Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

September 29, 2021

Robbins-Monro Algorithm

In the deterministic case, Newton's method allowed us to find the zeros of a function $f : \mathbb{R}^n \to \mathbb{R}^n$ using the iterates:

$$x^{(k+1)} = x^{(k)} - [Df(x^{(k)})]^{-1} f(x^{(k)}) \quad .$$

If we're close enough to $x^\star$ in $f(x^\star) = 0$, then the inverse Jacobian plays a negligeable role (the slope is very weak) and we might as well set:

$$x^{(k+1)} = x^{(k)} - \eta_k f(x^{(k)}) \quad .$$

for $\eta_k > 0$ sufficiently small. The above doesn't require differentiability!

Averaging across iterates

Now consider a setting where $f(x)$ is not observed directly but where instead we observe some r.v. Y_k. A possible idea would be to average N of them at every $k = 0, 1, \ldots$ and compute:

$$x^{(k+1)} = x^{(k)} - \frac{\eta_k}{N} \sum_{i=0}^{N} y_i^{(k)} \quad .$$

The observation made by Robbins and Monro (1951) was that you might as well use only one realization of Y_k:

$$x^{(k+1)} = x^{(k)} - \eta_k y_k \quad .$$

- The sequence $\{x^{(k)}\}$ is only intermediary in finding $x^\star$: we don't care about being precise for each of them, all we care about is the end result.

The RM conditions

In order for the sequence:

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

to converge, we also need the step sizes η_k to satisfy the following conditions (the “RM” conditions):

1. $\eta_k > 0, k = 0, 1, \ldots$
2. $\eta_k \to 0$
3. $\sum_{i=0}^{\infty} \eta_k = \infty$

A fourth condition is sometimes added:

4. $\sum_{i=0}^{\infty} \eta_k^2 < \infty$

but can be weaked under some assumptions.

Implicit averaging

The importance of the decreasing steps is that it provides us with an implicit form of averaging **across iterates**.

To gain some intuition, consider the case were we want to compute the sample mean estimator online. Let $f(x) \triangleq \mathbb{E}[Y] - x$, so that $f(x) = 0$ for $x = \mathbb{E}[Y]$.

The root-finding SA algorithm is of the form:

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

If we set $\eta_k = 1/(k+1)$ and $x^{(0)} = 0$, this coincides exactly with the sample mean estimator $x^{(k)} = (1/k) \sum_{i=1}^{k} y_k$.

Root-finding Stochastic Approximation

We might as well consider problems of the form $f(x) = \bar{c}$, which are also root-finding problems: ie. find x such that $f(x) - \bar{c} = 0$.

The SA recursion then reads:

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

where y_k is a noisy observation of $f(x^{(k)})$.

Noise decomposition

To better understand the effect of noise in the SA recursion, we can write:

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \eta_k \left(\bar{c} - y_k\right) \\ &= x^{(k)} + \eta_k \left(\bar{c} - f(x^{(k)})\right) + \eta_k \underbrace{\left(f(x^{(k)}) - y_k\right)}_{\text{noise}} \ , \end{aligned}$$

where we just added and subtracted.

What assumption do we need on the noise term so that it can wash away/average out through time?

- Commmon assumption: we noise term is a **Martingale**

Martingale

A sequence of random variables $X_1, X_2, \dots$ is called a Martingale if:

$$\mathbb{E}\left[|X_i|\right] < \infty \text{ and } \mathbb{E}\left[X_{i+1} \,|\, X_1, \dots, X_i\right] = X_i, \quad i = 1, 2, \dots$$

ie. when conditioning on the past, the expected next value coincides exactly with the last one of the sequence. We define the Martingale difference as $\Delta_i = X_{i+1} - X_i$.

The Martingale difference of interest for us in the analysis of SA is $\Delta_k = Y_k - f(x^{(k)})$, so that:

$$\mathbb{E}\left[\Delta_{i+1} \,|\, \Delta_1, \dots, \Delta_i\right] = 0$$

The Martingale assumption in SA then allows us to say that the “mean change” in $\{x^{(k)}\}$ over small intervals is more important than the noise.

