ECE-GY 9223 Reinforcement Learning Spring 2019

Lecture 6 — March 8, 2019

Prof. Quanyan Zhu *Scribe: Kanishk Gandhi*

1 Overview

In the last lecture we completed our discussion on the use of Neural Networks for function approximation. We saw how we could learn the different parameters for the neural network using gradient descent and how some of the hyper-parameters involved were chosen. Then, we reviewed a deterministic iterative gradient descent based numerical optimization method and looked at its convergence properties. Subsequently, while looking at how these gradient based iterative optimization methods could be applied to fixed point equations, we segued into introducing the stochastic approximation method. We ended by introducing the Martingale Convergence Theorem, the Cronwall inequalities and a theorem showing the convergence of the Robins Monro scheme of Stochastic Approximation.

In this lecture we start by looking at how we can use iterative methods to solve a fixed point equation. We illustrate the use of these methods using the example of fictitious play. We then go back to studying the Robin's Monro Stochastic approximation method and proving the theorem of convergence for the same.

2 Iterative Methods for Fixed Point Equations

Consider the fixed point equation that we have been looking at through the course:

$$\mathbb{E}[g(r, v)] = V$$

where V is a random Variable and g is a function that depends on the random variable, but is unknown to us but its outputs can be observed. To solve the above fixed point equation, we can use iterative methods as follows:

$$r_{t+1} = (1 - \gamma)r_t + \gamma\, \mathbb{E}[g(r, v)] \tag{1}$$

Not knowing g could make solving the above fixed point equation more involved. But fortunately, we can take the aid of the Monte Carlo approximation of the function to obtain its expected value.

$$\mathbb{E}[g(r, v)] \approx \frac{1}{k} \sum_{i=1}^{K} g(r, \tilde{v_i})$$

if we choose a single sample, as we saw in the previous lecture, then the above equation becomes the Robins-Monro Stochastic Approximation method.

$$r_{t+1} = (1 - \gamma)r_t + \gamma g(r, \tilde{v}) \tag{2}$$

$$g(r, \tilde{v}) = \mathbb{E}[g(r, v)] + g(r, \tilde{v}) - \mathbb{E}[g(r, v)] \tag{3}$$

Observation 1. *Using the Monte Carlo approximation of g has advantages that are two fold:*

1. *The underlying probability distribution of the random variable V is not required to be known; we should only be able to sample from it.*
2. *We do not need to know the the underlying function g. It can be treated as blackbox as we only need to observe the output for certain samples.*

The dynamical system can further be written as :

$$r_{t+1} = (1-\gamma)r_t + \gamma(\mathbb{E}[g(r,v)]) + w_t \tag{4}$$

where, w_t is the noise represented by :

$$w_t = g(r,\tilde{v}) - \mathbb{E}[g(r,v)]$$

The noise term w_t has some nice properties that help our approximation:

1. Zero mean : $\mathbb{E}[w_t|F_t] = 0$
2. Bounded variance : $\mathbb{E}[w_t^2|F_t] < \infty$

We can then find a solution to a stochastic difference equation that is given as follows:

$$x_{n+1} = x_n + \epsilon_n(h(x_n) + \mu_n) \tag{5}$$
$$\dot{x} = h(x) \tag{6}$$

Here, $x \in \mathbb{R}^d$ and $h(.) : \mathbb{R}^d \to \mathbb{R}^d$ and x_0 is known to us. We showed in the previous class that the above difference equation converges. We now look at the example of fictitious play to see how stochastic sampling can be used in the context of matrix games.

2.1 Fictitious Play

Consider the scenario of fictitious play in a 2-person matrix game. The setting is sch that the players do not know what the payoff matrix is. They try to approximate their expected payoffs using the samples that they see. This is similar to the previous section where the function g was unknown.

The problem can be set up as follows:

- At time $t = k$, player $i = 1$ *or* 2 chooses an action from the action space $a_k^i \in A^i = \{1, \dots n^i\}$. They then receive a reward $\pi^i_{a_k^1, a_k^2}$ that can be observed.
- Let $p^i(a,k) = \frac{1}{k}\sum_{t=1}^{k} \mathbb{1}[a_t^i = a]$
- At time $k+1$, a_{k+1}^i is chosen by assuming that player i' choose an action a_k with probability $p^i(a,k)$

  $$max_{p^i \in \mathcal{P}} \sum_{a_1}\sum_{a_2} \pi^i_{a_1,a_2} p^i(a_1) p^{i'}(a_2)$$

  Where the probability of the other player is modelled using the empirical frequency at step k.

