LECTURE OUTLINE

• We start a nine-lecture sequence on advanced infinite horizon DP and approximate solution methods

• We allow infinite state space, so the stochastic shortest path framework cannot be used any more

• Results are rigorous assuming a countable disturbance space
  – This includes deterministic problems with arbitrary state space, and countable state Markov chains
  – Otherwise the mathematics of measure theory make analysis difficult, although the final results are essentially the same as for countable disturbance space

• The discounted problem is the proper starting point for this analysis

• The central mathematical structure is that the DP mapping is a contraction mapping (instead of existence of a termination state)
DISCOUNTED PROBLEMS/BOUNDED COST

• Stationary system with arbitrary state space

\[ x_{k+1} = f(x_k, u_k, w_k), \quad k = 0, 1, \ldots \]

• Cost of a policy \( \pi = \{\mu_0, \mu_1, \ldots\} \)

\[
J_\pi(x_0) = \lim_{N \to \infty} \mathbb{E}_{w_k} \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}
\]

with \( \alpha < 1 \), and for some \( M \), we have \( |g(x, u, w)| \leq M \) for all \( (x, u, w) \)

• **Shorthand notation for DP mappings** (operate on functions of state to produce other functions)

\[
(TJ)(x) = \min_{u \in \mathcal{U}(x)} \mathbb{E}_w \left\{ g(x, u, w) + \alpha J(f(x, u, w)) \right\}, \forall x
\]

\( TJ \) is the optimal cost function for the one-stage problem with stage cost \( g \) and terminal cost \( \alpha J \).

• For any stationary policy \( \mu \)

\[
(T\mu J)(x) = \mathbb{E}_w \left\{ g(x, \mu(x), w) + \alpha J(f(x, \mu(x), w)) \right\}, \forall x
\]
“SHORTHAND” THEORY – A SUMMARY

• Cost function expressions [with $J_0(x) \equiv 0$]

$$J_{\pi}(x) = \lim_{k \to \infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J_0)(x), \quad J_\mu(x) = \lim_{k \to \infty} (T_{\mu}^k J_0)(x)$$

• Bellman’s equation: $J^* = TJ^*, \quad J_\mu = T_\mu J_\mu$

• Optimality condition:

$$\mu: \text{optimal} \quad \Longleftrightarrow \quad T_\mu J^* = TJ^*$$

• Value iteration: For any (bounded) $J$ and all $x$,

$$J^*(x) = \lim_{k \to \infty} (T^k J)(x)$$

• Policy iteration: Given $\mu^k$,

  – Policy evaluation: Find $J_{\mu^k}$ by solving

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k}$$

  – Policy improvement: Find $\mu^{k+1}$ such that

$$T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k}$$
**TWO KEY PROPERTIES**

- **Monotonicity property:** For any functions $J$ and $J'$ such that $J(x) \leq J'(x)$ for all $x$, and any $\mu$

  \[ (TJ)(x) \leq (TJ')(x), \quad \forall \ x, \]

  \[ (T\mu J)(x) \leq (T\mu J')(x), \quad \forall \ x. \]

- **Additivity property:** For any $J$, any scalar $r$, and any $\mu$

  \[ (T(J + re))(x) = (TJ)(x) + \alpha r, \quad \forall \ x, \]

  \[ (T\mu(J + re))(x) = (T\mu J)(x) + \alpha r, \quad \forall \ x, \]

  where $e$ is the unit function $[e(x) \equiv 1]$. 
CONVERGENCE OF VALUE ITERATION

- If $J_0 \equiv 0$,

$$J^*(x) = \lim_{N \to \infty} (T^N J_0)(x), \quad \text{for all } x$$

Proof: For any initial state $x_0$, and policy $\pi = \{\mu_0, \mu_1, \ldots\}$,

$$J_\pi(x_0) = E \left\{ \sum_{k=0}^{\infty} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}$$

$$= E \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}$$

$$+ E \left\{ \sum_{k=N}^{\infty} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}$$

The tail portion satisfies

$$\left| E \left\{ \sum_{k=N}^{\infty} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\} \right| \leq \frac{\alpha^N M}{1 - \alpha},$$

where $M \geq |g(x, u, w)|$. Take the min over $\pi$ of both sides. Q.E.D.
**BELLMAN’S EQUATION**

- The optimal cost function $J^*$ satisfies Bellman’s Eq., i.e. $J^* = T(J^*)$.

