

# **6.231 DYNAMIC PROGRAMMING**

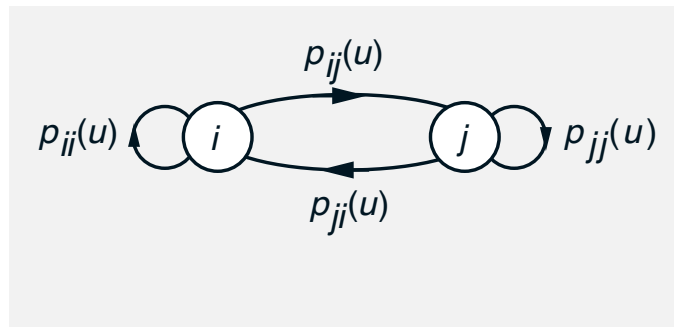
## **LECTURE 4**

### **LECTURE OUTLINE**

- Review of approximation in value space
- Approximate VI and PI
- Projected Bellman equations
- Matrix form of the projected equation
- Simulation-based implementation
- LSTD and LSPE methods
- Optimistic versions
- Multistep projected Bellman equations
- Bias-variance tradeoff

# DISCOUNTED MDP

- System: Controlled Markov chain with **states**  $i = 1, \dots, n$  and finite set of controls  $u \in U(i)$
- **Transition probabilities:**  $p_{ij}(u)$



- Cost of a policy  $\pi = \{\mu_0, \mu_1, \dots\}$  starting at state  $i$ :

$$J_\pi(i) = \lim_{N \rightarrow \infty} E \left\{ \sum_{k=0}^N \alpha^k g(i_k, \mu_k(i_k), i_{k+1}) \mid i = i_0 \right\}$$

with  $\alpha \in [0, 1)$

- **Shorthand notation for DP mappings**

$$(TJ)(i) = \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \alpha J(j)), \quad i = 1, \dots, n,$$

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

## “SHORTHAND” THEORY – A SUMMARY

- **Bellman's equation:**  $J^* = TJ^*$ ,  $J_\mu = T_\mu J_\mu$  or

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \alpha J^*(j)), \quad \forall i$$

$$J_\mu(i) = \sum_{j=1}^n p_{ij}(\mu(i)) (g(i, \mu(i), j) + \alpha J_\mu(j)), \quad \forall i$$

- **Optimality condition:**

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

i.e.,

$$\mu(i) \in \arg \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \alpha J^*(j)), \quad \forall i$$

# THE TWO MAIN ALGORITHMS: VI AND PI

- **Value iteration:** For any  $J \in \mathbb{R}^n$

$$J^*(i) = \lim_{k \rightarrow \infty} (T^k J)(i), \quad \forall i = 1, \dots, n$$

- **Policy iteration:** Given  $\mu^k$ 
  - **Policy evaluation:** Find  $J_{\mu^k}$  by solving

$$J_{\mu^k}(i) = \sum_{j=1}^n p_{ij}(\mu^k(i)) (g(i, \mu^k(i), j) + \alpha J_{\mu^k}(j)), \quad i = 1, \dots, n$$

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

- **Policy improvement:** Let  $\mu^{k+1}$  be such that

$$\mu^{k+1}(i) \in \arg \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \alpha J_{\mu^k}(j)), \quad \forall i$$

$$\text{or } T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}$$

- **Policy evaluation is equivalent to solving an  $n \times n$  linear system of equations**
- **For large  $n$ , exact PI is out of the question (even though it terminates finitely)**

# APPROXIMATION IN VALUE SPACE

- Approximate  $J^*$  or  $J_\mu$  from a parametric class  $\tilde{J}(i, r)$ , where  $i$  is the current state and  $r = (r_1, \dots, r_m)$  is a vector of “tunable” scalars weights.
- By adjusting  $r$  we can change the “shape” of  $\tilde{J}$  so that it is close to the true optimal  $J^*$ .
- Any  $r \in \mathbb{R}^s$  defines a (suboptimal) one-step lookahead policy

$$\tilde{\mu}(i) = \arg \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \alpha \tilde{J}(j, r)), \quad \forall i$$

- We will focus mostly on linear architectures

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

where  $\Phi$  is an  $n \times s$  matrix whose columns are viewed as basis functions

- Think  $n$ : HUGE,  $s$ : (Relatively) SMALL
- For  $\tilde{J}(r) = \Phi r$ , approximation in value space means approximation of  $J^*$  or  $J_\mu$  within the sub-space

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

## APPROXIMATE VI

- Approximates sequentially  $J_k(i) = (T^k J_0)(i)$ ,  $k = 1, 2, \dots$ , with  $\tilde{J}_k(i, r_k)$
- The starting function  $J_0$  is given (e.g.,  $J_0 \equiv 0$ )
- After a large enough number  $N$  of steps,  $\tilde{J}_N(i, r_N)$  is used as approximation  $\tilde{J}(i, r)$  to  $J^*(i)$
- **Fitted Value Iteration:** A sequential “fit” to produce  $\tilde{J}_{k+1}$  from  $\tilde{J}_k$ , i.e.,  $\tilde{J}_{k+1} \approx T\tilde{J}_k$  or (for a single policy  $\mu$ )  $\tilde{J}_{k+1} \approx T_\mu \tilde{J}_k$ 
  - For a “small” subset  $S_k$  of states  $i$ , compute

$$(T\tilde{J}_k)(i) = \min_{u \in U(i)} \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \alpha \tilde{J}_k(j, r))$$