To find the optimal solution, first consider:

- $g^i(., p^{i'})$ to be the probability distribution of player i
- Player i plays by sampling an action from the distribution $g^i(., p^{i'})$. Thus, a^i_{k+1} is sampled from $g^i(., p^{i'})$

$$a^i_{k+1} \sim g^i(., p^{i'})$$

We can update our previous equations to get recursive update for the policy distribution g, using results obtained in the previous section.

$$(k+1)p^i(a, k+1) = kp^i(a, k) + \mathbb{1}(a^i_{k+1} = a) \tag{7}$$

$$p^i(a, k+1) = p^i(a, k) + \frac{1}{k+1}(\mathbb{1}(a^i_{k+1} = a) - p^i(a, k)) \tag{8}$$

Let $\epsilon_k = \frac{1}{k+1}$ and we know that

$$\mathbb{E}[\mathbb{1}(a^i_{k+1} = a)] = p(a^i_{k+1}) = g^i(., p^{i'})$$

then we get the ODE associated with the stochastic sample,

$$\dot{p^i} = g(p^i, p^{i'}) - p^i \tag{9}$$

thus, we have now found a recursive formulation for the policy in the form that was presented in the previous section. We have already seen in the previous lecture how this form converges.

3 Stochastic Approximation (Robins Monro)

Consider the following assumptions,

Assumption 1. *(Lipshcitz continuity of h) There exists $L \geq 0$ s.t. $\forall x, y \in \mathbb{R}^d$:*

$$||h(x) - h(y)|| \leq L||x - y||$$

Intuitively, this assumption says that h does not grow too quickly

Assumption 2. *(Step Size) Consider the step size ϵ_n*

$$\lim_{n\to\infty} \sum_{n\geq 0} \epsilon_n = \infty$$

$$\lim_{n\to\infty} \sum_{n\geq 0} \epsilon_n^2 < \infty$$

Assumption 3. *(Martingale Difference Noise) Consider the martingale difference noise M_n such that:*

- $\mathbb{E}[M_n|F_n] = 0$
- $\mathbb{E}[||M_n||^2|F_n] \leq K||1 + X_n||$

where, $F_n = \sigma(X_0, M_0 \dots X_1, M_1)$

Assumption 4. *(Bounded Iterates)* $\sup_{n\geq 0} ||X_n|| < \infty$ *a.s.*

Assumption 5. *(Lyapunov Crierion) There exists a positive radially unbounded continuous differentiable function* $V : \mathbb{R}^d \to \mathbb{R}$ *such that* $\forall x \in \mathbb{R}$,

$$\langle \nabla V(x), h(x) \rangle \leq 0$$

with the strict inequality if $V(x) \neq 0$

Theorem 1. *If the above assumptions hold, then* $V(x_n) \xrightarrow{n\to\infty} 0$ *a.s.*

Corollary 1. *As* $n \to \infty$*, and the above aassumptions hold,* X_n *converges to the stationary point of the ODE a.s.*

Proof. We begin by describing the overview of the approach that the proof is going to take.

- First, we are going to show that the ODE trajectory, $\dot{x} = h(x)$, is arbitrarily close to $\{X_n\}$ with suitable interpolation.
- Next, we show that there is a Lyapunov function that allows the above ODE to converge to the stationary point.

Let the timeline be $t_n = \sum_{k=0}^{n-1} \epsilon_k$. One can notice immediately that the step size decreases over time. We are interpolating using linear functions between any two points. Let $\bar{X}(t_n) = X_n$, then $\bar{X}(t)$ is linearly interpolated at $t \neq t_0, t_1, \ldots$.