**Proof:** For all $x$ and $N$,

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

where $J_0(x) \equiv 0$ and $M \geq |g(x, u, w)|$. Applying $T$ to this relation, and using Monotonicity and Additivity,

$$(TJ^*)(x) - \frac{\alpha^{N+1} M}{1 - \alpha} \leq (T^{N+1} J_0)(x) \leq (TJ^*)(x) + \frac{\alpha^{N+1} M}{1 - \alpha}$$

Taking the limit as $N \to \infty$ and using the fact

$$\lim_{N \to \infty} (T^{N+1} J_0)(x) = J^*(x)$$

we obtain $J^* = TJ^*$. **Q.E.D.**
THE CONTRACTION PROPERTY

- **Contraction property:** For any bounded functions $J$ and $J'$, and any $\mu$,

\[
\max_x |(TJ)(x) - (TJ')(x)| \leq \alpha \max_x |J(x) - J'(x)|,
\]

\[
\max_x |(T\mu J)(x) - (T\mu J')(x)| \leq \alpha \max_x |J(x) - J'(x)|.
\]

**Proof:** Denote $c = \max_{x \in S} |J(x) - J'(x)|$. Then

\[
J(x) - c \leq J'(x) \leq J(x) + c, \quad \forall \ x
\]

Apply $T$ to both sides, and use the Monotonicity and Additivity properties:

\[
(TJ)(x) - \alpha c \leq (TJ')(x) \leq (TJ)(x) + \alpha c, \quad \forall \ x
\]

Hence

\[
|(TJ)(x) - (TJ')(x)| \leq \alpha c, \quad \forall \ x.
\]

**Q.E.D.**
We can strengthen our earlier result:

Bellman’s equation $J = TJ$ has a unique solution, namely $J^*$, and for any bounded $J$, we have

$$\lim_{k \to \infty} (T^k J)(x) = J^*(x), \quad \forall x$$

Proof: Use

$$\max_x |(T^k J)(x) - J^*(x)| = \max_x |(T^k J)(x) - (T^k J^*)(x)|$$

$$\leq \alpha^k \max_x |J(x) - J^*(x)|$$

Special Case: For each stationary $\mu$, $J_\mu$ is the unique solution of $J = T_\mu J$ and

$$\lim_{k \to \infty} (T^k_\mu J)(x) = J_\mu(x), \quad \forall x,$$

for any bounded $J$.

Convergence rate: For all $k$,

$$\max_x |(T^k J)(x) - J^*(x)| \leq \alpha^k \max_x |J(x) - J^*(x)|$$
NEC. AND SUFFICIENT OPT. CONDITION

• A stationary policy $\mu$ is optimal if and only if $\mu(x)$ attains the minimum in Bellman’s equation for each $x$; i.e.,

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

Proof: If $TJ^* = T_\mu J^*$, then using Bellman’s equation ($J^* = TJ^*$), we have

$$J^* = T_\mu J^*,$$

so by uniqueness of the fixed point of $T_\mu$, we obtain $J^* = J_\mu$; i.e., $\mu$ is optimal.

• Conversely, if the stationary policy $\mu$ is optimal, we have $J^* = J_\mu$, so

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

Combining this with Bellman’s equation ($J^* = TJ^*$), we obtain $TJ^* = T_\mu J^*$. Q.E.D.
COMPUTATIONAL METHODS

• Value iteration and variants
  – Gauss-Seidel version
  – Approximate value iteration

• Policy iteration and variants
  – Combination with value iteration
  – Modified policy iteration
  – Asynchronous policy iteration

• Linear programming

\[
\begin{align*}
\text{maximize} & \quad \sum_{i=1}^{n} J(i) \\
\text{subject to} & \quad J(i) \leq g(i, u) + \alpha \sum_{j=1}^{n} p_{ij}(u) J(j), \quad \forall (i, u)
\end{align*}
\]

• Approximate linear programming: use in place of \( J(i) \) a low-dim. basis function representation

\[
\tilde{J}(i, r) = \sum_{k=1}^{m} r_k w_k(i)
\]

and low-dim. LP (with many constraints)
6.231 Dynamic Programming and Stochastic Control
Fall 2011

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.