I call the "on-policy inequality", the fact that

$$\|Pz\|_x \leq \|z\|_x, \quad \forall z \in \mathbb{R}^n$$

if x is the stationary distribution of the Markov chain under P.

Error bound

We can show that (prop 6.3.1) that the error can be bounded by:

$$\|\underbrace{v_d}_{\text{true value function}} - \Phi \overbrace{w^\star}^{\text{TD(0) solution}} \|_x \leq \frac{1}{\sqrt{1-\gamma^2}} \|v_d - \underbrace{Tv_d}_{\text{projection of the true value function}} \|_x \quad .$$

Reducing the bias

We can control the bias using a variant of TD called TD(λ). The idea is to use a multi-step policy evaluation operator:

$$L_d^{(\lambda)} \triangleq (1-\lambda)\sum_{k=0}^{\infty} \lambda^k L_d^{k+1} \quad ,$$

with $\lambda \in [0,1]$). Note that L_d^k denotes the k-application of the *single*-step operator L_d, ie: $L^1 = L, L^2 = LL, \ldots$. We can then consider the fixed-point problem $L_d^{(\lambda)} v = v$ where:

$$\begin{aligned}
L_d^{(\lambda)} v &= r_d^{(\lambda)} + \gamma P_d^{(\lambda)} v \\
r_d^{(\lambda)} &\triangleq \sum_{k=0}^{\infty} (\gamma \lambda P_d)^k \, r_d = (I - \gamma\lambda P_d)^{-1} r_d \\
P_d^{(\lambda)} &\triangleq (1-\lambda) \sum_{k=0}^{\infty} (\gamma\lambda)^k \, P^{k+1} = (I - \gamma\lambda P_d)^{-1}(1-\lambda) P_d
\end{aligned}$$

Matrix splitting interpretation

Let $M_d^{(\lambda)} \triangleq I - \gamma\lambda P_d$ and $N_d^{(\lambda)} \triangleq \gamma(1-\lambda)P_d$, we have that:

$$I - \gamma P_d = M_d^{(\lambda)} - N_d^{(\lambda)} \quad .$$

The pair $M_d^{(\lambda)}, N_d^{(\lambda)}$ is said to be a *matrix splitting* (Varga, 1961) of $I - \gamma P_d$. Therefore:

$$\begin{aligned} L_d^{(\lambda)} v &= r_d^{(\lambda)} + \gamma P^{(\lambda)} v \\ &= \left(M_d^{(\lambda)}\right)^{-1} \left(r_d + N_d^{(\lambda)} v\right) \quad . \end{aligned}$$

Two extremes

If $\lambda = 0$, we get the usual single-step policy evaluation operator:

$$L^{(0)}v = L_d v = r_d + \gamma P_d v \quad .$$

If $\lambda = 1$, we solve for the value of d^∞ in one application of $L_d^{(1)}$:

$$L^{(1)}v = (I - \gamma P_d)^{-1} r_d = v_d \quad .$$

Matrix splitting methods are **consistent**, ie:

$$\begin{aligned} v &= M^{-1} r_d + M^{-1} N v \\ \Leftrightarrow (I - M^{-1}N)v &= M^{-1} r_d \\ \Leftrightarrow (M - N) &= r_d \quad . \end{aligned}$$

(I dropped the sub/superscripts for clarity)

Combination with function approximation

Compositing the projection operator T with $L_d^{(\lambda)}$ gives us a linear system of equations of the form:

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

In this case, we get an error bound of the form:

$$\|v_d - \Phi w^\star\|_x \leq \frac{1}{\sqrt{1-\beta^2}} \|v_d - T v_d\|_x \quad .$$

where the only thing that has now changed is the coefficient:

$$\beta = \frac{\gamma(1-\lambda)}{1-\gamma\lambda} \quad .$$

Geometry

Consequence: the contraction factor decreases with λ increasing and the error/bias decreases.

With $\lambda = 1$, we get the best achievable error: ie. the projection of v_d onto the representable subspace.

(Picture to be drawn on the board)

Stochastic approximation counterpart

What we talked about so-far can be thought as the linear ODE corresponding to the the SA counterpart that we call TD(λ), whose iterates are of the form:

$$
\begin{aligned}
w^{(t+1)} &= w^{(t)} + \eta_t z_t \delta_t \\
\delta_t &= r_t + \gamma \phi_{t+1} w_t - \phi_t w_t \\
z_t &= \sum_{k=0}^{t} (\gamma\lambda)^{t-k} \phi_k \quad ,
\end{aligned}
$$

or equivalently $z_t = \gamma\lambda z_{t-1} + \phi_t \quad .$

Fitted Value Methods

Remember that when writting:

$$\Phi w = T L_d \Phi w \quad ,$$

T can be conceptualized as an optimization procedure which solves an L_2 minimization problem. For example, imagine that we're at the k iterate with $v^{(k)} = \Phi w^{(k)}$, T is the operator which returns the unique minimizer $v^{(k+1)}$ of:

$$\text{minimize} \;\; J(v; v^{(k)}) \triangleq \|v - L_d v^{(k)}\|_x^2 = \mathbb{E}\left[\left(v(S_t) - \left(L_d v^{(k)}\right)(S_t)\right)^2\right] \quad ,$$

where the expectation is taken under the stationary distribution x.

Fitted Value Methods

$$
\begin{aligned}
\text{minimize } \; & \mathbb{E}\left[\left(v(S_t; w) - \left(L_d v^{(k)}\right)(S_t)\right)^2\right] \\
&= \mathbb{E}\left[\left(v(S_t; w) - \mathbb{E}\left[r(S_t, A_t) + \gamma v^{(k+1)}(S_{t+1}) \,\middle|\, S_t\right]\right)^2\right] \quad .
\end{aligned}
$$

where w is the optimization variable.

Key idea: **given** $v^{(k)}$, we can approximately compute TL_d as a supervised learning problem.

There is a supervised learning problem for each step of successive approximation; not a single static objective.