6.231 Dynamic Programming and Stochastic Control
Fall 2008

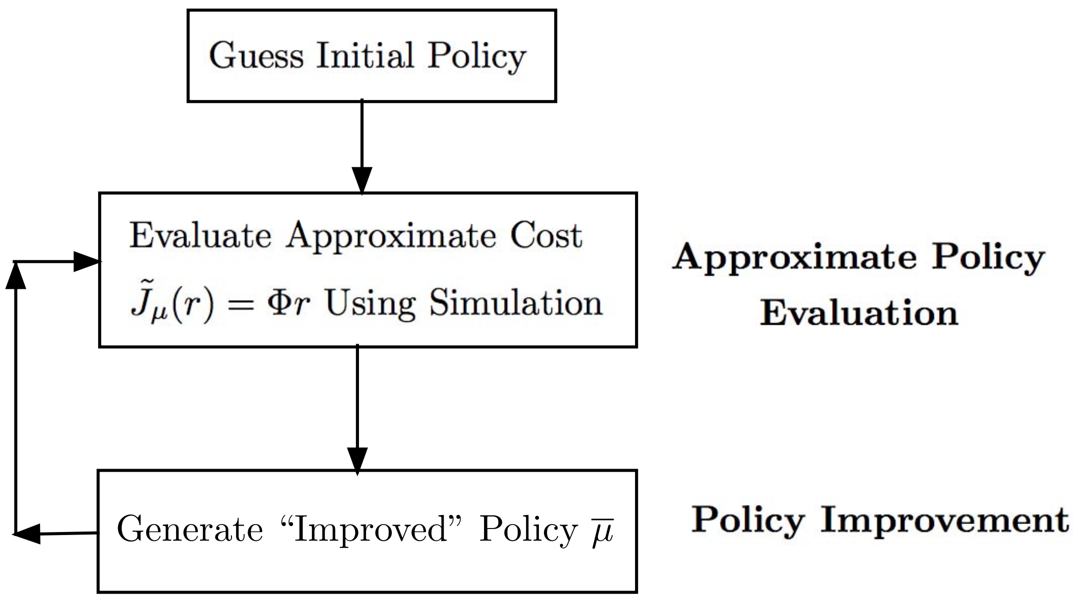
6.231 DYNAMIC PROGRAMMING

LECTURE 22

LECTURE OUTLINE

- Discounted problems - Approximate policy evaluation/policy improvement
- Indirect approach - The projected equation
- Contraction properties - Error bounds
- PVI (Projected Value Iteration)
- LSPE (Least Squares Policy Evaluation)
- Tetris - A case study

POLICY EVALUATION/POLICY IMPROVEMENT



- Linear cost function approximation

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

where Φ is full rank $n \times s$ matrix with columns the basis functions, and ith row denoted $\phi(i)'$.

- Policy "improvement"

$$\overline{\mu}(i) = \arg \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) \big(g(i,u,j) + \alpha \phi(j)' r \big)$$

- Indirect methods find Φr by solving a projected equation.

WEIGHTED EUCLIDEAN PROJECTIONS

- Consider a weighted Euclidean norm

$$\|J\|_v = \sqrt{\sum_{i=1}^{n} v_i \big(J(i)\big)^2},$$

where v is a vector of positive weights $v_1, \ldots, v_n$.

- Let Π denote the projection operation onto

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

with respect to this norm, i.e., for any $J \in \Re^n$,

$$\Pi J = \Phi r_J$$

where

$$r_J = \arg\min_{r \in \Re^s} \|J - \Phi r\|_v$$

- Π and r_J can be written explicitly:

$$\Pi = \Phi(\Phi' V \Phi)^{-1} \Phi' V, \qquad r_J = (\Phi' V \Phi)^{-1} \Phi' V J,$$

where V is the diagonal matrix with v_i, $i = 1, \ldots, n$, along the diagonal.

THE PROJECTED BELLMAN EQUATION

- For a fixed policy μ to be evaluated, consider the corresponding mapping T:

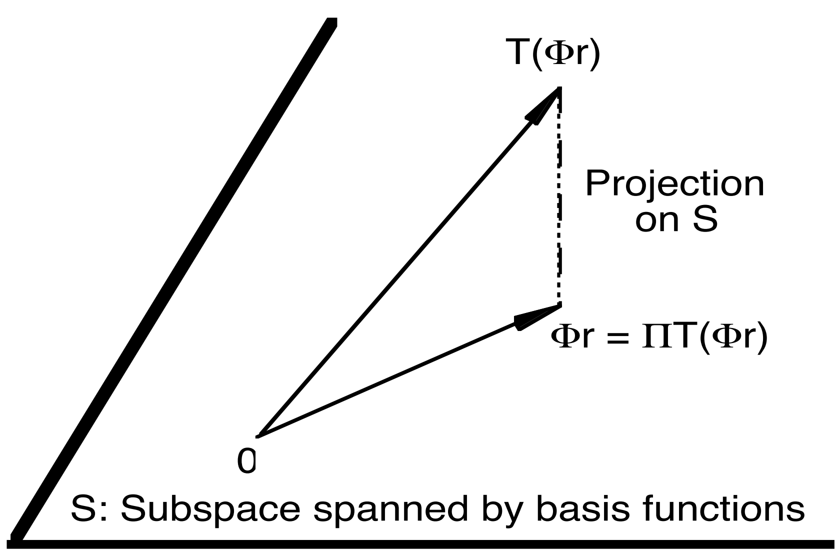
$$(TJ)(i) = \sum_{i=1}^{n} p_{ij}\big(g(i,j)+\alpha J(j)\big), \qquad i = 1,\ldots,n,$$

or more compactly,

$$TJ = g + \alpha PJ$$

- The solution J_μ of Bellman's equation $J = TJ$ is approximated by the solution of

$$\Phi r = \Pi T(\Phi r)$$



Indirect method: Solving a projected form of Bellman's equation

KEY QUESTIONS AND RESULTS

- Does the projected equation have a solution?
- Under what conditions is the mapping ΠT a contraction, so ΠT has unique fixed point?
- Assuming ΠT has unique fixed point Φr^*, how close is Φr^* to J_μ?
- **Assumption:** P has a single recurrent class and no transient states, i.e., it has steady-state probabilities that are positive

$$\xi_j = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} P(i_k = j \mid i_0 = i) > 0, \quad j = 1, \ldots, n$$

- **Proposition:** ΠT is contraction of modulus α with respect to the weighted Euclidean norm $\|\cdot\|_\xi$, where $\xi = (\xi_1, \ldots, \xi_n)$ is the steady-state probability vector. The unique fixed point Φr^* of ΠT satisfies

$$\|J_\mu - \Phi r^*\|_\xi \leq \frac{1}{\sqrt{1-\alpha^2}} \|J_\mu - \Pi J_\mu\|_\xi$$

ANALYSIS

- Important property of the projection Π on S with weighted Euclidean norm $\|\cdot\|_v$. For all $J \in \Re^n$, $\overline{J} \in S$, the *Pythagorean Theorem* holds:

$$\|J - \overline{J}\|_v^2 = \|J - \Pi J\|_v^2 + \|\Pi J - \overline{J}\|_v^2$$

- Proof: Geometrically, $(J - \Pi J)$ and $(\Pi J - \overline{J})$ are orthogonal in the scaled geometry of the norm $\|\cdot\|_v$, where two vectors $x, y \in \Re^n$ are orthogonal if $\sum_{i=1}^n v_i x_i y_i = 0$. Expand the quadratic in the RHS below:

$$\|J - \overline{J}\|_v^2 = \|(J - \Pi J) + (\Pi J - \overline{J})\|_v^2$$

- The Pythagorean Theorem implies that the projection is *nonexpansive*, i.e.,

$$\|\Pi J - \Pi \bar{J}\|_v \leq \|J - \bar{J}\|_v, \qquad \text{for all } J, \bar{J} \in \Re^n.$$

To see this, note that

$$\begin{aligned}
\left\|\Pi(J - \overline{J})\right\|_v^2 &\leq \left\|\Pi(J - \overline{J})\right\|_v^2 + \left\|(I - \Pi)(J - \overline{J})\right\|_v^2 \\
&= \|J - \overline{J}\|_v^2
\end{aligned}$$

PROOF OF CONTRACTION PROPERTY

- Lemma: We have

$$\|Pz\|_\xi \leq \|z\|_\xi, \qquad z \in \Re^n$$

- Proof of lemma: Let p_{ij} be the components of P. For all $z \in \Re^n$, we have

$$\|Pz\|_\xi^2 = \sum_{i=1}^n \xi_i \left(\sum_{j=1}^n p_{ij} z_j \right)^2 \leq \sum_{i=1}^n \xi_i \sum_{j=1}^n p_{ij} z_j^2$$

$$= \sum_{j=1}^n \sum_{i=1}^n \xi_i p_{ij} z_j^2 = \sum_{j=1}^n \xi_j z_j^2 = \|z\|_\xi^2,$$

where the inequality follows from the convexity of the quadratic function, and the next to last equality follows from the defining property $\sum_{i=1}^n \xi_i p_{ij} = \xi_j$ of the steady-state probabilities.

