LECTURE OUTLINE

- Undiscounted total cost problems
- Positive and negative cost problems
- Deterministic optimal cost problems
- Adaptive (linear quadratic) DP
- Affine monotonic and risk sensitive problems

Reference:
Updated Chapter 4 of Vol. II of the text: Noncontractive Total Cost Problems
On-line at:
http://web.mit.edu/dimitrib/www/dpchapter.html
Check for most recent version
Infinite horizon total cost DP theory divides in
- “Easy” problems where the results one expects hold (uniqueness of solution of Bellman Eq., convergence of PI and VI, etc)
- “Difficult” problems where one of more of these results do not hold

“Easy” problems are characterized by the presence of strong contraction properties in the associated algorithmic maps $T$ and $T_\mu$

A typical example of an “easy” problem is discounted problems with bounded cost per stage (Chs. 1 and 2 of Voll. II) and some with unbounded cost per stage (Section 1.5 of Voll. II)

Another is semicontractive problems, where $T_\mu$ is a contraction for some $\mu$ but is not for other $\mu$, and assumptions are imposed that exclude the “ill-behaved” $\mu$ from optimality

A typical example is SSP where the improper policies are assumed to have infinite cost for some initial states (Chapter 3 of Vol. II)

In this lecture we go into “difficult” problems
UNDISCOUNTED TOTAL COST PROBLEMS

• Beyond problems with strong contraction properties. One or more of the following hold:
  – No termination state assumed
  – Infinite state and control spaces
  – Either no discounting, or discounting and unbounded cost per stage
  – Risk-sensitivity/exotic cost functions (e.g., SSP problems with exponentiated cost)

• Important classes of problems
  – SSP under weak conditions (e.g., the previous lecture)
  – Positive cost problems (control/regulation, robotics, inventory control)
  – Negative cost problems (maximization of positive rewards - investment, gambling, finance)
  – Deterministic positive cost problems - Adaptive DP
  – A variety of infinite-state problems in queueing, optimal stopping, etc
  – Affine monotonic and risk-sensitive problems (a generalization of SSP)
POS. AND NEG. COST - FORMULATION

• System $x_{k+1} = f(x_k, u_k, w_k)$ and cost

$$J_{\pi}(x_0) = \lim_{N \to \infty} E_{w_k} \left\{ \sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), w_k) \right\}$$

Discount factor $\alpha \in (0, 1]$, but $g$ may be unbounded

• **Case P**: $g(x, u, w) \geq 0$ for all $(x, u, w)$

• **Case N**: $g(x, u, w) \leq 0$ for all $(x, u, w)$

• Summary of analytical results:
  - Many of the strong results for discounted and SSP problems fail
  - Analysis more complex; need to allow for $J_{\pi}$ and $J^*$ to take values $+\infty$ (under P) or $-\infty$ (under N)
  - However, $J^*$ is a solution of Bellman’s Eq. (typically nonunique)
  - Opt. conditions: $\mu$ is optimal if and only if $T_{\mu}J^* = TJ^*$ (P) or if $T_{\mu}J_\mu = TJ_\mu$ (N)
**SUMMARY OF ALGORITHMIC RESULTS**

- Neither VI nor PI are guaranteed to work

- **Behavior of VI**
  - P: $T^k J \rightarrow J^*$ for all $J$ with $0 \leq J \leq J^*$, if $U(x)$ is finite (or compact plus more conditions - see the text)
  - N: $T^k J \rightarrow J^*$ for all $J$ with $J^* \leq J \leq 0$

- **Behavior of PI**
  - P: $J_{\mu_k}$ is monotonically nonincreasing but may get stuck at a nonoptimal policy
  - N: $J_{\mu_k}$ may oscillate (but an optimistic form of PI converges to $J^*$ - see the text)

- These anomalies may be mitigated to a greater or lesser extent by exploiting special structure, e.g.
  - Presence of a termination state
  - Proper/improper policy structure in SSP

- **Finite-state problems under P** can be transformed to equivalent SSP problems by merging (with a simple algorithm) all states $x$ with $J^*(x) = 0$ into a termination state. They can then be solved using the powerful SSP methodology (see updated Ch. 4, Section 4.1.4)
EXAMPLE FROM THE PREVIOUS LECTURE

