EE 223: Stochastic Estimation and Control **Spring 2007**

Lecture 18 — March 20

Lecturer: Venkat Anantharam *Scribe: Ching-ling Huang*

18.1 Discounted Dynamic Programming

Consider a fully observed dynamical system, with time-invariant state transition function f:

$$x_{k+1} = f(x_k, u_k, w_k), k \geq 0, \tag{18.1}$$

where $(w_k, k \geq 0)$ is an *i.i.d.* sequence of random variables. Let $0 < \alpha < 1$, and try to minimize (informally):

$$minE[\sum_{k=0}^{\infty} \alpha^k g(X_k, U_k, w_k)] \tag{18.2}$$

where $g(\cdot)$ is some given time-invariant one step cost function.

Let us fix a function $J(x)$ on the state space $\mathcal{X}$ and consider the finite horizon problem for N steps followed by paying $\alpha^N J(x)$ at time N. To solve this finite horizon problem, we would set:

$$J_N^{(N)}(x) = \alpha^N J(x), \tag{18.3}$$

$$J_k^{(N)}(x) = min_u E[\alpha^k g(x, u, w) + J_{k+1}^{(N)}(f(x, u, w))] \ , \quad 0 \leq k \leq N-1, \tag{18.4}$$

where the expectation is over the random variable w having the distribution of the system noise random variables. If we consider $\frac{1}{\alpha^k} J_k^{(N)}(x)$, then the DP recursion looks like:

$$\frac{1}{\alpha^k} J_k^{(N)}(x) = min_u E[g(x, u, w) + \alpha \frac{J_{k+1}^{(N)}(f(x, u, w))}{\alpha^{k+1}}] \ , \quad 0 \leq k \leq N-1, \tag{18.5}$$

where the expectation is over the random variable w having the distribution of the system noise random variables. We are thus led to consider the mapping T from functions on $\mathcal{X}$ to functions on $\mathcal{X}$:

$$(TJ)(x) = min_u E[g(x, u, w) + \alpha J(f(x, u, w))]. \tag{18.6}$$

For example, if the evolution is governed by a finite state controlled Markov chain with controlled transition probability matrix $\mathbf{P}(u) = [P_{ij}(u)]$ and costs $g(i, u)$, then

$$(TJ)(i) = min_u [g(i, u) + \alpha \sum_j P_{ij}(u) J(j))]. \tag{18.7}$$

Given any stationary strategy, $\mu : \mathcal{X} \mapsto \mathcal{U}$, we can similarly define a map T_μ, from functions on $\mathcal{X}$ to functions on $\mathcal{X}$:

$$T_\mu J(x) = E[g(x, \mu(x), w) + \alpha J(f(x, \mu(x), w))]. \tag{18.8}$$

We make the basic observation that T and each T_μ are monotone, i.e. if $J_1(x) \leq J_2(x)$ for all x, then $(TJ_1)(x) \leq (TJ_2)(x)$ for all x, and $(T_\mu J_1)(x) \leq (T_\mu J_2)(x)$ for all x.

Let us assume that there is some $0 < M < \infty$ such that

$$|g(x, u, w)| \leq M \ . \tag{18.9}$$

Theorem: There is a function $J^*(x)$ on $\mathcal{X}$ such that for any bounded $J(x)$ on $\mathcal{X}$ (say, $|J(x)| \leq L$ for all x):

$$lim_{N\to\infty}|J_0^{(N)}(x) - J^*(x)| = 0, \textit{uniformly on } x, \tag{18.10}$$

(Here we have made the boundedness assumption on $g(x, u, w)$.) $J^*(x)$ is the optimal cost at the discounted problem. Further, $(TJ^*) = J^*$ and J^* is the unique fixed point of the equation $TJ = J$.

Proof:

Consider applying the following strategy in the infinite horizon discounted problem: Apply the optimal strategy for the "horizon N + terminal cost J" problem at the first N steps and make arbitrary choices of controls from time N onwards. From the initial state x, the cost associated to this strategy is at most

$$J_0^{(N)}(x) + \alpha^N L + \frac{\alpha^N M}{1-\alpha}. \tag{18.11}$$

This expression thus serves as an upper bound for $J^*(x)$, where $J^*(x)$ is defined as the optimal overall cost of the discounted problem from the initial condition x. Conversely, fix $\epsilon > 0$ and consider a strategy for the discounted problem achieving cost at most $J^*(x) + \epsilon$ (initial condition x). Then

$$J_0^{(N)}(x) \leq J^*(x) + \epsilon + \alpha^N L + \frac{\alpha^N M}{1-\alpha}, \tag{18.12}$$

because we could have applied the first N steps of this strategy on the "finite horizon N + terminal cost J" problem. So for all $\epsilon > 0$,

$$lim_{N\to\infty}|J^*(x) - J_0^{(N)}(x)| < \epsilon, \textit{uniformly on } x. \tag{18.13}$$

