LECTURE OUTLINE

- Review of stochastic shortest path problems
- Computational methods for SSP
  - Value iteration
  - Policy iteration
  - Linear programming
- Computational methods for discounted problems
STOCHASTIC SHORTEST PATH PROBLEMS

- Assume finite-state system: States 1, \ldots, n and special cost-free termination state \( t \)
  - Transition probabilities \( p_{ij}(u) \)
  - Control constraints \( u \in U(i) \)
  - Cost of policy \( \pi = \{\mu_0, \mu_1, \ldots\} \) is

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

- Optimal policy if \( J_{\pi}(i) = J^*(i) \) for all \( i \).
- Special notation: For stationary policies \( \pi = \{\mu, \mu, \ldots\} \), we use \( J_{\mu}(i) \) in place of \( J_{\pi}(i) \).

- Assumption (Termination inevitable): There exists integer \( m \) such that for every policy and initial state, there is positive probability that the termination state will be reached after no more than \( m \) stages; for all \( \pi \), we have

\[
\rho_{\pi} = \max_{i=1, \ldots, n} P\{x_m \neq t \mid x_0 = i, \pi \} < 1
\]
MAIN RESULT

• Given any initial conditions $J_0(1), \ldots, J_0(n)$, the sequence $J_k(i)$ generated by value iteration

$$J_{k+1}(i) = \min_{u \in U(i)} \left[ g(i, u) + \sum_{j=1}^{n} p_{ij}(u) J_k(j) \right], \ \forall \ i$$

converges to the optimal cost $J^*(i)$ for each $i$.

• Bellman’s equation has $J^*(i)$ as unique solution:

$$J^*(i) = \min_{u \in U(i)} \left[ g(i, u) + \sum_{j=1}^{n} p_{ij}(u) J^*(j) \right], \ \forall \ i$$

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

• Key proof idea: The “tail” of the cost series,

$$\sum_{k=mK}^{\infty} E \left\{ g(x_k, \mu_k(x_k)) \right\}$$

vanishes as $K$ increases to $\infty$. 

BELLMAN’S EQUATION FOR A SINGLE POLICY

- Consider a stationary policy \( \mu \)
- \( J_\mu(i), \ i = 1, \ldots, n, \) are the unique solution of the linear system of \( n \) equations

\[
J_\mu(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) J_\mu(j), \quad \forall \ i = 1, \ldots, n
\]

- **Proof:** This is just Bellman’s equation for a modified/restricted problem where there is only one policy, the stationary policy \( \mu \), i.e., the control constraint set at state \( i \) is \( \tilde{U}(i) = \{ \mu(i) \} \)
- The equation provides a way to compute \( J_\mu(i), \ i = 1, \ldots, n, \) but the computation is substantial for large \( n \) \( [O(n^3)] \)
- For large \( n \), value iteration may be preferable. (Typical case of a large linear system of equations, where an iterative method may be better than a direct solution method.)
- For VERY large \( n \), exact methods cannot be applied, and approximations are needed. (We will discuss these later.)
POLICY ITERATION

• It generates a sequence $\mu^1, \mu^2, \ldots$ of stationary policies, starting with any stationary policy $\mu^0$.

• At the typical iteration, given $\mu^k$, we perform a policy evaluation step, that computes the $J_{\mu^k}(i)$ as the solution of the (linear) system of equations

$$J(i) = g(i, \mu^k(i)) + \sum_{j=1}^{n} p_{ij}(\mu^k(i))J(j), \quad i = 1, \ldots, n,$$

in the $n$ unknowns $J(1), \ldots, J(n)$. We then perform a policy improvement step, which computes a new policy $\mu^{k+1}$ as

$$\mu^{k+1}(i) = \arg \min_{u \in U(i)} \left[ g(i, u) + \sum_{j=1}^{n} p_{ij}(u)J_{\mu^k}(j) \right], \quad \forall i$$

• The algorithm stops when $J_{\mu^k}(i) = J_{\mu^{k+1}}(i)$ for all $i$

• Note the connection with the rollout algorithm, which is just a single policy iteration
JUSTIFICATION OF POLICY ITERATION