• This is essentially a shortest path example with termination state $t$

\[ J_{\mu'} = J^* = (0,0) \]
\[ J_{\mu} = (b,0) \]

$J(1) = \min[J(1), b + J(t)]$, \quad $J(t) = J(t)$
DETERM. OPT. CONTROL - FORMULATION

- System: \( x_{k+1} = f(x_k, u_k) \), arbitrary state and control spaces \( X \) and \( U \)
- Cost positivity: \( 0 \leq g(x, u), \forall x \in X, u \in U(x) \)
- No discounting:

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

- “Goal set of states” \( X_0 \)
  - All \( x \in X_0 \) are cost-free and absorbing
- A shortest path-type problem, but with possibly infinite number of states
- A common formulation of control/regulation and planning/robotics problems
- Example: Linear system, quadratic cost (possibly with state and control constraints), \( X_0 = \{0\} \) or \( X_0 \) is a small set around 0
- Strong analytical and computational results
DETERM. OPT. CONTROL - ANALYSIS

- Bellman’s Eq. holds (for not only this problem, but also all deterministic total cost problems)

\[ J^*(x) = \min_{u \in U(x)} \{ g(x, u) + J^*(f(x, u)) \}, \quad \forall \ x \in X \]

- Definition: A policy \( \pi \) terminates starting from \( x \) if the state sequence \( \{x_k\} \) generated starting from \( x_0 = x \) and using \( \pi \) reaches \( X_0 \) in finite time, i.e., satisfies \( x_{\bar{k}} \in X_0 \) for some index \( \bar{k} \)

- Assumptions: The cost structure is such that
  - \( J^*(x) > 0, \ \forall \ x \notin X_0 \) (termination incentive)
  - For every \( x \) with \( J^*(x) < \infty \) and every \( \epsilon > 0 \), there exists a policy \( \pi \) that terminates starting from \( x \) and satisfies \( J_\pi(x) \leq J^*(x) + \epsilon \).

- Uniqueness of solution of Bellman’s Eq.: \( J^* \) is the unique solution within the set

\[ J = \{ J \mid 0 \leq J(x) \leq \infty, \ \forall \ x \in X, \ J(x) = 0, \ \forall \ x \in X_0 \} \]

- Counterexamples: Earlier SP problem. Also linear quadratic problems where the Riccati equation has two solutions (observability not satisfied).
The sequence \( \{ T^k J \} \) generated by VI starting from a \( J \in \mathcal{J} \) with \( J \geq J^* \) converges to \( J^* \).

If in addition \( U(x) \) is finite (or compact plus more conditions - see the text), the sequence \( \{ T^k J \} \) generated by VI starting from any function \( J \in \mathcal{J} \) converges to \( J^* \).

A sequence \( \{ J_{\mu k} \} \) generated by PI satisfies \( J_{\mu k}(x) \downarrow J^*(x) \) for all \( x \in X \).

PI counterexample: The earlier SP example.

Optimistic PI algorithm: Generates pairs \( \{ J_k, \mu^k \} \) as follows: Given \( J_k \), we generate \( \mu^k \) according to

\[
\mu^k(x) = \arg \min_{u \in U(x)} \{ g(x, u) + J_k(f(x, u)) \}, \quad x \in X
\]

and obtain \( J_{k+1} \) with \( m_k \geq 1 \) VIs using \( \mu^k \):

\[
J_{k+1}(x_0) = J_k(x_{m_k}) + \sum_{t=0}^{m_k-1} g(x_t, \mu^k(x_t)), \quad x_0 \in X
\]

If \( J_0 \in \mathcal{J} \) and \( J_0 \geq T J_0 \), we have \( J_k \downarrow J^* \).