- Using the lemma, the nonexpansiveness of Π, and the definition $TJ = g + \alpha PJ$, we have

$$\|\Pi TJ - \Pi T\bar{J}\|_\xi \leq \|TJ - T\bar{J}\|_\xi = \alpha \|P(J - \bar{J})\|_\xi \leq \alpha \|J - \bar{J}\|_\xi$$

for all $J, \bar{J} \in \Re^n$. Hence T is a contraction of modulus α.

PROOF OF ERROR BOUND

- Let Φr^* be the fixed point of ΠT. We have

$$\|J_\mu - \Phi r^*\|_\xi \leq \frac{1}{\sqrt{1-\alpha^2}} \|J_\mu - \Pi J_\mu\|_\xi.$$

Proof: We have

$$\begin{aligned}
\|J_\mu - \Phi r^*\|_\xi^2 &= \|J_\mu - \Pi J_\mu\|_\xi^2 + \big\|\Pi J_\mu - \Phi r^*\big\|_\xi^2 \\
&= \|J_\mu - \Pi J_\mu\|_\xi^2 + \big\|\Pi T J_\mu - \Pi T(\Phi r^*)\big\|_\xi^2 \\
&\leq \|J_\mu - \Pi J_\mu\|_\xi^2 + \alpha^2 \|J_\mu - \Phi r^*\|_\xi^2,
\end{aligned}$$

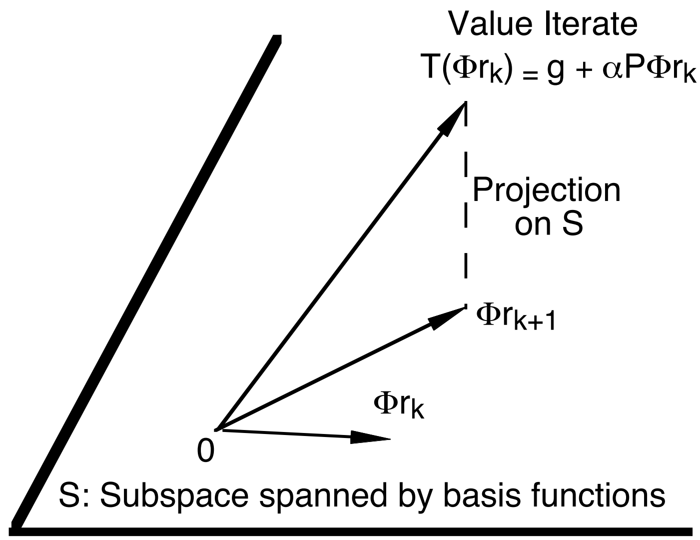
where the first equality uses the Pythagorean Theorem, the second equality holds because J_μ is the fixed point of T and Φr^* is the fixed point of ΠT, and the inequality uses the contraction property of ΠT. From this relation, the result follows.

- Note: The factor $1/\sqrt{1-\alpha^2}$ in the RHS can be replaced by a factor that is smaller and computable. See
H. Yu and D. P. Bertsekas, "New Error Bounds for Approximations from Projected Linear Equations," Report LIDS-P-2797, MIT, July 2008.

PROJECTED VALUE ITERATION (PVI)

- Given the projection property of ΠT, we may consider the PVI method

$$\Phi r_{k+1} = \Pi T(\Phi r_k)$$



- Question: Can we implement PVI using simulation, without the need for n-dimensional linear algebra calculations?

- LSPE (Least Squares Policy Evaluation) is a simulation-based implementation of PVI.

LSPE - SIMULATION-BASED PVI

- PVI, i.e., $\Phi r_{k+1} = \Pi T(\Phi r_k)$ can be written as

$$r_{k+1} = \arg\min_{r \in \Re^s} \left\| \Phi r - T(\Phi r_k) \right\|_\xi^2,$$

from which by setting the gradient to 0,

$$\left(\sum_{i=1}^{n} \xi_i \, \phi(i) \phi(i)' \right) r_{k+1} = \left(\sum_{i=1}^{n} \xi_i \, \phi(i) \sum_{j=1}^{n} p_{ij} \Big(g(i,j) + \alpha \phi(j)' r_k \Big) \right)$$

- For LSPE we generate an infinite trajectory $(i_0, i_1, \ldots)$ and update r_k after transition (i_k, i_{k+1})

$$\left(\sum_{t=0}^{k} \phi(i_t) \phi(i_t)' \right) r_{k+1} = \left(\sum_{t=0}^{k} \phi(i_t) \Big(g(i_t, i_{t+1}) + \alpha \phi(i_{t+1})' r_k \Big) \right)$$

- LSPE can equivalently be written as

$$\left(\sum_{i=1}^{n} \hat{\xi}_{i,k} \, \phi(i) \phi(i)' \right) r_{k+1} = \Bigg(\sum_{i=1}^{n} \hat{\xi}_{i,k} \, \phi(i) \sum_{j=1}^{n} \hat{p}_{ij,k} \Big(g(i,j) + \alpha \phi(j)' r_k \Big) \Bigg),$$

