IE 3186: Approximate Dynamic Programming **Fall 2018**

Lecture 2: Discounted Infinite Horizon Model

Lecturer: Daniel Jiang *Scribes: Ziyue Sun*

References:

D.P. Bertsekas. *Dynamic Programming and Optimal Control: Approximate Dynamic Programming*, Vol. 2, 4th ed, Athena Scientific, Belmont MA, 2012. (§1.2)

W.B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, 2nd ed, Wiley & Sons, 2007. (Chapter 3)

2.1 Discounted Infinite Horizon Model

This is the most popular model; alternatives are the average cost or stochastic shortest path model. Descriptions of these can be found in the Bertsekas textbook. Infinite number of stages + stationarity lead to elegant algorithms and solutions.

- Only need to learn one value function since in infinite horizon, every time period is now identical.
- Good approximation for any problem with a long horizon even if not truly infinite.

The objective function is:

$$J_\pi(x) = \lim_{N\to\infty} \mathbf{E}\left[\sum_{k=0}^{N} \gamma^k g(x_k, \mu_k(x_k), w_{k+1}) \,\middle|\, x_0 = x\right] \tag{2.1}$$

over policies $\pi = (\mu_0, \mu_1, ...)$. The optimal cost is denoted $J^*(x) = \inf_{\pi\in\Pi} J_\pi(x)$, where Π is the set of all admissible policies.

- State space $\mathcal{X}$.
- Control space $\mathcal{U}(x)$ for $x \in \mathcal{X}$.
- Noise process $\{w_1, w_2, \ldots\}$ i.i.d. across time.

- Transition function $x_{k+1} = f(x_k, u_k, w_{k+1})$ for $x_k \in \mathcal{X},$, $u_k \in \mathcal{U}(x_k)$.
- Bounded costs $|g(x, u, w)| < M$.

Proposition 2.1 (Boundedness of Value Function)**.** *For any policy π, its infinite horizon value is bounded.*

Proof.

$$|J_\pi(x)| \leqslant M + \gamma M + \gamma^2 M + \cdots = \frac{M}{1-\gamma}.$$

Proposition 2.2 (Basic Properties of T, T_μ)**.** *The following hold:*

1. *Monotonicity: Suppose $J(x) \leqslant J^{'}(x)$ for $\forall\, x \in X$. Then,*

$$(T_\mu J)(x) \leqslant (T_\mu J^{'})(x), \;\; (TJ)(x) \leqslant (TJ^{'})(x), \forall\, x \in X.$$

Proof. Consider fixed μ

$$\begin{aligned}(T_\mu J)(x) &= \mathbf{E}\left[g(x, \mu, w) + \gamma J\left(f(x, \mu(x), w)\right]\right. \\ &\leqslant \mathbf{E}\left[g(x, \mu, w) + \gamma J^{'}\left(f(x, \mu(x), w)\right]\right. = (T_\mu J^{'})(x).\end{aligned}$$

Minimize over μ to get the second part. □

2. *Constant shift: For any J, scalar r, and policy μ:*

$$(T_\mu(J + re))(x) = (T_\mu J)(x) + \gamma r, \;\; (T(J + re))(x) = (TJ)(x) + \gamma r.$$

for all x where e is the ones vector. Proof is clear by definition.

Some intuition: in the finite horizon case, we iterate T a total of N times to get optimal value. Analogously, we should expect to now require iterating it ∞ times.

Theorem 2.3 (Convergence of Value Iteration)**.** *For all bounded J,*

$$J^*(x) = (T^k J)(x), \forall\, x \tag{2.2}$$

Proof. Let $J \equiv 0$ for simplicity. For any $\pi = (\mu_0, \mu_1, ...)$, we have

$$\begin{aligned}J_\pi(x) &= \mathbf{E}\left[\sum_{l=0}^{\infty} \gamma^l g(x_k, \mu_k(x_k), w_{k+1})\right] \\ &= \mathbf{E}\left[\sum_{l=0}^{k-1} \gamma^l g(x_l, \mu_l(x_l), w_{l+1})\right] + \mathbf{E}\left[\sum_{l=k}^{\infty} \gamma^l g(x_l, \mu_l(x_l), w_{l+1})\right]\end{aligned}$$