Rollout with terminating heuristic (e.g., MPC).
LINEAR-QUADRATIC ADAPTIVE CONTROL

- **System:** \( x_{k+1} = Ax_k + Bu_k, \ x_k \in \mathbb{R}^n, u_k \in \mathbb{R}^m \)
- **Cost:** \( \sum_{k=0}^{\infty} (x_k'Qx_k + u_k'Ru_k), \ Q \geq 0, \ R > 0 \)
- **Optimal policy is linear:** \( \mu^*(x) = Lx \)
- **The Q-factor of each linear policy \( \mu \) is quadratic:**
  \[
  Q_\mu(x, u) = (x' \ u') K_\mu \begin{pmatrix} x \\ u \end{pmatrix} \quad (*)
  \]
- **We will consider \( A \) and \( B \) unknown**
- **We use as basis functions all the quadratic functions involving state and control components**
  \[ x^i x^j, \ u^i u^j, \ x^i u^j, \quad \forall \ i, j \]
  These form the “rows” \( \phi(x, u)' \) of a matrix \( \Phi \)
- **The Q-factor \( Q_\mu \) of a linear policy \( \mu \) can be exactly represented within the subspace spanned by the basis functions:**
  \[
  Q_\mu(x, u) = \phi(x, u)' r_\mu
  \]
  where \( r_\mu \) consists of the components of \( K_\mu \) in (*)
- **Key point:** Compute \( r_\mu \) by simulation of \( \mu \) (Q-factor evaluation by simulation, in a PI scheme)
PI FOR LINEAR-QUADRATIC PROBLEM

- **Policy evaluation**: \( r_\mu \) is found (exactly) by least squares minimization

\[
\min_{\mu} \sum_{(x_k, u_k)} \left| \phi(x_k, u_k)' r - (x_k' Q x_k + u_k' R u_k + \phi(x_{k+1}, \mu(x_{k+1})))' r \right|^2
\]

where \((x_k, u_k, x_{k+1})\) are “enough” samples generated by the system or a simulator of the system.

- **Policy improvement**:

\[
\overline{\mu}(x) \in \arg \min_{u} (\phi(x, u)' r_\mu)
\]

- **Knowledge of \(A\) and \(B\) is not required**

- If the policy evaluation is done exactly, this becomes exact PI, and convergence to an optimal policy can be shown

- The basic idea of this example has been generalized and forms the starting point of the field of adaptive DP

- This field deals with adaptive control of continuous-space (possibly nonlinear) dynamic systems, in both discrete and continuous time
• Generalization of positive cost finite-state stochastic total cost problems where:
  – In place of a transition prob. matrix $P_\mu$, we have a general matrix $A_\mu \geq 0$
  – In place of 0 terminal cost function, we have a more general terminal cost function $\bar{J} \geq 0$

• Mappings

\[
T_\mu J = b_\mu + A_\mu J, \quad (TJ)(i) = \min_{\mu \in \mathcal{M}} (T_\mu J)(i)
\]

• Cost function of $\pi = \{\mu_0, \mu_1, \ldots \}$

\[
J_\pi(i) = \limsup_{N \to \infty} (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(i), \quad i = 1, \ldots, n
\]

• Special case: An SSP with an exponential risk-sensitive cost, where for all $i$ and $u \in U(i)$

\[
A_{ij}(u) = p_{ij}(u)e^{g(i,u,j)}, \quad b(i,u) = p_{it}(u)e^{g(i,u,t)}
\]

• Interpretation:

\[
J_\pi(i) = E\left\{e^{\text{(length of path of } \pi \text{ starting from } i)}\right\}
\]
AFFINE MONOTONIC PROBLEMS: ANALYSIS

- The analysis follows the lines of analysis of SSP
- Key notion (generalizes the notion of a proper policy in SSP): A policy $\mu$ is **stable** if $A_{\mu}^k \to 0$; else it is called **unstable**
- We have

\[
T_{\mu}^N J = A_{\mu}^N J + \sum_{k=0}^{N-1} A_{\mu}^k b_{\mu}, \quad \forall J \in \mathbb{R}^n, \ N = 1, 2, \ldots ,
\]

- For a stable policy $\mu$, we have for all $J \in \mathbb{R}^n$

\[
J_{\mu} = \limsup_{N \to \infty} T_{\mu}^N J = \limsup_{N \to \infty} \sum_{k=0}^{\infty} A_{\mu}^k b_{\mu} = (I - A_{\mu})^{-1} b_{\mu}
\]

- Consider the following assumptions:

1. There exists at least one stable policy
2. For every unstable policy $\mu$, at least one component of $\sum_{k=0}^{\infty} A_{\mu}^k b_{\mu}$ is equal to $\infty$

- Under (1) and (2) the strong SSP analytical and algorithmic theory generalizes
- Under just (1) the weak SSP theory generalizes.