Hence the limit is 0, uniformly on x. To see that $TJ^* = J^*$, observe that

$$J^*(x) - \alpha^N L - \frac{\alpha^N M}{1-\alpha} \leq J_0^{(N)}(x) \leq J^*(x) + \epsilon + \alpha^N L + \frac{\alpha^N M}{1-\alpha}. \tag{18.14}$$

Apply T, also noting that

$$T(H + c)(x) = min_u E[g(x, u, w) + \alpha(H + c)(f(x, u, w))] = (TH)(x) + \alpha c, \tag{18.15}$$

where H is a function on x, and c is a constant. Hence,

$$(TJ^*)(x) - \alpha(\alpha^N L + \frac{\alpha^N M}{1-\alpha}) \leq (TJ_0^{(N)})(x) \leq (TJ^*)(x) + \alpha(\epsilon + \alpha^N L + \frac{\alpha^N M}{1-\alpha}). \tag{18.16}$$

Thus, letting $N \to \infty$, we conclude that

$$TJ^*(x) \leq J^*(x) \leq TJ^*(x). \tag{18.17}$$

Further, the fixed point equation (which we just saw has J^* as a fixed point),

$$TJ = J \tag{18.18}$$

must have a unique fixed point because any fixed point of it is the overall cost in the discounted problem. Moreover, if $\mu^* : \mathcal{X} \mapsto \mathcal{U}$ is such that

$$(TJ^*)(x) = (T_{\mu^*}J^*)(x), \tag{18.19}$$

i.e. $J^*(x) = E[g(x, \mu^*(x), w) + \alpha J^*(f(x, \mu^*(x), w))]$, then μ^* is a stationary Markov optimal strategy.

18.2 Bellman's Equation

The fixed point equation, $TJ = J$, is called Bellman's equation. The result can be rephrased as: The optimal overall cost is the unique fixed point of Bellman's equation, and any control choice describing this fixed point in terms of itself defines an optimal strategy. There are several ways to solve (in principle) Bellman's equation. We now discuss some of them. Often, however, the state space of the problem is too large to allow these methods to be considered practical.

18.2.1 Value Iteration

For the finite state space controlled Markov chain case, start with some

$$J = \begin{pmatrix} J(1) \\ J(2) \\ \vdots \\ J(d) \end{pmatrix} \tag{18.20}$$

where $|\mathcal{X}| = d$. Here we also assume the control space $\mathcal{U}$ is finite. Then find

$$TJ = \begin{pmatrix} TJ(1) \\ TJ(2) \\ \vdots \\ TJ(d) \end{pmatrix} \tag{18.21}$$

where for each i,

$$TJ(i) = min_u[g(i,u) + \alpha \sum_j P_{ij}(u)J(j)]. \tag{18.22}$$

Let $\mu^{(0)}$ denote the minimizer, then iterate this process (replacing J with TJ) until the successive functions get close enough according to whatever convergence criterion one has.

Claim: The sequence $(T^k J, k \geq 0)$ converges to J^*, which we already proved. If $\mu^{(k)}$ denotes the minimizer in the step from $T^k J$ to $T^{k+1} J$, then there is some K such that $\mu^{(k)}$ are always optimal for $k \geq K$.

To see the truth of this claim, suppose μ is strictly suboptimal, i.e. there is some $1 \leq i \leq d$ with $T_\mu J^*(i) > TJ^*(i) + \Delta$, for some $\Delta > 0$. Now let K_μ is so large that,

$$\left\| T^k J - J^* \right\|_\infty < \frac{\Delta}{3}, \tag{18.23}$$

for all $k \geq K_\mu$. Suppose now that μ shows up infinitely often as one of the $\mu^{(k)}$. For $k > K_\mu$ we can arrive at a contradiction. We write:

$$(T_\mu T^k J)(i) > (T_\mu J^*)(i) - \frac{\Delta}{3} i > (TJ^*)(i) + \frac{2}{3}\Delta = J^*(i) + \frac{2}{3}\Delta \tag{18.24}$$

But we have $T_\mu T^k J = T^{k+1} J$, because μ is a choice for the minimizer $\mu^{(k)}$. Hence

$$(T_\mu T^k J)(i) = (T^{k+1} J)(i) < J^*(i) + \frac{1}{3}\Delta \ , \tag{18.25}$$

which is the desired contradiction.

Since there are only finitely many ($d^{|\mathcal{U}|}$ many) strategies, and each nonoptimal strategy can show up only finitely many times as a minimizer by virtue of this argument, it follows that after a certain stage in value iteration any minimizing strategy is optimal. One could thus find an optimal strategy by running a side loop that occasionally checks if the currently chosen minimizing strategy μ is optimal (by checking if the unique fixed point J_μ^* of T_μ is also a fixed point of T).