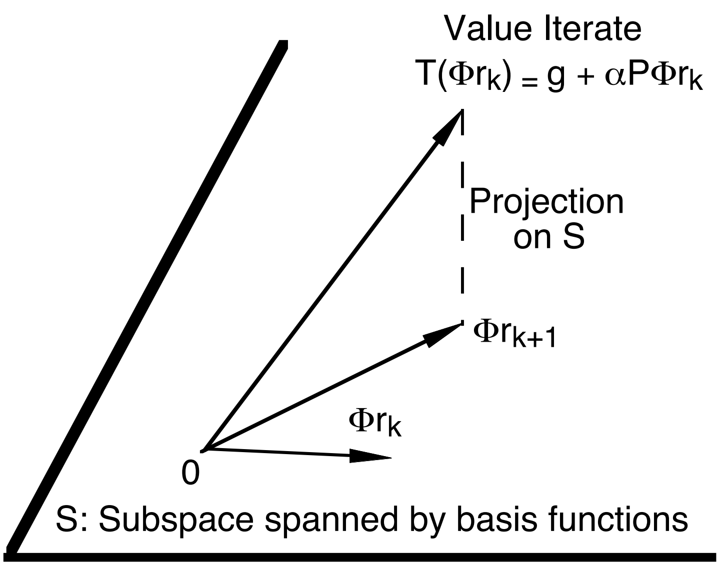
where $\hat{\xi}_{i,k}, \hat{p}_{ij,k}$: empirical frequencies of state i and transition (i, j), based on $(i_0, \ldots, i_{k+1})$.

LSPE INTERPRETATION

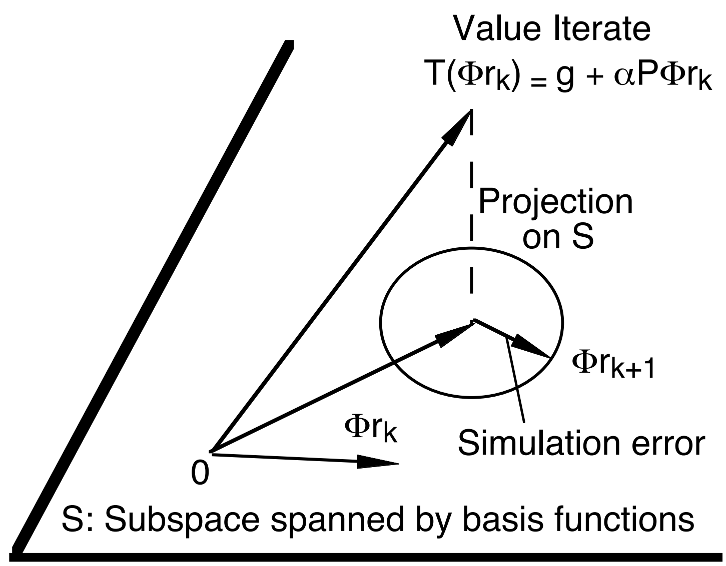
- LSPE can be written as PVI with sim. error:

$$\Phi r_{k+1} = \Pi T(\Phi r_k) + e_k$$

where e_k diminishes to 0 as the empirical frequencies $\hat{\xi}_{i,k}$ and $\hat{p}_{ij,k}$ approach ξ and p_{ij}.



Projected Value Iteration (PVI)

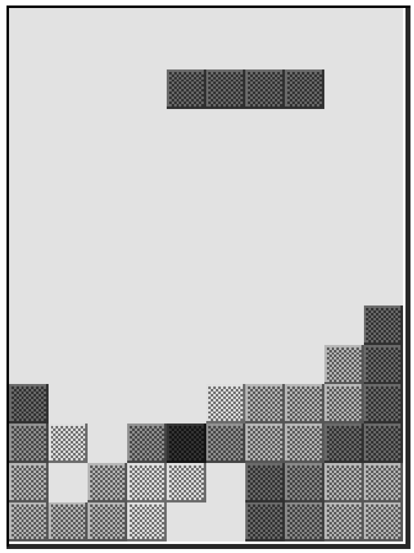


Least Squares Policy Evaluation (LSPE)

- Convergence proof is simple: Use the law of large numbers.

- Optimistic LSPE: Changes policy prior to convergence - behavior can be very complicated.

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.
- LSPE (its SSP version) was used for approximate policy evaluation.
- Both regular and optimistic versions were used.
- See: Bertsekas and Ioffe, "Temporal Differences-Based Policy Iteration and Applications in Neuro-Dynamic Programming," LIDS Report, 1996. Also the NDP book.