Asymptotic behavior

If the mean change in our estimate of the root dominates over the noise, we might as well approximate our recursion by a **deterministic system**.

An because that mean change property is only valid over small intervals of time, it makes sense to take the **continuous-time limit**.

We then model our SA recursion with an Ordinary Differential Equation (ODE).

$$x^{(k+1)} = x^{(k)} + \eta_k (\bar{c} - y_k) \qquad \text{(discrete time)}$$

$$\dot{x}(t) = \bar{c} - f(x(t)) \qquad \text{(continuous time approximation)}$$

Convergence

We can show that if $x^\star$ is an asymptotically stable point of the ODE, then $x^{(k)} \to x^\star$ with probability one.

Definition A point is asymptotically stable if:

1. **Lyapunov stable**: For every $\epsilon > 0$, there exists a $\delta(\epsilon)$ such that $\|x(t) - x^\star\| \leq \epsilon$ for all $t > 0$ when $\|x(t_0) - x^\star\| \leq \delta(\epsilon)$
2. **Asymptotically stable**: There exists an ϵ_0 such that $x(t) \to x^\star$ as $t \to \infty$ if $\|x(t_0) - x^\star\| < \epsilon_0$

Concretely: Lyapunov stable means that if you're close enough to the equilibrium, your dynamical system stays close to it. Asymptotically stable means that you also converge to the equilibrium; not just stay in the viscinity.

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.

Asymptotic stability in nonlinear ODEs

Consider a nonlinear ODE of the form:

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

with equilibrium $x^\star$ (ie. $f(x^\star) = 0$) By the Hartman–Grobman theorem, we can study the properties of this system locally around $x^\star$ via linearization.

- $x^\star$ is locally asymptotically stable if the eigenvalues of $Df(x^\star)$ all have a negative real part.

Assumptions for SA convergence by ODE method

The root-finding SA recursion

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

converges to $x^\star, f(x^\star) = \bar{c}$ under the following assumptions.

1. Gain sequence: $\eta_k > 0, \eta_k \to 0, \sum_{k=0}^{\infty} \eta_k = \infty$
2. Asymptotic stability: $x^\star$ is an asymptotically stable equilibrium of the ODE $\dot{x}(t) = \bar{c} - f(x(t))$
3. Bounded iterates: $\sup_{k \geq 0} \|x^{(k)}\| < \infty$ almost surely. The iterates $x^{(k)}$ fall within a compact subset of the domain of attraction of the ODE infinitely often.
4. Bounded noise variance
5. Vanishing noise

Q-learning and SA

The Q-learning update (tabular) was of the form:

$$Q^{(k+1)}(s,a) = (1-\eta_k)Q^{(k)}(s,a) + \eta_k(F_k Q^{(k)})(s,a), \quad s \in \mathcal{S}, a \in \mathcal{A}(s)$$

$$(F_k Q^{(k)})(s,a) = \begin{cases} r(s_t,a_t) + \gamma \max_{a' \in \mathcal{A}(s_t)} Q^{(k)}(s_t,a') & \text{if } (s,a) = (s_t,a_t) \\ Q^{(k)}(s,a) & \text{otherwise} \end{cases}$$

How is this a root-finding SA problem?

Root-finding formulation

- In this discrete case, $v_\gamma^\star$ is a fixed-point of L, that is: $Lv_\gamma^\star = v_\gamma^\star$.
- We also have an optimality operator F for Q-factors and: $FQ_\gamma^\star = Q_\gamma^\star$.

Let's view this above as a root-finding problem: ie. we want to find a Q such that $FQ - Q = 0$. The optimality operator F is defined as:

$$(FQ)(s,a) = \mathbb{E}\left[r(S_t, A_t) + \gamma \max_{a' \in \mathcal{A}(S_{t+1})} Q(S_{t+1}, a') \,\middle|\, S_t = s, A_t = a \right] \quad .$$

But we cannot (large state space, or unknown transition probability function) compute $FQ - Q$ exactly. Instead, we only have **noisy measurements** via the Monte Carlo method:

$$Y_t = r(S_t, A_t) + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t),$$

$$\mathbb{E}\left[Y_t \mid S_t = s, A_t = a\right] = (FQ - Q)(s,a) \quad .$$