6.231 DYNAMIC PROGRAMMING

LECTURE 21

LECTURE OUTLINE

- With this lecture, we start a four-lecture sequence on advanced dynamic programming and neuro-dynamic programming topics. References:
  – Dynamic Programming and Optimal Control, Vol. II, by D. Bertsekas
  – Neuro-Dynamic Programming, by D. Bertsekas and J. Tsitsiklis
- **1st Lecture:** Discounted problems with infinite state space, stochastic shortest path problem
- **2nd Lecture:** DP with cost function approximation
- **3rd Lecture:** Simulation-based policy and value iteration, temporal difference methods
- **4th Lecture:** Other approximation methods: Q-learning, state aggregation, approximate linear programming, approximation in policy space

DISCOUNTED PROBLEMS W/ BOUNDED COST

- System

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k = 0, 1, \ldots,$$

- Cost of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$

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

with $g(x, u, w)$: bounded over (x, u, w), and $\alpha < 1$.

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

$$(TJ)(x) = \min_{u \in U(x)} \mathop{E}_{w} \Big\{ g(x, u, w) + \alpha J\big(f(x, u, w)\big) \Big\}, \ \forall \, x$$

TJ is the optimal cost function for the one-stage problem with stage cost g and terminal cost αJ.

- For any stationary policy μ

$$(T_\mu J)(x) = \mathop{E}_{w} \Big\{ g\big(x, \mu(x), w\big) + \alpha J\big(f(x, \mu(x), w)\big) \Big\}, \ \forall \, x$$

“SHORTHAND” THEORY

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

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

- Bellman’s equation: $J^* = TJ^*$, $J_\mu = T_\mu J_\mu$
- Optimality condition:

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

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

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

- Policy iteration steps: Given μ^k,
  - Policy evaluation: Find J_{μ^k} by solving

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

  - Policy improvement: Find μ^{k+1} such that

$$T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k}$$

THE THREE KEY PROPERTIES

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

$$(TJ)(x) \le (TJ')(x), \qquad \forall\, x,$$

$$(T_\mu J)(x) \le (T_\mu J')(x), \qquad \forall\, x$$

- **Additivity property:** For any J, any scalar r, and any μ

$$\big(T(J+re)\big)(x) = (TJ)(x) + \alpha r, \qquad \forall\, x,$$

$$\big(T_\mu(J+re)\big)(x) = (T_\mu J)(x) + \alpha r, \qquad \forall\, x,$$

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

- **Contraction property:** For any (bounded) functions J and J', and any μ,

$$\max_x \big|(TJ)(x) - (TJ')(x)\big| \le \alpha \max_x \big|J(x) - J'(x)\big|,$$

$$\max_x \big|(T_\mu J)(x) - (T_\mu J')(x)\big| \le \alpha \max_x \big|J(x) - J'(x)\big|.$$

"SHORTHAND" ANALYSIS

- **Contraction mapping theorem:** The contraction property implies that:
  - T has a unique fixed point, J^*, which is the limit of T^kJ for any (bounded) J.
  - For each μ, T_μ has a unique fixed point, J_μ, which is the limit of T_μ^kJ for any J.