The last term satisfies

$$\left|\mathbf{E}\left[\sum_{l=k}^{\infty}\gamma^l g(x_l,\mu_l(x_l),w_{l+1})\right]\right| \leqslant \frac{\gamma^k M}{1-\gamma}$$

So:

$$J_\pi(x)-\frac{\gamma^k M}{1-\gamma} \leqslant \mathbf{E}\left[\sum_{l=0}^{k-1}\gamma^l g(x_l,\mu_l(x_l),w_{l+1})\right] \leqslant J_\pi(x)+\frac{\gamma^k M}{1-\gamma} \tag{2.3}$$

Since $J\equiv 0$, we can add it everywhere and write:

$$J_\pi(x)-\frac{\gamma^k M}{1-\gamma} \leqslant (T_{\mu_0}T_{\mu_1}\dots T_{\mu_k}J)(x) \leqslant J_\pi(x)+\frac{\gamma^k M}{1-\gamma}.$$

Taking minimum over π and then $k\to\infty$, we get

$$J^*(x) \leqslant \lim_{k\to\infty}(T^kJ)(x) \leqslant J^*(x),$$

completing the proof. □

Theorem 2.4. *(Optimal Policy and Bellman Equation) The optimal value function J^* satisfies $J=TJ$. It is called the "fixed point" of T.*

Proof. Let $J_0\equiv 0$. As before:

$$J^*(x)-\frac{\gamma^k M}{1-\gamma} \leqslant (T^kJ_0)(x) \leqslant J^*(x)-\frac{\gamma^k M}{1-\gamma}$$

By monotonicity and constant shift,

$$\begin{aligned} T\left(J^*(x)-\frac{\gamma^k M}{1-\gamma}\right) &\leqslant (T(T^kJ_0))(x) \leqslant T\left(J^*(x)-\frac{\gamma^k M}{1-\gamma}\right)\\ (TJ^*)(x)-\frac{\gamma^k M}{1-\gamma} &\leqslant (T^{k+1}J_0)(x) \leqslant (TJ^*)(x)+\frac{\gamma^k M}{1-\gamma}. \end{aligned}$$

Let $k\to\infty$ to get $(TJ^*)(x)\leqslant J^*(x)\leqslant (TJ^*)(x)$. □

Proposition 2.5 (Contraction Property)**.** *For any arbitrary value functions J and J' and a policy μ, we have*

$$\left\|T_\mu J - T_\mu J'\right\|_\infty \leqslant \gamma\left\|J-J'\right\|_\infty$$

where $\left\|J\right\|_\infty = \max_x |J(x)|$. *Similarly,*

$$\left\|TJ-TJ'\right\| \leqslant \gamma\left\|J-J'\right\|_\infty$$

Proof. Let $c = \|J - J^{'}\|_\infty$. So we have, for all x:

$$J(x) - c \leqslant J^{'}(x) \leqslant J(x) + c$$
$$(TJ)(x) - \gamma c \leqslant (TJ^{'})(x) \leqslant (TJ)(x) + \gamma c$$

This implies that $|(TJ)(x) - (TJ^{'})(x)| \leqslant \gamma c$, which completes the proof. The T_μ part can be proved by considering a modified problem where the only action available is $\mu(x)$. □

By Banach's fixed point theorem, this contraction property implies a unique J^* satisfying $J^* = TJ^*$.

Corollary 2.6. *$J_\mu(x) = \lim_{N\to\infty} (T_\mu^N J_0)(x)$ and the associated Bellnab equation is $J_\mu = T_\mu J_\mu$, which is uniquely satisfied.*

Theorem 2.7 (Optimal Policy)**.** *A stationary policy μ is optimal iff $\mu(x)$ attains the minimum in Bellman's equation for each x:*

$$TJ^* = T_\mu J^*.$$

Without shorthand notation, this means:

$$\mu(x) \in \arg\min_{u \in \mathcal{U}(x)} \mathbf{E}\left[g(x,u,w) + \gamma J^*(f(x,u,w))\right]$$

Proof. ($\Leftarrow$) If we have $TJ^* = T_\mu J^*$, then using $TJ^* = J^*$, it holds that $J^* = T_\mu J^*$. So by uniqueness of fixed point of T_μ, it must be that $J_\mu = J^*$.

