2.997 Decision-Making in Large-Scale Systems — February 25
MIT, Spring 2004 — Handout #10

Lecture Note 7

1 Real-Time Value Iteration

Recall the real-time value iteration (RTVI) algorithm

$$\begin{array}{ll} \text{choose} & x_{k+1} = f(x_k, u_k, w_k) \\ \text{choose} & u_t \text{ in some fashion} \\ \text{update} & J_{k+1}(x_k) = (TJ_k)(x_k), \quad J_{k+1}(x) = (TJ_k)(x), \forall\, x \neq x_k \end{array}$$

We thus have

$$TJ_k(x_k) = \min_a \left\{ g_a(x_k) + \alpha \sum_y P_a(x_k, y) J_k(y) \right\}$$

We encounter the following two questions in this algorithm.

1. what if we do not know $P_a(x, y)$?
2. even if we know/can simulate $P_a(x, y)$, computing $\sum_y P_a(x, y)J(y)$ may be expensive.

To overcome these two problems, we consider the Q-learning approach.

2 Q-Learning

2.1 Q-factors

For every state-action pair, we consider

$$\begin{aligned} Q^*(x, a) &= g_a(x) + \alpha P_a(x_k, y) J^*(y) && (1) \\ J^*(x) &= \min_a Q^*(x, a) && (2) \end{aligned}$$

We can interpret these equations as Bellman's equations for an MDP with expanded state space. We have the original states $x \in \mathcal{S}$, with associated sets of feasible actions $\mathcal{A}_x$, and extra states $(x, a), x \in \mathcal{S}, a \in \mathcal{A}_x$, corresponding to state-action pairs, for which there is only one action available, and no decision must be made. Note that, whenever we are in a state x where a decision must be made, the system transitions deterministically to state (x, a) based on the state and action a chosen. Therefore we circumvent the need to perform expectations $\sum_y P_a(x, y)J(y)$ associated with greedy policies.

We define the operator

$$(HQ)(x, a) = g_a(x) + \alpha \sum_y P_a(x, y) \min_{a'} Q(y, a') \tag{3}$$

It is easy to show that the operator H has the same properties as operator T defined in previous lectures for discounted-cost problems:

Monotonicity	$\forall\, Q$, and $\bar{Q}$ such that $Q \leq \bar{Q}$, $HQ \leq H\bar{Q}$.
Offset	$H(Q + Ke) = HQ + \alpha Ke$.
Contraction	$\|HQ - H\bar{Q}\|_\infty \leq \alpha \|\|Q - \bar{Q}\|\|_\infty$, $\forall\, Q, \bar{Q}$

It follows that H has a unique fixed point, corresponding to the Q factor Q^*.

2.2 Q-Learning

We now develop a real-time value iteration algorithm for computing Q^*. An algorithm analogous to RTVI for computing the cost-to-go function is as follows:

$$Q_{t+1}(x_t, u_t) = g_{u_t}(x_t) + \alpha \sum_y P_{u_t}(x, y) \min_{a'} Q_t(y, a').$$

However, this algorithm undermines the idea that Q-learning is motivated by situations where we do not know $P_a(x, y)$ or find it expensive to compute expectations $\sum_a P_a(x, y) J(y)$. Alternatively, we consider variants that implicitly estimate this expectation, based on state transitions observed in system trajectories. Based on this idea, one possibility is to utilize a scheme of the form

$$Q_{t+1}(x_t, a_t) = g_{a_t}(x_t) + \alpha \min_{a'} Q_t(x_{t+1}, a')$$

However, note that such an algorithm should not be expected to converge; in particular, $Q_t(x_{t+1}, a')$ is a noisy estimate of $\sum_y P_{u_t}(x, y) \min_{a'} Q_t(y, a')$. We consider a *small-step* version of this scheme, where the noise is attenuated:

$$Q_{t+1}(x_t, a_t) = (1 - \gamma_t) Q_t(x_t, a_t) + \gamma_t \left[g_{a_t}(x_t) + \alpha \min_{a'} Q_t(x_{t+1}, a') \right]. \tag{4}$$

We will study the properties of (4) under the more general framework of *stochastic approximations*, which are at the core of many simulation-based or real-time dynamic programming algorithms.

