6.231 DYNAMIC PROGRAMMING

LECTURE 22

LECTURE OUTLINE

- Approximate DP for large/intractable problems
- Approximate policy iteration
- Simulation-based policy iteration
- Actor-critic interpretation
- Learning how to play tetris: A case study
- Approximate value iteration with function approximation

APPROX. POLICY ITERATION - DISCOUNTED CASE

- Suppose that the policy evaluation is approximate, according to,

$$\max_x |J_k(x) - J_{\mu^k}(x)| \leq \delta, \qquad k = 0, 1, \ldots$$

and policy improvement is also approximate, according to,

$$\max_x |(T_{\mu^{k+1}} J_k)(x) - (T J_k)(x)| \leq \epsilon, \qquad k = 0, 1, \ldots$$

where δ and ϵ are some positive scalars.

- **Error Bound:** The sequence $\{\mu^k\}$ generated by the approximate policy iteration algorithm satisfies

$$\limsup_{k \to \infty} \max_{x \in S} \big(J_{\mu^k}(x) - J^*(x) \big) \leq \frac{\epsilon + 2\alpha\delta}{(1 - \alpha)^2}$$

- Typical practical behavior: The method makes steady progress up to a point and then the iterates J_{μ^k} oscillate within a neighborhood of J^*.

APPROXIMATE POLICY ITERATION - SSP

- Suppose that the policy evaluation is approximate, according to,

$$\max_{i=1,\ldots,n} |J_k(i) - J_{\mu^k}(i)| \leq \delta, \qquad k = 0, 1, \ldots$$

and policy improvement is also approximate, according to,

$$\max_{i=1,\ldots,n} |(T_{\mu^{k+1}} J_k)(i) - (T J_k)(i)| \leq \epsilon, \qquad k = 0, 1, \ldots$$

where δ and ϵ are some positive scalars.

- Assume that all policies generated by the method are proper (they are guaranteed to be if $\delta = \epsilon = 0$, but not in general).

- **Error Bound:** The sequence $\{\mu^k\}$ generated by approximate policy iteration satisfies

$$\limsup_{k\to\infty} \max_{i=1,\ldots,n} \big(J_{\mu^k}(i) - J^*(i)\big) \leq \frac{n(1-\rho+n)(\epsilon+2\delta)}{(1-\rho)^2}$$

where $\rho = \max_{\substack{i=1,\ldots,n \\ \mu:\ \text{proper}}} P\{x_n \neq t \,|\, x_0 = i, \mu\}$

SIMULATION-BASED POLICY EVALUATION

- Given μ, suppose we want to calculate J_μ by simulation.
- Generate by simulation sample costs. Approximation:

$$J_\mu(i) \approx \frac{1}{M_i} \sum_{m=1}^{M_i} c(i,m)$$

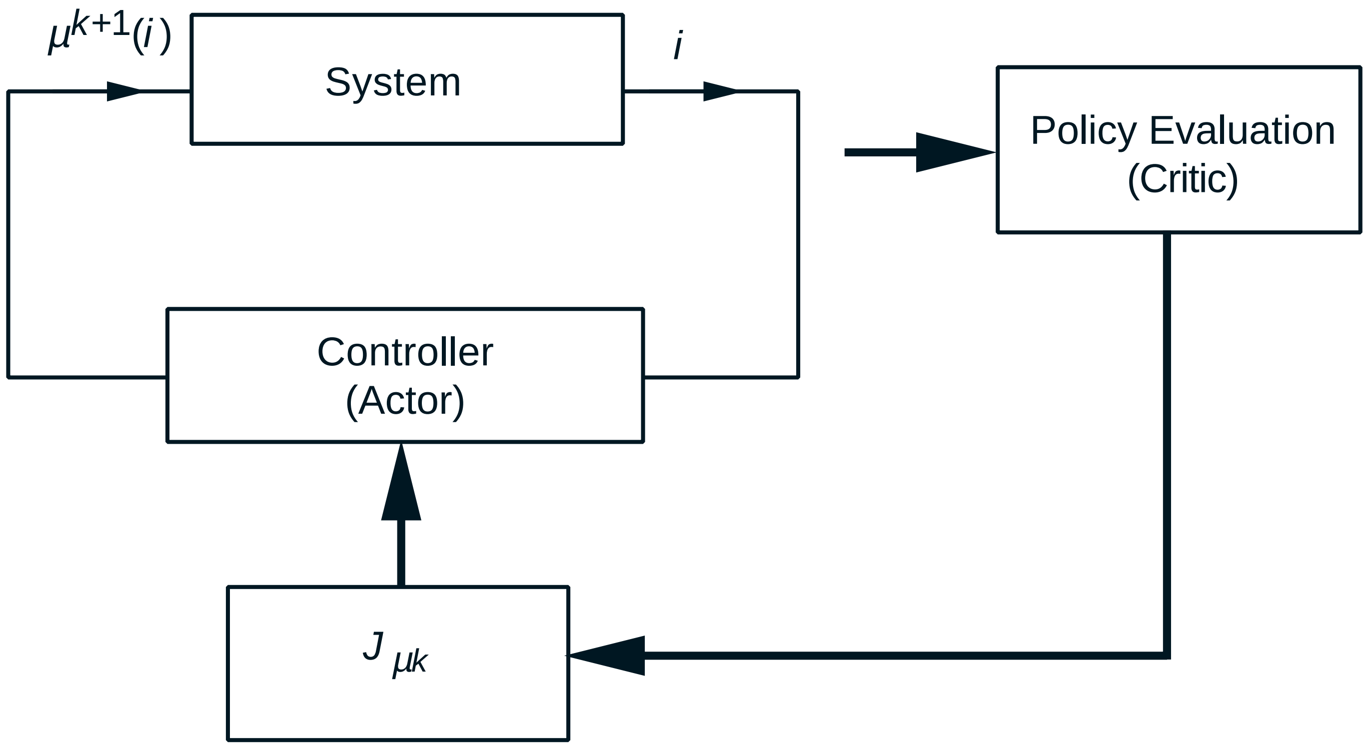
$c(i,m)$: mth sample cost starting from state i

