CLASS WILL START AT 5PM

Spring 2020

Lecture 11
Section 5b:
Discounted-Cost Discrete-Time Stochastic Systems: Value Iteration and Policy Iteration
Discounted-Cost Problem Class and the DPE

- Dynamics, admissible feedback policies, discounted-cost infinite time-horizon payoff \((\alpha \in (0, 1))\) and value function
  \[
  \begin{align*}
  \xi_{t+1} &= f(\xi_t, \mu_t(\xi_t), w_t), \\
  \xi_0 &= x \in \mathbb{R}^n, \\
  \mathcal{M}_0 &= \{ \{\mu_t\}_{t=0}^{\infty} \mid \mu_t : \mathbb{R}^n \to U \text{ s.t. } J(x, \mu_t) \in \mathbb{R} \forall t \geq 0 \}, \\
  J(x, \mu_t) &= E|_{\xi_0=x} \left\{ \sum_{t=0}^{\infty} \alpha^t L(\xi_t, \mu_t(\xi_t)) \right\} = E|_{\{w_t\}_{t=0}^{\infty}} \left\{ \sum_{t=0}^{\infty} \alpha^t L(\xi_t, \mu_t(\xi_t)) \right\}, \\
  V(x) &= \inf_{\mu_t \in \mathcal{M}_0} J(x, \mu_t).
  \end{align*}
  \]

- Here, \(\{w_t\} = \{w_t\}_{t=0}^{T-1}\) IID; \(w_t(\omega) \in \mathbb{R}^k; U \subseteq \mathbb{R}^m\).

- Assume
  \[
  \exists D < \infty \text{ s.t. } 0 \leq L(x, v) \leq D \quad \forall x \in \mathbb{R}^n, v \in U.
  \]

- The following will be proved as part of the main theorem.

**Theorem (1)**

*Under Assumption (A.1), the value function is the unique solution of the DPE given by*

\[
V(x) = \inf_{v \in U} \left\{ L(x, v) + \alpha E[V(f(x, v, w))] \right\}.
\]

(DPE)
Two common solution approaches are Value Iteration and Policy Iteration.

Value iteration idea: Write the DPE as \( V(x) = G[V](x) \), and solve this. The actual algorithm is as follows.

1. Guess some \( V_0 \).
2. Given \( V_n \), \( V_{n+1} = G[V_n] \). Repeat...
3. Stop at some convergence criterion, yielding approximate solution, \( V_N \approx \bar{V} \) (where \( \bar{V} = G[\bar{V}] \)) and approximate optimal control \( \bar{u}(x) \approx u_N(x) \in \arg\min_{v \in U} \{ L(x, v) + \alpha E[V_N(f(x, v, w))] \} \).

**Theorem (Banach Fixed-Point Theorem)**

Let \( (\mathcal{Y}, \| \cdot \|) \) be a Banach Space. Suppose \( F : \mathcal{Y} \to \mathcal{Y} \) is a contraction. Then, there exists unique \( \bar{y} \in \mathcal{Y} \) such that \( \bar{y} = F(\bar{y}) \). Further, if \( y_0 \in \mathcal{Y} \) and \( y_{n+1} = F(y_n) \) for all \( n \geq 0 \), then \( y_n \to \bar{y} \).

- We showed that, under some nice conditions, the result implies that Value Iteration converges to the unique true solution of the DPE from any starting guess. (It may take longer if we start far from the solution, of course.)
- Also, if the \( \min \) exists (likely), then the optimal control is 
  \[
  \mu_t^*(x) = \bar{u}(x) \in \arg\min_{v \in U} \{ L(x, v) + \alpha E[V_N(f(x, v, w))] \} \quad (\text{for all times, } t). 
  \]
First Example

- We'll do a simple, one-dimensional problem.
- Dynamics:
  \[
  \xi_{t+1} = \xi_t + u_t + w_t, \quad \xi_0 = x \in [0, \infty),
  \]
  \[
  U = \{-1, 0\},
  \]
  \[
  L(x, v) = \begin{cases} 
  e^x & \text{if } v = -1, \\
  0.5e^x & \text{if } v = 0.
  \end{cases}
  \]
- \{w_t\} IID, with uniform density over [0, 1).
- Payoff and value:
  \[
  J(x, \{\mu_t\}_{t=0}^\infty) = \mathbb{E}\left[\xi_0 = x \left[\sum_{t=0}^\infty \alpha^t L(\xi_t, \mu_t(\xi_t))\right]\right],
  \]
  \[
  V(x) = \inf_{\{\mu_t\}_{t=0}^\infty \in \mathcal{M}_0^f} J(x, \{\mu_t\}_{t=0}^\infty).
  \]
- Do a couple steps.
Consider the discounted-cost, infinite time horizon, linear quadratic Gaussian regulator.