Let $X^n(t)$ be the actual solution to the ODE for $t \geq t_n$ with the initial condition $X^n(t = t_0) = X_n$

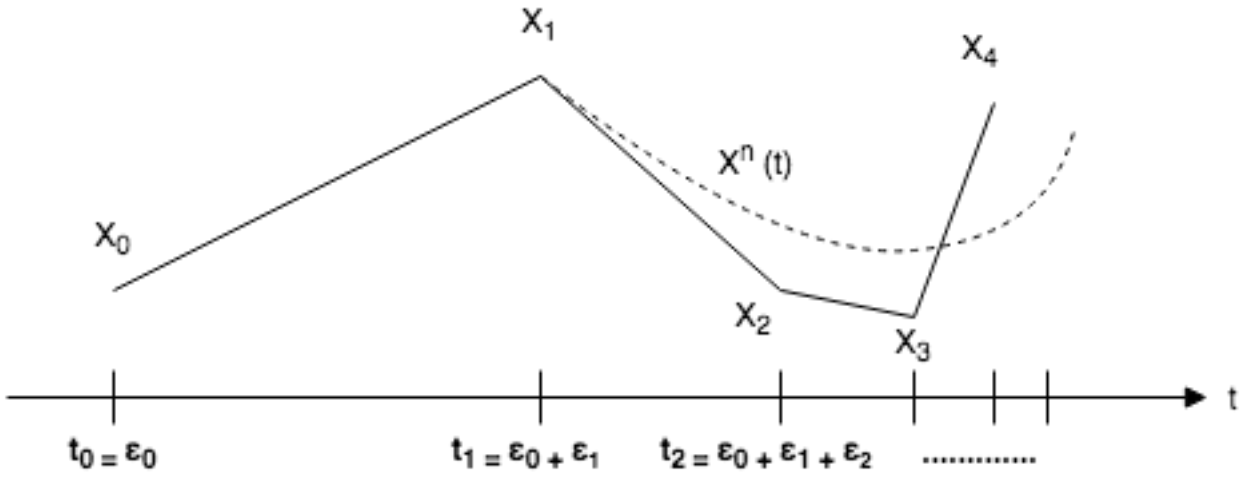

Figure 1: An illustration of the approximation

Claim 1. *For any $T > 0$, $\sup ||X^n(t) - \bar{X}(t)|| = 0$, $\forall t \in [t_n, t_{n+T}]$ as $n \to \infty$*

Proof. We want to show that when n is large, then our approximation is arbitrarily close to the actual trajectory.

Let $m = \inf\{k : t_k > t_n + T\}$, then we first want to show that

$$\sup_{n \le k \le m} ||X^n(t_k) - X_k|| \xrightarrow{n \to \infty} 0 \tag{10}$$

We can write the difference as:

$$\begin{aligned}
X^n(t_k) &= X_n + \int_{t_n}^{t_k} h(X^n(s))ds \\
&= X_n + \sum_{l=n}^{k-1} \int_{t_l}^{t_{l+1}} h(X^n(s))ds \\
&= X_n + \sum_{l=n}^{k-1} [\int_{t_l}^{t_{l+1}} h(X^n(t_l))ds + \int_{t_l}^{t_{l+1}} (h(X^n(s)) - h(X^n(t_l))ds] \\
&= X_n + \sum_{l=n}^{k-1} h(X^n(t_l))(t_{l+1} - t_l) + \sum_{l=n}^{k-1} \int_{t_l}^{t_{l+1}} (h(X^n(s)) - h(X^n(t_l))ds
\end{aligned}$$

We know that,

$$X_{n+1} = X_n + \epsilon_n(h(X_n) + \mu_n)$$

and that X_k is generated by

$$X_k = X_n + \sum_{l=n}^{k-1} \epsilon_l h(X_l) + \sum_{l=n}^{k-1} \epsilon_l \mu_l$$

where $\epsilon_l = (t_{l+1} - t_l)$ We take the difference to get:

$$||X^n(t_k) - X_k|| = ||\sum_{l=n}^{k-1} \epsilon_l (h(X^n(t_l)) - h(X_l)) + \sum_{l=n}^{k-1} \int_{t_l}^{t_{l+1}} (h(X^n(s)) - h(X^n(t_l)))ds - \sum_{l=n}^{k-1} \epsilon_l \mu_l|| \tag{11}$$

$$\le \sum_{l=n}^{k-1} \epsilon_l ||(h(X^n(t_l)) - h(X_l))|| + ||\sum_{l=n}^{k-1} \epsilon_l \mu_l|| + \sum_{l=n}^{k-1} \int_{t_l}^{t_{l+1}} ||h(X^n(s)) - h(X^n(t_l))||ds \tag{12}$$