3 Stochastic Approximation

In the stochastic approximation setting, the goal is to solve a system of equations

$$r = Hr,$$

where r is a vector in $\Re^n$ for some n and H is an operator defined in $\Re^n$. If we know how to compute Hr for any given r, it is common to try to solve this sytem of equations by value iteration:

$$r_{k+1} = Hr_k. \tag{5}$$

Now suppose that we cannot compute Hr but have noisy estimates $(Hr + w)$ with $\mathrm{E}[w] = 0$. One alternative is to approximate (5) by drawing several samples $Hr + w_i$ and averaging them, in order to obtain an estimate of Hr. In this case, we would have

$$r_{t+1} = \frac{1}{k} \sum_{i=1}^{k} (Hr_t + w_i)$$

We can also do the summation recursively by setting

$$\begin{aligned} r_t^{(i)} &= \frac{1}{i}\sum_{j=1}^{i}(Hr_t + w_i), \\ r_t^{(i+1)} &= \frac{i}{i+1}r_t^{(i)} + \frac{1}{i+1}(Hr_t + w_{i+1}). \end{aligned}$$

Therefore, $r_{t+1} = r_t^{(k)}$. Finally, we may consider replacing samples $Hr_t + w_i$ with samples $Hr_t^{(i-1)} + w_i$, obtaining the final form

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

A simple application of these ideas involves estimating the expected value of a random variable by drawing i.i.d. samples.

Example 1 *Let $v_1, v_2, \ldots$ be i.i.d. random variables. Given*

$$r_{t+1} = \frac{t}{t+1}r_t + \frac{1}{t+1}v_{t+1}$$

we know that $r_t \to \bar{v}$ by strong law of large numbers. We can actually prove

$$\textit{(General Version)}\ r_{t+1} = (1-\gamma_t)r_t + \gamma_t v_{t+1} \to \bar{v} \ \ \textit{w.p. } 1,$$

if $\sum_{t=1}^{\infty}\gamma_t = \infty$ and $\sum_{t=1}^{\infty}\gamma_t^2 < \infty$.

The conditions on the step sizes γ_t

$$\sum_{t=1}^{\infty}\gamma_t = \infty \tag{6}$$

and

$$\sum_{t=1}^{\infty}\gamma_t^2 < \infty \tag{7}$$

are standard in stochastic approximation algorithms. A simple argument illustrates the need for condition (6): if the total sum of step sizes is finite, iterates r_t are confined to a region around the initial guess r_0, so that, if r_0 is far enough from any solution of $r = Hr$, the algorithm cannot possibly converge. Moreover, since we have noisy estimates of Hr, convergence of $r_{t+1} = (1-\gamma_t)r_t + \gamma H_{r_t} + \gamma_t w$ requires that the noisy term $\gamma_t w$ decreases with time, motivating the condition (7).

We will consider two approaches to analyzing the stochastic approximation algorithm

$$\begin{aligned} r_{t+1} &= (1-\gamma_t)r_t + \gamma_t(Hr_t + w_t) \\ &= r_t + \gamma_t(Hr_t + w_t - r_t) \\ &= r_t + \gamma_t S(r_t, w_t) \end{aligned} \tag{8}$$

where we define $S(r_t, w_t) = Hr_t + w_t - r_t$. The two approaches are

1. Lyapunov function analysis

2. ODE approach

3.1 Lyapunov function analysis

The question we try to answer is "Does (8) converge? If so, where does it converge to?"

We will first illustrate the basic ideas of Lyapunov function analysis by considering a deterministic case.

3.1.1 Deterministic Case

In deterministic case, we have $S(r,w) = S(r)$. Suppose there exists some unique r^* such that

$$S(r^*) = Hr^* - r^* = 0.$$

The basic idea is to show that a certain measure of distance between r_t and r^* is decreasing.

Example 2 *Suppose that F is a contraction with respect to $\|\cdot\|_2$. Then*

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

converges.

Proof: Since F is a contraction, there exists a unique r^* s.t. $Fr^* = r^*$. Let

$$V(r) = \|r - r^*\|_2.$$

We will show $V(r_t) \geq V(r_{t+1})$. Observe

$$\begin{aligned}
V(r_{t+1}) &= \|r_{t+1} - r^*\|_2 \\
&= \|r_t + \gamma_t(Fr_t - r_t) - r^*\|_2 \\
&= \|(1-\gamma_t)(r_t - r^*) + \gamma_t(Fr_t - r^*)\|_2 \\
&\leq (1-\gamma_t)\|r_t - r^*\|_2 + \gamma_t\|Fr_t - r^*\|_2 \\
&\leq (1-\gamma_t)\|r_t - r^*\|_2 + \alpha\gamma_t\|r_t - r^*\|_2 \\
&= \|r_t - r^*\|_2 - (1-\alpha)\gamma_t\|r_t - r^*\|_2.
\end{aligned}$$