- We can show that $J_{\mu^{k+1}}(i) \leq J_{\mu^k}(i)$ for all $i, k$
- Fix $k$ and consider the sequence generated by

$$J_{N+1}(i) = g(i, \mu^{k+1}(i)) + \sum_{j=1}^{n} p_{ij}(\mu^{k+1}(i)) J_{N}(j)$$

where $J_0(i) = J_{\mu^k}(i)$. We have

$$J_0(i) = g(i, \mu^k(i)) + \sum_{j=1}^{n} p_{ij}(\mu^k(i)) J_0(j)$$

$$\geq g(i, \mu^{k+1}(i)) + \sum_{j=1}^{n} p_{ij}(\mu^{k+1}(i)) J_0(j) = J_1(i)$$

Using the monotonicity property of DP,

$$J_0(i) \geq J_1(i) \geq \cdots \geq J_N(i) \geq J_{N+1}(i) \geq \cdots, \quad \forall i$$

Since $J_N(i) \to J_{\mu^{k+1}}(i)$ as $N \to \infty$, we obtain

$J_{\mu^k}(i) = J_0(i) \geq J_{\mu^{k+1}}(i)$ for all $i$. Also if $J_{\mu^k}(i) = J_{\mu^{k+1}}(i)$ for all $i$, $J_{\mu^k}$ solves Bellman’s equation and is therefore equal to $J^*$

- A policy cannot be repeated, there are finitely many stationary policies, so the algorithm terminates with an optimal policy
• We claim that $J^*$ is the “largest” $J$ that satisfies the constraint

$$J(i) \leq g(i, u) + \sum_{j=1}^{n} p_{ij}(u) J(j),$$

for all $i = 1, \ldots, n$ and $u \in U(i)$.

• **Proof:** If we use value iteration to generate a sequence of vectors $J_k = (J_k(1), \ldots, J_k(n))$ starting with a $J_0$ such that

$$J_0(i) \leq \min_{u \in U(i)} \left[ g(i, u) + \sum_{j=1}^{n} p_{ij}(u) J_0(j) \right], \ \forall \ i$$

Then, $J_k(i) \leq J_{k+1}(i)$ for all $k$ and $i$ (monotonicity property of DP) and $J_k \to J^*$, so that $J_0(i) \leq J^*(i)$ for all $i$.

• So $J^* = (J^*(1), \ldots, J^*(n))$ is the solution of the linear program of maximizing $\sum_{i=1}^{n} J(i)$ subject to the constraint (1).
- Drawback: For large $n$ the dimension of this program is very large. Furthermore, the number of constraints is equal to the number of state-control pairs.
DISCOUNTED PROBLEMS

- Assume a discount factor $\alpha < 1$.
- Conversion to an SSP problem.

- Value iteration converges to $J^*$ for all initial $J_0$:

$$J_{k+1}(i) = \min_{u \in U(i)} \left[ g(i, u) + \alpha \sum_{j=1}^{n} p_{ij}(u)J_k(j) \right], \forall i$$

- $J^*$ is the unique solution of Bellman’s equation:

$$J^*(i) = \min_{u \in U(i)} \left[ g(i, u) + \alpha \sum_{j=1}^{n} p_{ij}(u)J^*(j) \right], \forall i$$
DISCOUNTED PROBLEMS (CONTINUED)

- Policy iteration converges finitely to an optimal policy, and linear programming works.

- **Example:** Asset selling over an infinite horizon. If accepted, the offer $x_k$ of period $k$, is invested at a rate of interest $r$.

- By depreciating the sale amount to period 0 dollars, we view $(1 + r)^{-k}x_k$ as the reward for selling the asset in period $k$ at a price $x_k$, where $r > 0$ is the rate of interest. So the discount factor is $\alpha = 1/(1 + r)$.

- $J^*$ is the unique solution of Bellman’s equation

\[
J^*(x) = \max \left[ x, \frac{E\{J^*(w)\}}{1 + r} \right].
\]

- An optimal policy is to sell if and only if the current offer $x_k$ is greater than or equal to $\bar{\alpha}$, where

\[
\bar{\alpha} = \frac{E\{J^*(w)\}}{1 + r}.
\]
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.