Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

October 6, 2021

Recap: stochastic approximation

In root-finding SA were we want to find a solution to:

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

but only via noisy observations of $f(x)$. This leads to the SA iterates:

$$x^{(k+1)} = x^{(k)} + \eta_k \left(\bar{c} - y_k\right) \quad ,$$

Under some assumptions on the type of noise, we saw that we can approximate the above by the ODE:

$$\dot{x}(t) = \left(\bar{c} - f(x(t))\right) \quad .$$

If $x^\star$ is an asymptotically stable equilibrium of the ODE, then $x^{(k)} \to x^\star$ with probability one.

TD(0) witih linear function approximation

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

How can we think of this as a stochastic root-finding problem? Noisy observations of which function? Conceptually, we want to find a w such that $f(w) = 0$, but instead of observing $f(w)$, we only get to observe $y_t = \delta_t \phi_t$ and have a SA recursion of the form $w^{(t+1)} = w^{(t)} + \eta_t \delta_t \phi_t = w^{(t)} + \eta_t y_t$.

⚠ We don't know what that f is just yet! This is what we are about to find out in the next slides.

Mean iterates

Let's average out the iterates under the stationary distribution of d^{∞}:

$$\bar{w}^{(k+1)} = \bar{w}^{(k)} + \eta_k \mathbb{E}\left[\left(R_t + \phi_t^\top \bar{w}^{(k)} - \gamma \phi_{t+1}^\top \bar{w}^{(k)}\right)\phi_t\right] \quad .$$

Here $\phi_t \triangleq \phi(S_t)$, $\phi_{t+1} \triangleq \phi(S_{t+1})$, $R_t \triangleq r(S_t, A_t)$ are random variables.

The above expectation is linear function of $\bar{w}^{(k)}$, therefore, we can also write it in matrix form as:

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

TD(0) ODE

We therefore have a linear ODE of the form:

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

and if $w^\star$ is an asymptotically stable equilibrium of f, then $w^{(k)} \to w^\star$ with probability one.

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.

Operator-theoretic viewpoint

Instead of the above two analysis methods, we are instead going to leverage an operator theoretic perspective on our problem. Consider again the deterministic iterates:

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

This can be seen as an instance of Richardson iteration for solving the linear system of equations:

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

Or equivalently:

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

Weighted Euclidean norm

Definition We write $\| \cdot \|_x$ to denote the weithed Euclidean norm on $\mathbb{R}^n$. That is, if $v \in \mathbb{R}^n$, then:

$$\|v\|_x \triangleq \sqrt{\sum_{i=1}^{n} x_i v_i^2}$$

Normal equation

The key observation is that:

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

is a normal equation corresponding to a projection. More precisely, if we find a $\hat{w}$ that satisfies the above, then it must also be that:

$$\hat{w} = \arg\min_{w \in \mathbb{R}^m} \|\Phi w - (r_d + \gamma P_d \Phi \bar{w})\|_X^2$$

We made the assumption that Φ is full rank, which means that the set of minimizer is a singleton.

Variational problem

Let T be an operator projecting onto the space $\mathcal{B}$ spanned by the columns of Φ (ie. any vector in that space can be written as a unique linear combination of the vectors in the basis).

The meaning of T being a projection is that that is given any $v \in \mathbb{R}^{|\mathcal{S}|}$, Tv returns the unique vector from $\mathcal{B}$ that minimizes $\|v - \hat{v}\|_x^2$ for any $\hat{v} \in \mathcal{B}$. That is:

$$Tv = \Phi\hat{w} \;\text{ where }\; \hat{w} = \arg\min_{w \in \mathbb{R}^m} \|v - \Phi w\|_x^2$$

Composition of operators

In our case, we want to project $L_d(\Phi w) \in \mathbb{R}^{|\mathcal{S}|}$. This means that we want to find a $\hat{w} \in \mathbb{R}^m$ such that:

$$\hat{w} = \arg\min_{w \in \mathbb{R}^m} \|\Phi w - (r_d + \gamma P \Phi \hat{w})\|_x^2$$

The **projected policy evaluation operator** is the composition of the projection operator T with the policy evaluation operator L_d. The corresponding fixed-point problem is then to find a $w \in \mathbb{R}^k$ such that:

$$TL_d(\Phi w) = \Phi w \quad .$$

But do we have a contraction?

Wouldn't be nice if TL_d were to be a contraction? We could then leverage Banach's fixed-point theorem to prove the existence of a unique solution + get an algorithm to find it for free. (Spoiler: yes, it can be).

Two notions to see before we get there: 1. Projections are nonexpansives 2. On-policy inequality

1+2 + contractivity of L_d will allows to build our proof.

Nonexpanive mapping

Projections are nonexpansive, this means that:

$$\|Tv - Tu\|_x \leq \|v - u\|_x, \forall v \in \mathbb{R}^{|\mathcal{S}|}, u \in \mathbb{R}^{|\mathcal{S}|} \quad .$$

Also, by the Pythagorean theorem:

$$\|Tv - Tu\|_x^2 = \|T(v - u)\|_x^2 \leq \|T(v - u)\|^2 + \|(I - T)(v - u)\|_x^2 = \|v - u\|_x^2$$

Therefore TL_d is a contraction with respect to the norm $\|\cdot\|_x$ if T is a contraction with respect to $\|\cdot\|_x$ because:

$$\|TL_dv - TL_du\|_x \leq \|Tv - Tu\|_x \leq \gamma\|v - u\|_x \ .$$

We saw that L_d is γ-contraction with respect to the sup-norm, but it doesn't have to be the case for any weighted norm $\|\cdot\|_x$. Because of that, we need to impose conditions on x to ensure that it's the case.

On-policy inequality

Therorem Let P be the transition matrix of some Markov chain with stationary distribution x, then:

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