Dynamics, payoff and value:

\[ \xi_{t+1} = A \xi_t + B \mu_t(\xi_t) + \sigma w_t, \quad \xi_0 = x \in \mathbb{R}^n, \quad (D) \]

\[ J(x, \{\mu_t(\cdot)\}) = \mathbb{E} \left|_{\xi_0=x} \left\{ \sum_{t=0}^{\infty} \alpha^t L(\xi_t, \mu_t(\xi_t)) \right\} \right|, \quad (P) \]

where \( L(x, v) = \frac{1}{2} [x^T C x + v^T D v] \),

\[ V(x) = \inf_{\{\mu_t(\cdot)\} \in \tilde{M}_0} J(x, \{\mu_t(\cdot)\}). \quad (V) \]

\( \mu_t(x) \in \mathbb{R}^m \) for all \( t \); \( \{w_t\} \) are IID with \( w_t \sim \mathcal{N}(0, Q) \) with values in \( \mathbb{R}^k \) for all \( t \).

Actually, we’ll get \( \mu_t(x) = \bar{u}(x) \) for all \( t \), and likely, we’ll have \( \bar{u}(x) = Kx \) for some \( K \).

DPE:

\[ V(x) = \inf_{v \in \mathbb{R}^k} \left\{ L(x, v) + \alpha \mathbb{E}[V(f(x, v, w))] \right\} \]

\[ = \inf_{v \in \mathbb{R}^k} \left\{ \frac{1}{2} [x^T C x + v^T D v] + \alpha \mathbb{E}[V(Ax + Bv + \sigma w)] \right\}. \]

Do some of this example.
Convergence Issue

- Note 1: Convergence of the control can occur much sooner than that of the value.
- Note 2: Actually, generally do this on a grid over space. We'll delay that sort of thing to the continuous-time/continuous-space deterministic problem class.
The second method for discounted-cost problems is **Policy Iteration**.

Suppose you have some (time-independent) feedback control policy, \( \mu : \mathcal{X} \rightarrow \mathcal{U} \). Let

\[
G^\mu [V](x) \doteq L(x, \mu(x)) + \alpha E[V(f(x, \mu(x), w))].
\]

\( G^\mu \) is an operator. (DPE if we only had one option for the controller.)

**Idea:**

- Guess some time-independent f/b policy, say \( \mu^0 : \mathcal{X} \rightarrow \mathcal{U} \). (The “zero” superscript is iteration number - not time!)
- Let \( V^0 \) be the solution of \( V^0 = G^{\mu^0} [V^0] \).
- Then, for \( x \in \mathcal{X} \), let

\[
\mu^1(x) \in \text{argmin} \, v \in \mathcal{U} \{ L(x, v) + \alpha E[V^0(f(x, v, w))] \}.
\]

- Let \( V^1 \) be the solution of \( V^1 = G^{\mu^1} [V^1] \).
- Repeat...
Policy Iteration

- Policy Iteration has two sub-steps per step.

1. Policy “improvement”: For each $x \in \mathcal{X}$, let
   \[ \mu^{n+1}(x) \in \arg\min_{v \in U} \{ L(x, v) + \alpha E[V^n(f(x, v, w))] \}. \]  
   (1)

2. Value computation:
   \[ V^{n+1}(x) = G_{\mu^{n+1}}[V^{n+1}](x) \text{ for all } x \in \mathcal{X}. \]  
   (2)

- Regarding the second step, suppose there was only one allowable control policy, $\mu^{n+1}$. Then, (2) is the DPE for
   \[ V(x) = \inf_{\mu \in \{\mu^{n+1}\}} E \left\{ \sum_{t=0}^{\infty} \alpha^t L(\xi_t, \mu^{n+1}(\xi_t)) \right\}. \]

- Consequently, by our value-iteration/BFPTTh results, there is a unique solution of (2).