Neuro Dynamic Programming

Peng Lingwei

August 31, 2019

Contents

1 Introduction

2 Dynamic Programming

Definition 1. ***(Proper stationary policy).*** *(Reach termination state 0 w.p.1)*

$$\rho^\pi = \max_{i=1,\dots,n} P^\pi \{s_n \neq 0 | i_0 = i\} < 1$$

In stochastic shortest path problems, we have two assumptions:

- There exists at least one proper policy;
- For every improper policy π, the corresponding cost-to-oo $J^\pi(i)$ is infinite for at least one state i.

Policy Iteration as an Actor-Critic System

- Critic: policy evaluation;
- actor: policy improvement.

3 Neural Network Architectures and Training

Trivial. Using some function to approximate V^π, V^*, Q^π, Q^*. This book uses neural network.

4 Stochastic Iterative Algorithms

Suppose that we are interested in solving a system of equations of the form

$$Hr = r,$$

where H is a function from $\mathbb{R}^n$ into itself. If $Hr = r - \nabla f(r)$, the solution of the system $Hr = r$ is of the form

$$\nabla f(r) = 0,$$

Then it's sometime minimize the cost function f.

One possible algorithm for solving the system $Hr = r$ is provided by the iteration

$$r_{t+1} = Hr_t, \quad or \quad r_{t+1} = (1-\gamma)r_t + \gamma Hr_t.$$

the second method reduces to the gradient method if $Hr = r - \nabla f(r)$.

Sometimes an exact evaluation of Hr is difficult but that we have access to a random variable s of the form $s = Hr + w$, where w is a random noise term. Then we obtain stochastic iterative or stochastic approximation algorithm

$$r_{t+1} = (1-\gamma)r + \gamma(Hr + w).$$

A more concrete setting is obtained as follows. Let v be a random variable with a known probability distribution $p(v|r)$ that depends on r. Suppose that we are interested in solving:

$$\mathbb{E}_{v \sim p(v|r)} [g(r, v)] = r,$$

where g is a known function. We can use preceeding algorithm:

$$r_{t+1} = (1-\gamma)r_t + \gamma \mathbb{E}_{v\sim p(v|r)}\left[g(r,v)\right].$$

We can estimate $\mathbb{E}_{v\sim p(v|r)}\left[g(r,v)\right] \approx \frac{1}{k}\sum_{i=1}^{k} g(r,\tilde{v}_i)$. We get Robbins-Monro stochastic approximation algorithm ($k=1$),

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

which is a special case of the algorithm $r_{t+1} = (1-\gamma)r_t + \gamma\left(Hr_t + w\right)$, where $Hr = \mathbb{E}_{v\sim p(v|r)}\left[g(r,v)\right]$, and $w = g(r,\tilde{v}) - \mathbb{E}\left[g(r,v)\right]$.

4.1 THE BASIC MODEL

Let T^i be the set of times at which $r(i)$ updates:

$$r_{t+1}(i) = \begin{cases} r_t(i), & t \notin T^i \\ (1-\gamma_t(i))r_t(i) + \gamma_t(i)\left((Hr_t)(i) + w_t(i)\right), & t \in T^i \end{cases}$$

Assumption: $\sum_{t=0}^{\infty} \gamma_t(i) = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2(i) < \infty$.

4.2 CONVERGENCE BASED ON A SMOOTH POTHENTIAL FUNCTION

$$r_{t+1} = r_t + \gamma_t \delta_t, \quad \delta_t = Hr_t - r_t + w_t.$$

Let $\mathcal{H}_t$ denote the history of the algorithm

$$\mathcal{H}_t = \left\{r_0, \ldots, r_t, \delta_0, \ldots, \delta_{t-1}, \gamma_0, \ldots, \gamma_t\right\}.$$

Assumption 1. *Exist function $f : \mathbb{R}^n \mapsto \mathbb{R}$, with the following properties:*