- Approximating each $J_\mu(i)$ is impractical for a large state space. Instead, a "compact representation" $\tilde{J}_\mu(i,r)$ may be used, where r is a tunable parameter vector. We may calculate an optimal value r^* of r by a least squares fit

$$r^* = \arg\min_r \sum_{i=1}^{n} \sum_{m=1}^{M_i} \left| c(i,m) - \tilde{J}_\mu(i,r) \right|^2$$

- This idea is the starting point for more sophisticated simulation-related methods, to be discussed in the next lecture.

ACTOR-CRITIC INTERPRETATION

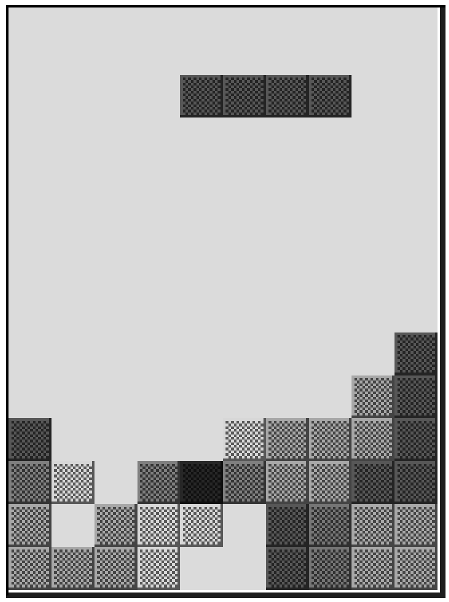


- The critic calculates approximately (e.g., using some form of a least squares fit) J_{μ^k} by processing state/sample cost pairs, which are generated by the actor by simulation

- Given the approximate J_{μ^k}, the actor implements the improved policy $J_{\mu^{k+1}}$ by

$$(T_{\mu^{k+1}} J_k)(i) = (T J_k)(i)$$

EXAMPLE: TETRIS I



- The state consists of the board position i, and the shape of the current falling block (astronomically large number of states).
- It can be shown that all policies are proper!!
- Use a linear approximation architecture with feature extraction

$$\tilde{J}(i, r) = \sum_{m=1}^{s} \phi_m(i) r_m,$$

where $r = (r_1, \ldots, r_s)$ is the parameter vector and $\phi_m(i)$ is the value of mth feature associated w/ i.

EXAMPLE: TETRIS II

- Approximate policy iteration was implemented with the following features:
  - The height of each column of the wall
  - The difference of heights of adjacent columns
  - The maximum height over all wall columns
  - The number of “holes” on the wall
  - The number 1 (provides a constant offset)

- Playing data was collected for a fixed value of the parameter vector r (and the corresponding policy); the policy was approximately evaluated by choosing r to match the playing data in some least-squares sense.

- The method used for approximate policy evaluation was the λ*-least squares policy evaluation method*, to be described in the next lecture.

- See: Bertsekas and Ioffe, “Temporal Differences-Based Policy Iteration and Applications in Neuro-Dynamic Programming,” in

http://www.mit.edu:8001//people/dimitrib/publ.html

VALUE ITERATION W/ FUNCTION APPROXIMATION

- Suppose we use a linear approximation architecture $\tilde{J}(i,r) = \phi(i)'r$, or

$$\tilde{J} = \Phi r$$

where $r = (r_1, \ldots, r_s)$ is a parameter vector, and Φ is a full rank $n \times s$ feature matrix.