($\Rightarrow$) Assume μ is optimal ($J^* = J_\mu$). We always have $J_\mu = T_\mu J_\mu$, so combining, we get $J^* = T_\mu J^* = TJ^*$, since J^* is also equal to TJ^*. □

2.1.1 Value Iteration Algorithm

1. Start with any J_0.
2. Let $J_{k+1} = TJ_k$.

By the theoren above, $J_{k+1} \to J^*$. Then use μ^* greedy with respect to J^*.

2.1.2 Policy Iteration Algorithm

Main idea: generate a sequence of stationary policies, each with improved cost.

1. Select any policy μ^0.
2. [Policy Evaluation] Compute the value of μ^k, i.e., find J_{μ^k} such that
$$J_{\mu^k} = T_{\mu^k} J_{\mu^k}.$$
3. [Policy Improvement] Find μ^{k+1} such that
$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}.$$
This can also be written as $\mu^{k+1} \leftarrow \arg\max_{u \in U(x)} \mathbf{E}\left[g(x,u,w) + \gamma J_{\mu^k}(f(x,u,w))\right]$.

Theorem 2.8. *(Policy Improvement Theorem) Consider two policies μ and μ' such that $T_{\mu'} J_\mu = T J_\mu$. Then the newer policy μ' is improved: $J_{\mu'} \leqslant J_\mu$. If μ is not optimal, there is strict improvement.*

Proof. Since μ' minimizes with respect to J_μ, we have:
$$\begin{aligned} J_\mu(x) &= \mathbf{E}\left[(g(x,\mu(x),w) + \gamma J_\mu(f(x,\mu(x),w)\right] \\ &\geqslant \mathbf{E}\left[(g(x,\mu'(x),w) + \gamma J_\mu(f(x,\mu'(x),w)\right] \\ &= T_{\mu'} V_\mu. \end{aligned}$$
Applying $T_{\mu'}$ to both sides and using monotonicity, we have:
$$T_{\mu'} J_\mu(x) \geqslant T_{\mu'} T_{\mu'} V_\mu \implies J_\mu \geqslant T_{\mu'} J_\mu \geqslant T_{\mu'} J_\mu \geqslant \cdots \geqslant J_{\mu'}$$
Now suppose there is no improvement, i.e., $J_\mu = J_{\mu'}$. Then
$$J_\mu = T_{\mu'} J_\mu = T J_\mu$$
and by uniqueness, $J_\mu = J^*$. □

2.2 Approximate Dynamic Programming

Approximate dynamic programming attempts to overcome the following three curses of dimensionality: the size of the state space, the size of the action space, and the size of the outcome space (if it is large, then the expectation cannot be computed). The next example illustrates all three issues.

Example 2.9. *Portfolio optimization: there are N stocks and each stock i has an associated Markov price process $\{P_{i,k}\}_{k=0}^{\infty}$. Decisions are how much of each to purchase or sell at each time period and the state variable is*
$$x_k = (r_{1,k}, r_{2,k}, \ldots, r_{N,k}, p_{1,k}, p_{2,k}, \ldots, p_{N,k})^T,$$

where $r_{i,k}$ represents the amount of stock i owned at time k and $p_{i,k}$ is the price of stock i at time k. The action at time k is $u_k = (u_{1,k}, u_{2,k}, \ldots, u_{N,k})^T$, where $u_{i,k} > 0$ means buy the stock and $u_{i,k} < 0$ means sell. The noise is the change in price between periods k and $k+1$: $w_{k+1} = (\hat{p}_{1,k+1}, \hat{p}_{2,k+1}, \ldots, \hat{p}_{N,k+1})^T$. The transitions are

$$\begin{aligned} r_{i,k+1} &= r_{i,k} + u_{i,k}, \\ p_{i,k+1} &= p_{i,k} + \hat{p}_{i,k+1}, \end{aligned}$$

and the cost function is $g(x_k, u_k, w_{k+1}) = \sum_k u_{i,k}\, p_{i,k}$.

Other issues affecting our ability to solve an MDP are continuous states, actions, or distributions; unknown transitions (the central assumption in reinforcement learning) and partial observability (we won't do this).