Using Asumption 1, we know that $h(.)$ is L-Lipshitz:

$$||X^n(t_k) - X_k|| \le \sum_{l=n}^{k-1} \epsilon_l L||(X^n(t_l) - X_l)|| + ||\sum_{l=n}^{k-1} \epsilon_l \mu_l|| + \sum_{l=n}^{k-1} \int_{t_l}^{t_{l+1}} L||X^n(s) - X^n(t_l)||ds \tag{13}$$

We now try and simplify the three terms in the above inequality one at a time. We first look at the $||\sum_{l=n}^{k-1} \epsilon_l \mu_l||$ term in 13.

Observation 2. *Let $S_n = ||\sum_{l=n}^{k-1} \epsilon_l \mu_l||$. S_n is a Martingale.*

Proof.

$$\mathbb{E}[S_{n+1} - S_n | F_n] = \mathbb{E}[\epsilon_{n+1}\mu_{n+1}|\mathcal{F}_n] = 0$$

Using Assumption 3,

$$\mathbb{E}[S_{n+1}|\mathcal{F}_n] = S_n$$

Therefore, we can say that S_n is a Martingale □

At this point, the reader is advised to take a small detour and revisit the Martingale Convergence Theorem and Gronwall's Inequalities from the previous lecture.

As S_n is a Martingale,

$$\begin{aligned}
\sum_{n\geq 0} \mathbb{E}[||S_{n+1} - S_n||^2|\mathcal{F}_n] &= \sum_{n\geq 0} \mathbb{E}[||\epsilon_{n+1}\mu_{n+1}||^2|\mathcal{F}_n] \\
&\leq K \sum_{n\geq 0} \epsilon_{n+1}^2(1 + ||X_n||) && \text{Using Assumption 3} \\
&\leq K(\sum_{n\geq 0} \epsilon_{n+1}^2)(1 + \sup_n ||X_n||) && \text{Using Assumption 4} \\
&\leq \infty && \text{Using Assumption 2}
\end{aligned}$$

This tells us that

$$S_n \to S_\infty$$

We can now simplify the term $||\sum_{l=n}^{k-1} \epsilon_l \mu_l||$ as follows:

$$\begin{aligned}
||\sum_{l=n}^{k-1} \epsilon_l \mu_l|| &= ||S_{k-1} - S_{n-1}|| \\
&\leq \sup ||S_{k-1} - S_\infty|| + \sup ||S_{n-1} - S_\infty|| \\
&\leq sup_{n'\geq n}||S_{n'} - S_\infty|| \to 0
\end{aligned}$$

Thus, for a sufficiently large n and for some $\delta_1 > 0$,

$$||\sum_{l=n}^{k-1} \epsilon_l \mu_l|| \leq \delta_1 \tag{14}$$

Let us now look at the final term in the inequality 13

$$\sum_{l=n}^{k-1} \int_{t_l}^{t_{l+1}} L||X^n(s) - X^n(t_l)||ds$$

Let us start by looking at the basic ODE,

$$\begin{aligned}
\dot{X(t)} &= h(x) \\
||X(t)|| &= ||X(0)|| + \int_0^t h(X(s))ds \\
&\leq ||X(0)|| + \int_0^t ||h(X(s)) - h(0)||ds + ||h(0)||t \\
&\leq ||X(0)|| + ||h(X(0))||t + \int_0^t L||X(s)||ds && \text{Using Assumption 3} \\
&\leq ||X(0)|| + ||h(X(0))||T + \int_0^t L||X(s)||ds && \text{Let } t \in [0,T] \\
&\leq B + + \int_0^t L||X(s)||ds
\end{aligned}$$