- **Convergence rate:** For all k,

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

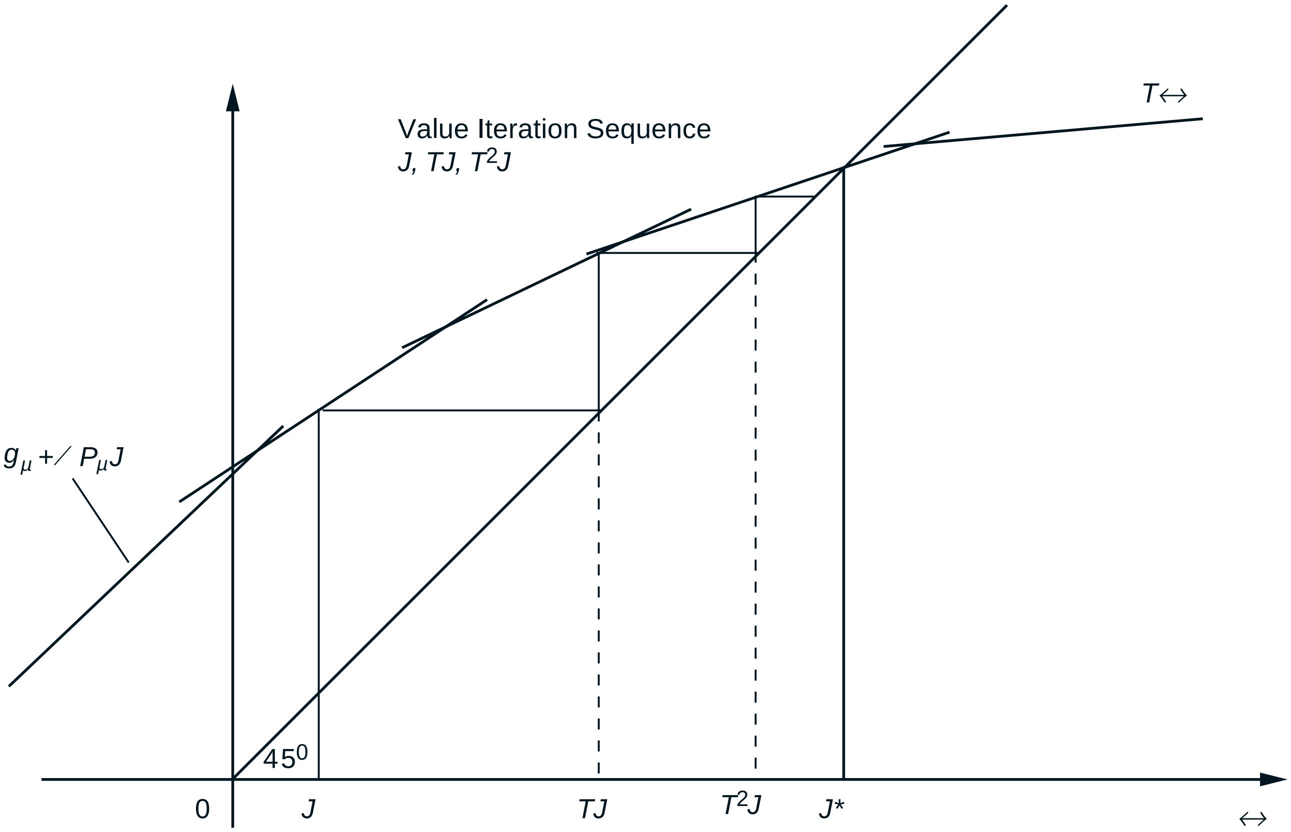
- An assortment of other analytical and computational results are based on the contraction property, e.g, error bounds, computational enhancements, etc.

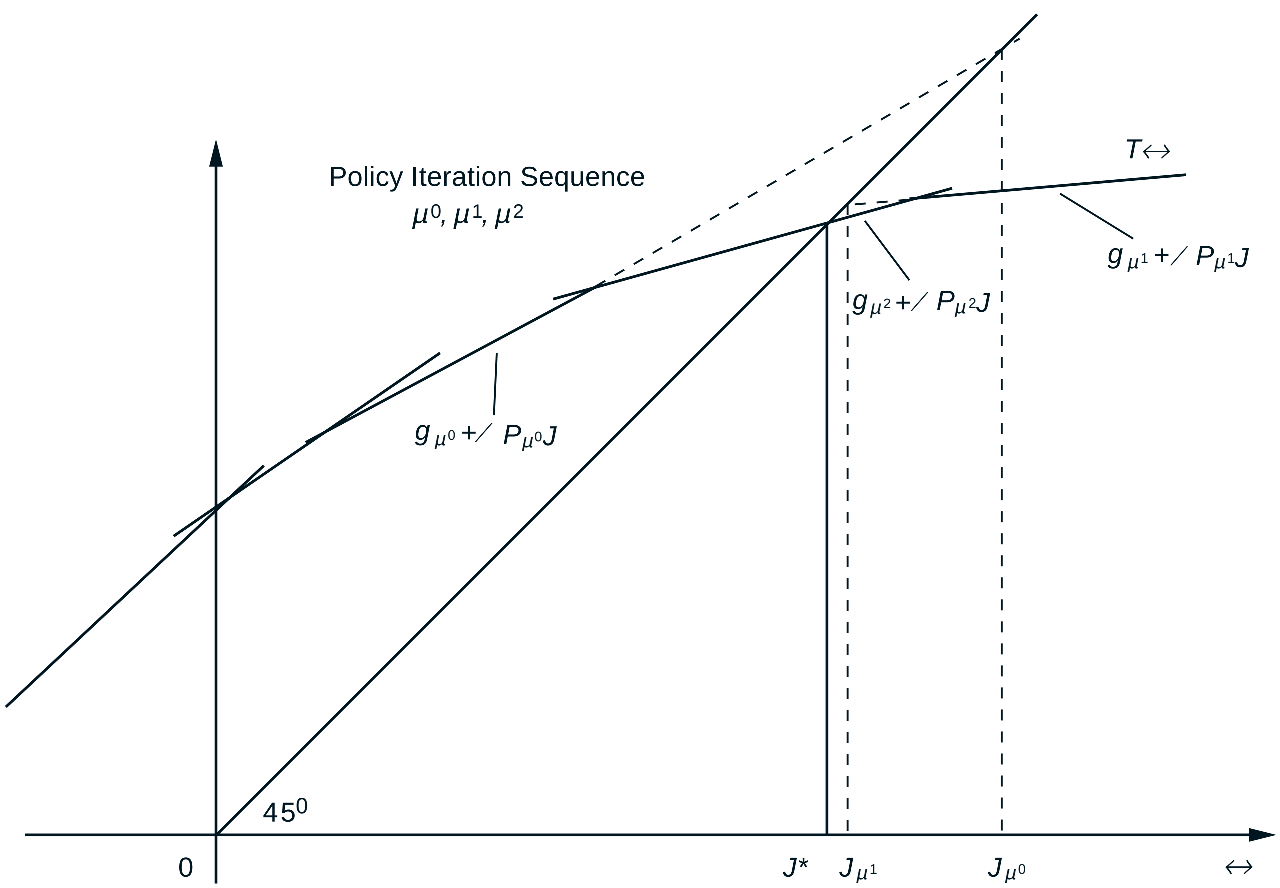
- Example: If we execute value iteration *approximately*, so we compute TJ within an ϵ-error, i.e.,