1. $\forall \mathbb{R}^n, f(r) \geq 0$*;*
2. $\|\nabla f(r_1) - \nabla f(r_2)\| \leq L\|r_1 - r_2\|_2$*;*
3. *(Pseudogradient property)* $c\|\nabla f(r_t)\|_2^2 + \langle \nabla f(r_t), \mathbb{E}\left[\delta_t|\mathcal{H}_t\right]\rangle \leq 0$
4. $\mathbb{E}\left[\|\delta_t\|_2^2|\mathcal{H}_t\right] \leq K_1 + K_2\|\nabla f(r_t)\|_2^2$

Proposition 1. *Consider the algorithm $r_{t+1} = r_t + \gamma_t s_t$, if $\sum_{t=0}^{\infty}\gamma_t = \infty$ and $\sum_{t=0}^{\infty}\gamma_t^2 < \infty$. Under preceeding assumption, the following hold with probability 1:*

- *The sequence $f(r_t)$ converges;*
- $\lim_{t\to\infty} \nabla f(r_t) = 0$*;*
- *Every limit point of r_t is a stationary point of f.*

Example 1. ***(Stochastic Gradient Algorithm).***

$$r_{t+1} = r_t + \gamma_t \delta_t, \quad \delta_t = -\left(\nabla f(r_t) + w_t\right)$$

Assumption:

1. $\sum_{t=0}^{\infty} \gamma_t = \infty, \quad \sum_{t=0}^{\infty} \gamma_t^2 < \infty$;

2. f *is nonnegative and has a Lipschitz continuous gradient;*

3. $\mathbb{E}\left[w_t|\mathcal{H}_t\right] = 0, \quad \mathbb{E}\left[\|w_t\|^2|\mathcal{H}_t\right] \leq A + B\|\nabla f(r_t)\|_2^2$;

We proof Assumption 1 is satisfied.

$$\langle \nabla f(r_t), \mathbb{E}\left[\delta_t|\mathcal{H}_t\right]\rangle = \langle \nabla f(r_t), -\nabla f(x_t) - \mathbb{E}\left[w_t|\mathcal{H}_t\right]\rangle = -\|\nabla f(r_t)\|_2^2$$

$$\begin{aligned}
\mathbb{E}\left[\|\delta_t\|_2^2|\mathcal{H}_t\right] =& \|\nabla f(r_t)\|_2^2 + \mathbb{E}\left[\|w_t\|_2^2|\mathcal{H}_t\right] + \langle 2\nabla f(r_t), \mathbb{E}\left[w_t|\mathcal{H}_t\right]\rangle \\
=& \|\nabla f(r_t)\|_2^2 + A + B\|\nabla f(r_t)\|_2^2 \\
=& A + (B+1)\|\nabla f(r_t)\|_2^2
\end{aligned}$$

Example 2. ***(Estimate of an Unknown Mean).*** *For random variables v with unknow mean μ and unit variance.*

$$r_{t+1} = (1-\gamma_t)r_t + \gamma_t v_t.$$

with assumption

1. $\sum_{t=0}^{\infty} \gamma_t = \infty$ *and* $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$;

Proof.

$$r_{t+1} = r_t - \gamma_t(r_t - \mu) + \gamma_t(v_t - \mu)$$

where $f(r) = (r-\mu)^2/2$, $\nabla f(r_t) = (r_t - \mu)$. (The other assumptions in stochastic gradient algorithm are sastified naturally.) $\square$

Example 3. ***(Euclidean Norm Pseudo-Contractions).***

$$r_{t+1} = (1-\gamma_t)r_t + \gamma_t(Hr_t + w_t),$$

Assuming:

1. $\|Hr - r^*\|_2 \leq \beta\|r - r^*\|_2, \forall r \in \mathbb{R}^n$, $0 \leq \beta < 1$;

2. $\mathbb{E}\left[w_t|\mathcal{H}_t\right] = 0$;

3. $\mathbb{E}\left[\|w_t\|_2^2|\mathcal{H}_t\right] \leq A + B\|r_t - r^*\|_2^2$