- **Approximate value iteration method:** Start with initial guess r_0; given r_t, generate r_{t+1} by

$$r_{t+1} = \arg\min_r \left\| \Phi r - T(\Phi r_t) \right\|$$

where $\|\cdot\|$ is some norm.

- Questions: Does r_t converge to some r^*? How close is Φr^* to J^*?

- **Convergence Result:** If T is a contraction with respect to a weighted Euclidean norm ($\|J\|^2 = J'DJ$, where D is positive definite, symmetric), then r_t converges to (the unique) r^* satisfying

$$r^* = \arg\min_r \left\| \Phi r - T(\Phi r^*) \right\|$$

GEOMETRIC INTERPRETATION

- Consider the feature subspace

$$S = \{\Phi r \mid r \in \Re^s\}$$

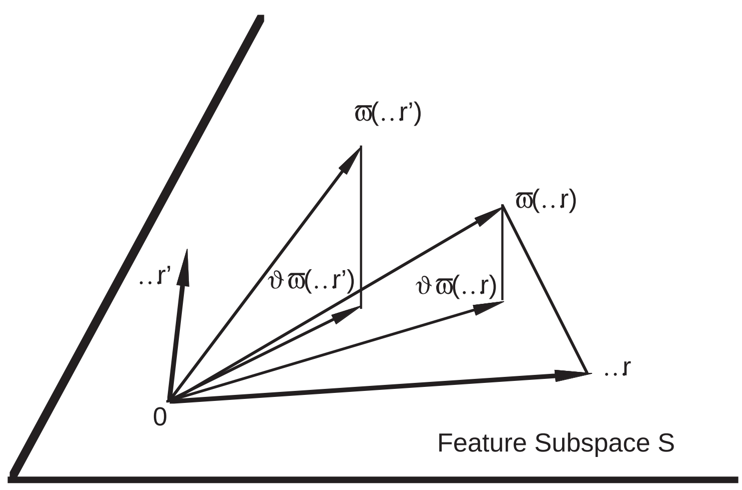
of all cost function approximations that are linear combinations of the feature vectors. Let Π denote projection on this subspace.

- The approximate value iteration is

$$r_{t+1} = \Pi T(\Phi r_t) = \arg\min_r \left\| \Phi r - T(\Phi r_t) \right\|$$

and amounts to starting at the point Φr_t of S applying T to it and then projecting on S.

- **Proof Idea:** Since T is a contraction with respect to the norm of projection, and projection is nonexpansive, ΠT (which maps S to S) is a contraction (with respect to the same norm).



PROOF

- Consider two vectors Φr and $\Phi r'$ in S. The (Euclidean) projection is a nonexpansive mapping, so

$$\|\Pi T(\Phi r) - \Pi T(\Phi r')\| \leq \|T(\Phi r) - T(\Phi r')\|$$

Since T is a contraction mapping (with respect to the norm of projection),

$$\|T(\Phi r) - T(\Phi r')\| \leq \beta \|\Phi r - \Phi r'\|$$

where $\beta \in (0, 1)$ is the contraction modulus, so

$$\|\Pi T(\Phi r) - \Pi T(\Phi r')\| \leq \beta \|\Phi r - \Phi r'\|$$

and it follows that ΠT is a contraction (with respect to the same norm and with the same modulus).

- In general, it is not clear how to obtain a Euclidean norm for which T is a contraction.

- **Important fact:** In the case where $T = T_\mu$, where μ is a stationary policy, T is a contraction for the norm $\|J\|^2 = J'DJ$, where D is diagonal with the steady-state probabilities along the diagonal.

ERROR BOUND

- If T is a contraction with respect to a weighted Euclidean norm $\|\cdot\|$ with modulus β, and r^* is the limit of r_t, i.e.,

$$r^* = \arg\min_r \big\|\Phi r - T(\Phi r^*)\big\|$$

then

$$\|\Phi r^* - J^*\| \leq \frac{\|\Pi J^* - J^*\|}{1-\beta}$$

where J^* is the fixed point of T, and ΠJ^* is the projection of J^* on the feature subspace S (with respect to norm $\|\cdot\|$).

Proof: Using the triangle inequality,

$$\begin{aligned}
\|\Phi r^* - J^*\| &\leq \|\Phi r^* - \Pi J^*\| + \|\Pi J^* - J^*\| \\
&= \|\Pi T(\Phi r^*) - \Pi T(J^*)\| + \|\Pi J^* - J^*\| \\
&\leq \beta\|\Phi r^* - J^*\| + \|\Pi J^* - J^*\| \qquad \text{Q.E.D.}
\end{aligned}$$

- Note that the error $\|\Phi r^* - J^*\|$ is proportional to $\|\Pi J^* - J^*\|$, which can be viewed as the "power of the approximation architecture" (measures how well J^* can be represented by the chosen features).