$$\max_x |\tilde{J}(x) - (TJ)(x)| \leq \epsilon,$$

in the limit we obtain J^* within an $\epsilon/(1-\alpha)$ error.

GEOMETRIC INTERPRETATIONS





UNDISCOUNTED PROBLEMS

- System

$$x_{k+1} = f(x_k, u_k, w_k), \qquad k = 0, 1, \ldots,$$

- Cost of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$

$$J_\pi(x_0) = \lim_{N \to \infty} \mathop{E}_{\substack{w_k \\ k=0,1,\ldots}} \left\{ \sum_{k=0}^{N-1} g\big(x_k, \mu_k(x_k), w_k\big) \right\}$$

- Shorthand notation for DP mappings

$$(TJ)(x) = \min_{u \in U(x)} \mathop{E}_{w} \Big\{ g(x, u, w) + J\big(f(x, u, w)\big) \Big\}, \ \forall\, x$$

- For any stationary policy μ

$$(T_\mu J)(x) = \mathop{E}_{w} \Big\{ g\big(x, \mu(x), w\big) + J\big(f(x, \mu(x), w)\big) \Big\}, \ \forall\, x$$

- Neither T nor T_μ are contractions in general. Some, but not all, of the nice theory holds, thanks to the monotonicity of T and T_μ.

- Some of the nice theory is recovered in SSP problems because of the termination state.

STOCHASTIC SHORTEST PATH PROBLEMS I

- Assume: Cost-free term. state t, a finite number of states $1, \ldots, n$, and finite number of controls
- Mappings T and T_μ (modified to account for termination state t):

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

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

- **Definition:** A stationary policy μ is called **proper**, if under μ, from every state i, there is a positive probability path that leads to t.
- Important fact: If μ is proper then T_μ is a contraction with respect to some weighted max norm

$$\max_i \frac{1}{v_i} |(T_\mu J)(i) - (T_\mu J')(i)| \le \alpha \max_i \frac{1}{v_i} |J(i) - J'(i)|$$

- If all μ are proper, then T is similarly a contraction (the case discussed in the text, Ch. 7).

STOCHASTIC SHORTEST PATH PROBLEMS II

- The theory can be pushed one step further. Assume that:
  - (a) There exists at least one proper policy
  - (b) For each improper μ, $T_\mu(i) = \infty$ for some i
- Then T is not necessarily a contraction, but:
  - J^* is the unique solution of Bellman's Equ.
  - μ^* is optimal if and only if $T_{\mu^*} J^* = T J^*$
  - $\lim_{k\to\infty}(T^k J)(i) = J^*(i)$ for all i
  - Policy iteration terminates with an optimal policy, if started with a proper policy
- **Example:** Deterministic shortest path problem with a single destination
  - States $<=>$ nodes; Controls $<=>$ arcs
  - Termination state $<=>$ the destination
  - Assumption (a) $<=>$ every node is connected to the destination
  - Assumption (b) $<=>$ all cycle costs > 0
  - Pathology: If there is a cycle cost $= 0$ (or < 0), Bellman's equation has an infinite number of solutions (no solution, respectively)

PATHOLOGIES: THE BLACKMAILER'S DILEMMA

- Two states, state 1 and the termination state t.
- At state 1, choose a control $u \in (0,1]$ (the blackmail amount demanded), and move to t at no cost with probability u^2, or stay in 1 at a cost $-u$ with probability $1-u^2$.
- Every stationary policy is proper, but the control set in not finite.
- For any stationary μ with $\mu(1) = u$, we have

$$J_\mu(1) = -(1-u^2)u + (1-u^2)J_\mu(1)$$

from which $J_\mu(1) = -\frac{1-u^2}{u}$

- Thus $J^*(1) = -\infty$, and there is no optimal stationary policy.
- It turns out that a *nonstationary* policy is optimal: demand $\mu_k(1) = \gamma/(k+1)$ at time k, with $\gamma \in (0, 1/2)$. (Blackmailer requests diminishing amounts over time, which add to ∞; the probability of the victim's refusal diminishes at a much faster rate.)