The potential function is $f(r) = \frac{1}{2}\|r - r^*\|_2^2.$, $\delta_t = -r_t + Hr_t + w_t$, *then* $\mathbb{E}\left[\delta_t|\mathcal{H}_t\right] = Hr_t - r_t$.

$$\begin{aligned}
\langle Hr - r^*, r - r^*\rangle \leq& \|Hr - r^*\|_2\|r - r^*\|_2 \leq \beta\|r - r^*\|_2^2 \\
\langle Hr - r, r - r^*\rangle \leq& -(1-\beta)\|r - r^*\|_2^2 \\
\langle \mathbb{E}\left[\delta_t|\mathcal{H}_t\right], \nabla f(r_t)\rangle \leq& -(1-\beta)\|\nabla f(r_t)\|_2^2
\end{aligned}$$

$$\mathbb{E}\left[\delta_t^2|\mathcal{H}_t\right] = \mathbb{E}\left[(-r_t + Hr_t)^2|\mathcal{H}_t\right] + \mathbb{E}\left[\|w_t\|^2|\mathcal{H}_t\right] \leq (Hr_t - r_t)^2 + A + B\|r_t - r^*\|_2^2$$

4.2.1 Convergence Proofs

In this section, we discarded a suitable set of measure zero, and don't keep repeating the qualification " with probability 1".

Theorem 1. ***(Supermartingale Convergence Theorem).*** *Here is three sequences of random variables $\{X_t\}, \{Y_t\}$ and $\{Z_t\}$. And let $\mathcal{F}_t$ be set of random variables and $\mathcal{F}_t \subset \mathcal{F}_{t+1}$. Suppose that*

1. *X_t, Y_t, Z_t are nonegative, and are functions of the random variables in $\mathcal{F}_t$;*
2. *$\forall t$, we have $\mathbb{E}[Y_{t+1}|\mathcal{F}_t] \leq Y_t - X_t + Z_t$;*
3. *$\sum_{t=0}^{\infty} Z_t < \infty$.*

Then we have $\sum_{t=0}^{\infty} X_t < \infty$, and the sequence Y_t converges to a nonegative random variable Y, w.p.1.

Theorem 2. ***(Martigale Convergence Theorem)*** *Let $\{X_t\}$ be a sequence of random variables and let $\mathcal{F}_t$ be set of random variables such that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$. Suppose that:*

1. *The random variable X_t is a function of the random variable in $\mathcal{F}_t$;*
2. *$\mathbb{E}[X_{t+1}|\mathcal{F}_t] = X_t$,*
3. *$\exists M < \infty$ such that $\mathbb{E}[|X_t|] \leq M$.*

Then, the sequence X_t converges to a random variable X, w.p.1.

Now we begin proof the preceeding section.

Proof. By assumption, we have $\|\nabla f(r_1) - \nabla f(r_2)\|_2 \leq L\|r_1 - r_2\|$, we have

$$f(r_{t+1}) \leq f(r_t) + \gamma_t \langle \nabla f(r), \delta_t \rangle + \frac{L}{2}\gamma_t^2 \|\delta_t\|_2^2$$

$$\begin{aligned}
\mathbb{E}[f(r_{t+1}|\mathcal{F}_t)] \leq & f(r_t) + \gamma_t \langle \nabla f(r_t), \mathbb{E}[\delta_t|\mathcal{F}_t] \rangle + \frac{L}{2}\gamma_t^2 \left(K_1 + K_2\|\nabla f(r_t)\|_2^2\right) \\
\leq & f(r_t) - \gamma_t \left(c - \frac{LK_2\gamma_t}{2}\right) \|\nabla f(r_t)\|_2^2 + \frac{LK_1\gamma_t^2}{2} \\
= & f(r_t) - X_t + Z_t,
\end{aligned}$$

where

$$X_t = \begin{cases} \gamma_t \left(c - \frac{LK_2\gamma_t}{2}\right) \|\nabla f(r_t)\|_2^2, & if\ LK_2\gamma_t \leq 2c, \\ 0, & otherwise. \end{cases}$$

and

$$Z_t = \begin{cases} \frac{LK_1\gamma_t^2}{2}, & if\ LK_2\gamma_t \leq 2c, \\ \frac{LK_1\gamma_t^2}{2} - \gamma_t \left(c - \frac{LK_2\gamma_t}{2}\right) \|\nabla f(r_t)\|_2^2, & otherwise \end{cases}$$