Therefore, $\|r_t - r^*\|_2$ is nonincreasing and bounded below by zero. Thus, $\|r_t - r^*\|_2 \xrightarrow{t\to 0} \epsilon \geq 0$. Then

$$\begin{aligned}
0 \leq \|r_{t+1} - r^*\|_2 &\leq \|r_t - r^*\|_2 - (1-\alpha)\gamma_t\|r_t - r^*\|_2 \\
&\leq \|r_t - r^*\|_2 - (1-\alpha)\gamma_t\epsilon \\
&\leq \|r_{t-1} - r^*\|_2 - (1-\alpha)(\gamma_t + \gamma_{t-1})\epsilon \\
&\vdots \\
&\leq \|r_0 - r^*\|_2 - (1-\alpha)\sum_{l=1}^{t}\gamma_l\epsilon
\end{aligned}$$

Hence

$$\epsilon \leq \frac{\|r_0 - r^*\|_2}{(1-\alpha)\sum_{l=1}^{t}\gamma_t}, \quad \forall\, t$$

we thus have $\epsilon = 0$. □

We can isolate several key aspects in the convergence argument used for the example above:

1. We define a "distance" $V(r_t) \geq 0$ indicating how far r_t is from a solution r^* satisfying $S(r) = 0$[1]

2. We show that the distance is "nonincreasing" in t

3. We show that the distance indeed converges to 0.

The argument also involves the basic result that "every nonincreasing sequence bounded below converges" to show that the distance converges

Motivated by these points, we introduce the notion of a *Lyapunov function*:

Definition 1 *We call function V a Lyapunov function if V satisfies*

(a) $V(\cdot) \geq 0$

(b) $(\nabla_r V)^T S(r) \leq 0$

(c) $V(r) = 0 \Leftrightarrow S(r) = 0$

3.1.2 Stochastic Case

The argument used for convergence in the stochastic case parallels the argument used in the deterministic case. Let $\mathcal{F}_t$ denote all information that is available at stage t, and let

$$\bar{S}_t(r) = \mathrm{E}\left[S(r, w_t)|\mathcal{F}_t\right].$$

Then we require a Lyapunov function V satisfying

$$V(\cdot) \geq 0 \tag{9}$$

$$(\nabla V(r_t))^T \bar{S}_t(r_t) \leq -c\|\nabla V(r_t)\|^2 \tag{10}$$

$$\|\nabla V(r) - \nabla V(\bar{r})\| \leq L\|r - \bar{r}\| \tag{11}$$

$$\mathrm{E}\left[S(r_t, w_t)^2|\mathcal{F}_t\right] \leq K_1 + K_2\|\nabla V(r_t)\|^2, \tag{12}$$

for some constants c,L,K_1 and K_2.

Note that (9) and (10) are direct analogues of requiring existence of a distance that is nonincreasing in t; moreover, (10) ensures that the distance decreases at a certain rate if r_t is far from a desired solution r^* satisfying $V(r^* = 0)$. Condition (11) imposes some regularity on V which is required to show that $V(r_t)$ does indeed converge to 0, and condition (12) imposes some control over the noise.

A last point worth mentioning is that (10) implies that the *expected value* of $V(r_t)$ is nonincreasing; however, we may have $V(r_{t+1}) > V(r_t)$ occasionally. Therefore we need an stochastic counterpart to the result that "every nonincreasing sequence bounded below converges." The stochastic counterpart of interest to our analysis is given below.

Theorem 1 (Supermartingale Convergence Theorem) *Suppose that X_t, Y_t and Z_t are nonnegative random variables and $\sum_{t=1}^{\infty} Y_t < \infty$ with probability 1. Suppose also that*

$$\mathrm{E}\left[X_{t+1}\middle|X_i, Y_i, Z_i, i \leq t\right] \leq X_t + Y_t - Z_t$$

Then

[1] $V(r) = \|r - r^*\|_2 \geq 0$ in the above example.

1. X_t *converges to a limit (which can be a random variable) with probability 1,*

2. $\sum_{t=1}^{\infty} Z_t < \infty$.

Theorem 2 *If (9), (10), (11), and (12) are satisfied and we have* $\sum_{t=1}^{\infty} \gamma_t = \infty$ *and* $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$*, then*

1. $V(r_t)$ *converges.*

2. $\lim_{t \to \infty} \nabla V(r_t) = 0$.

3. *Every limit point of* r_t *is a stationary point of* V.