We can now use Gronwall's inequality to get

$$\begin{aligned}
||X(t)|| &\leq B\exp(LT) = K_T && \forall t \in [0,T] \\
X^n(s) &= X^n(t_l) + \int_{t_l}^s h(X^n(u))du && s \in [t_l, t_{l+1}] \\
||X^n(s) - X^n(t)|| &\leq \int_{t_l}^s ||h(X^n(u)) - h(X(t_l))||du + ||h(X(t_l))||\epsilon_l \\
&\leq L\int_{t_l}^s ||X^n(u) - X(t_l)||du + ||h(X(t_l))||\epsilon_l && \text{Using Assumption 3} \\
&\leq L\int_{t_l}^s ||X^n(u) - X(t_l)||du + B\epsilon_l && \epsilon_l < \infty \text{ and } B < \infty
\end{aligned}$$

We can use Gronwall's inequality again to bound the above expression.

$$||X^n(s) - X^n(t)|| \leq B\epsilon_l \exp Lt_{l+1} \tag{15}$$

$$\leq \epsilon_l C_T \tag{16}$$

Going back to the final term in inequality 13,

$$\int_{t_l}^{t_{l+1}} L||X^n(s) - X^n(t_l)||ds \leq L\epsilon_l^2 C_T \tag{17}$$

$$\sum_{l=n}^{k-1}\int_{t_l}^{t_{l+1}} L||X^n(s) - X^n(t_l)||ds \leq \sum_{l=n}^{k-1} L\epsilon_l^2 C_T \tag{18}$$

$$= LC_T\sum_{l=n}^{k-1}\epsilon_l^2 \xrightarrow{n\to\infty} 0 \tag{19}$$

Thus, we can write for a sufficiently large n and $\delta_2 > 0$,

$$\sum_{l=n}^{k-1}\int_{t_l}^{t_{l+1}} L||X^n(s) - X^n(t_l)||ds < \delta_2 \tag{20}$$

We now return to finding bound for the first term in inequality 13,

$$\sum_{l=n}^{k-1} L\epsilon_l ||X^n(t_l) - X_l||$$

We define $\Delta_k = ||X^n(t_k) - X_k||$, then we can write the inequality 13 as,

$$\Delta_k \leq \sum_{l=1}^{k-1} L\epsilon_l \Delta_l + \delta_1 + \delta_2 \qquad \text{Using equation 14 and 20} \tag{21}$$

$$\leq (\delta_1 + \delta_2) \exp(\sum_{l=1}^{k-1} L\epsilon_l) \tag{22}$$

$$= \delta \exp(LT) \xrightarrow{n\to\infty} 0 \tag{23}$$

Now we go on to show that the error due to interpolation is small. We know that $X^n(t)$ is a solution to the ODE $\dot{X} = h(X)$. We can then write,

$$X^k(t) = X^k(t_k) + \int_{t_k}^{t} h(X^k(s))ds \tag{24}$$

$$= X^k(t_{k+1}) - \int_{t}^{t_{k+1}} h(X^k(s))ds \tag{25}$$

$$\tag{26}$$

therefore,

$$\begin{aligned} ||\bar{X}(t) - X^k(t)|| &= ||\lambda X_k + (1-\lambda)X_{k+1} - \lambda x^k(t) - (1-\lambda)X^k(t)|| \\ &\leq \lambda||X_k - X^k(t_k)|| + (1-\lambda)||X_{k+1} - X^k(t_{k+1})|| + \\ &\lambda \int_{t_l}^{t} ||h(X^k(s))||ds + (1-\lambda)\int_{t}^{t_{k+1}} h(X^k(s))ds \\ &= ||\lambda X_k + (1-\lambda)X_{k+1} - \lambda X^k(t_k) - \lambda \int_{t}^{t_{k+1}} h(X^k(s))ds \\ &- (1-\lambda)X^k(t_{k+1}) - (1-\lambda)\int_{t}^{t_{k+1}} h(X^k(s))ds|| \end{aligned}$$

Thus, we can finally write,

$$\sup_{t\in[t_k,t_m]} ||\bar{X}(t) - X^k(t)|| \leq \sup_{k\leq k'\leq m} ||X_{k'} - X^k(t_{k'})|| + \epsilon_n C_T \to 0 \text{ as } n \to \infty \tag{27}$$

□

We have proved that the arbitrary closeness of the linear approximation to the actual trajectory. In the next class we complete the proof for the theorem and start our discussion on Q-learing. □

References

[1] Robbins, H., Monro, S. (1951). A stochastic approximation method. The annals of mathematical statistics, 400-407.