Because $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$, so after some finite time $LK_2\gamma_t \leq 2c$, and $Z_t = LK_1\gamma_t^2/2$, and therefore $\sum_{t=0}^{\infty} Z_t < \infty$. Thus, the supermartingale convergence theorem applies and shows that $f(r_t)$ converges and $\sum_{t=0}^{\infty} X_t < \infty$.

Because $X_t = \gamma_t \left(c - \frac{LK_2\gamma_t}{2} \right) \|\nabla f(r_t)\|_2^2 \geq \frac{c}{2}\gamma_t\|\nabla f(r_t)\|_2^2$ after some finite time. Hence

$$\sum_{t=0}^{\infty} \gamma_t\|\nabla f(r_t)\|_2^2 < \infty$$

Because $\sum_{t=0}^{\infty} \gamma_t = \infty$, $\liminf_{t\to\infty} \|\nabla f(r_t)\|_2 = 0$

Let us denote $\bar{s}_t = \mathbb{E}\left[s_t|\mathcal{F}_t\right]$ and $w_t = s_t - \bar{s}_t$, then

$$\|\bar{s}_t\|_2^2 + \mathbb{E}\left[\|w_t\|_2^2|\mathcal{F}_t\right] = \mathbb{E}\left[\|s_t\|_2^2|\mathcal{F}_t\right] \leq K_1 + K_2\|\nabla f(r_t)\|_2^2$$

We need take a break and proof another lemma

Lemma 1. *$u_t = \sum_{\tau=0}^{t-1} \chi_\tau\gamma_\tau w_\tau$, converges w.p.1. where $\chi_t = 1_{\left[\|\nabla f(r_t)\|_2 \leq \epsilon\right]}$.*

Proof. We start the assumption $\sum_{t=0}^{\infty} \gamma_t^2 \leq A < \infty$.

$$\mathbb{E}\left[\chi_t\gamma_t w_t|\mathcal{F}_t\right] = \chi_t\gamma_t\mathbb{E}\left[w_t|\mathcal{F}_t\right] = 0 \Rightarrow \mathbb{E}\left[u_{t+1}|\mathcal{F}_t\right] = u_t$$

If $\chi_t = 0$, then $\mathbb{E}\left[\|u_{t+1}\|_2^2|\mathcal{F}_t\right] = \|u_t\|^2$. If $\chi_t = 1$, we have

$$\mathbb{E}\left[\|u_{t+1}\|_2^2|\mathcal{F}_t\right] = \|u_t\|_2^2 + \gamma_t^2\mathbb{E}\left[\|w_t\|_2^2|\mathcal{F}_t\right] \leq \|u_t\|_2^2 + \gamma_t^2(K_1 + K_2\epsilon^2)$$

$$\mathbb{E}\left[\|u_t\|_2^2\right] \leq (K_1 + K_2\epsilon^2)\mathbb{E}\left[\sum_{\tau=0}^{t-1} \gamma_\tau^2\right] \leq (K_1 + K_2\epsilon^2)A$$

$$\sup_t \mathbb{E}\left[\|u_t\|^2\right] \leq \sup_t \mathbb{E}\left[1 + \|u_t\|_2^2\right] < \infty$$

Then we can use Martigale convergence theorem to u_t and get that u_t converges, w.p.1.

We can assume that $\sum_{\tau=0}^{t-1} \gamma_\tau^2 \leq A < \infty$ and get the same result. □

I give up today.. □