- “Fit” the function  $\tilde{J}_{k+1}(i, r_{k+1})$  to the “small” set of values  $(T\tilde{J}_k)(i)$ ,  $i \in S_k$
- Simulation can be used for “model-free” implementation
- **Error Bound:** If the fit is uniformly accurate within  $\delta > 0$  (i.e.,  $\max_i |\tilde{J}_{k+1}(i) - T\tilde{J}_k(i)| \leq \delta$ ) then

$$\limsup_{k \rightarrow \infty} \max_{i=1, \dots, n} (\tilde{J}_k(i, r_k) - J^*(i)) \leq \frac{2\alpha\delta}{(1-\alpha)^2}$$

## AN EXAMPLE OF FAILURE

- Consider two-state discounted MDP with states 1 and 2, and a single policy.
  - Deterministic transitions:  $1 \rightarrow 2$  and  $2 \rightarrow 2$
  - Transition costs  $\equiv 0$ , so  $J^*(1) = J^*(2) = 0$ .
- Consider approximate VI scheme that approximates cost functions in  $S = \{(r, 2r) \mid r \in \mathbb{R}\}$  with a weighted least squares fit; here  $\Phi = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$
- Given  $J_k = (r_k, 2r_k)$ , we find  $J_{k+1} = (r_{k+1}, 2r_{k+1})$ , where for weights  $\xi_1, \xi_2 > 0$ ,  $r_{k+1}$  is obtained as

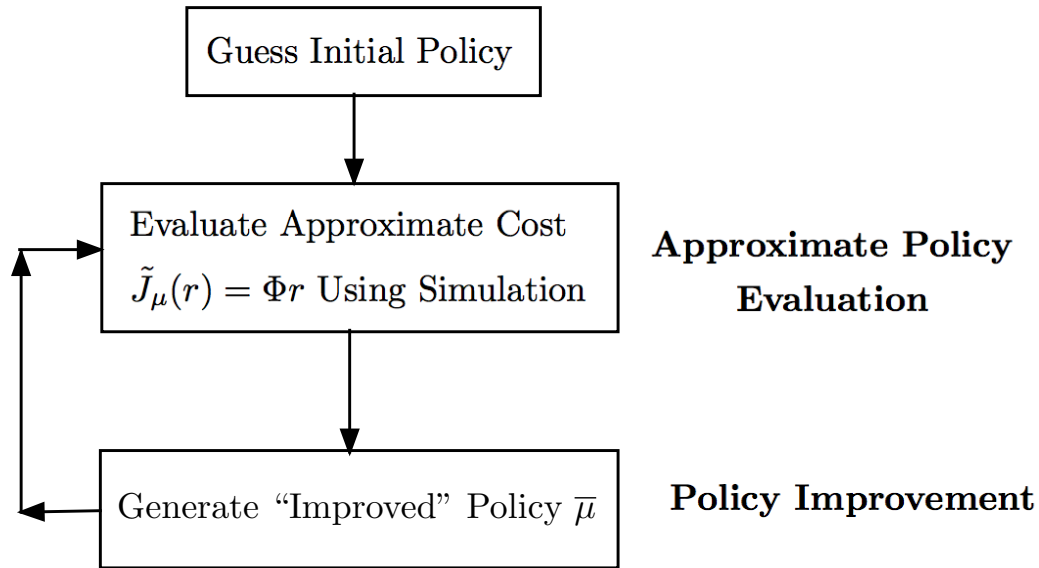
$$r_{k+1} = \arg \min_r \left[ \xi_1 (r - (TJ_k)(1))^2 + \xi_2 (2r - (TJ_k)(2))^2 \right]$$

- With straightforward calculation

$$r_{k+1} = \alpha \beta r_k, \quad \text{where } \beta = 2(\xi_1 + 2\xi_2) / (\xi_1 + 4\xi_2) > 1$$

- So if  $\alpha > 1/\beta$ , the sequence  $\{r_k\}$  diverges and so does  $\{J_k\}$ .
- Difficulty is that  $T$  is a contraction, but  $\Pi T$  (= least squares fit composed with  $T$ ) is not
- Norm mismatch problem

# APPROXIMATE PI



- **Evaluation of typical policy  $\mu$ :** Linear cost function approximation  $\tilde{J}_\mu(r) = \Phi r$ , where  $\Phi$  is full rank  $n \times s$  matrix with columns the basis functions, and  $i$ th row denoted  $\phi(i)'$ .
- **Policy “improvement”** to generate  $\bar{\mu}$ :

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

- **Error Bound:** If

$$\max_i |\tilde{J}_{\mu^k}(i, r_k) - J_{\mu^k}(i)| \leq \delta, \quad k = 0, 1, \dots$$

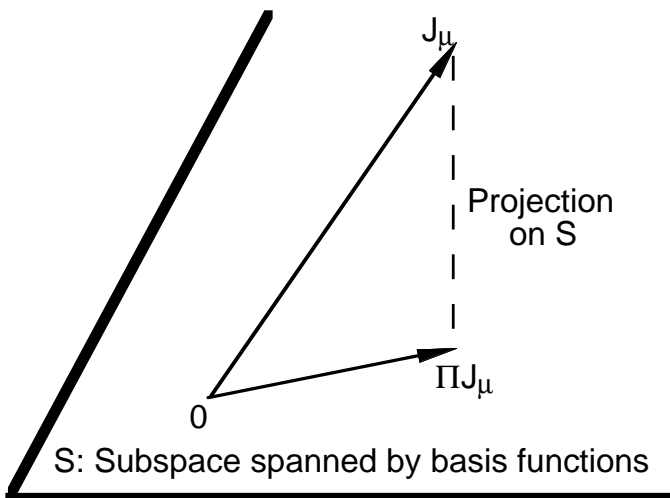
The sequence  $\{\mu^k\}$  satisfies

$$\limsup_{k \rightarrow \infty} \max_i (J_{\mu^k}(i) - J^*(i)) \leq \frac{2\alpha\delta}{(1 - \alpha)^2}$$

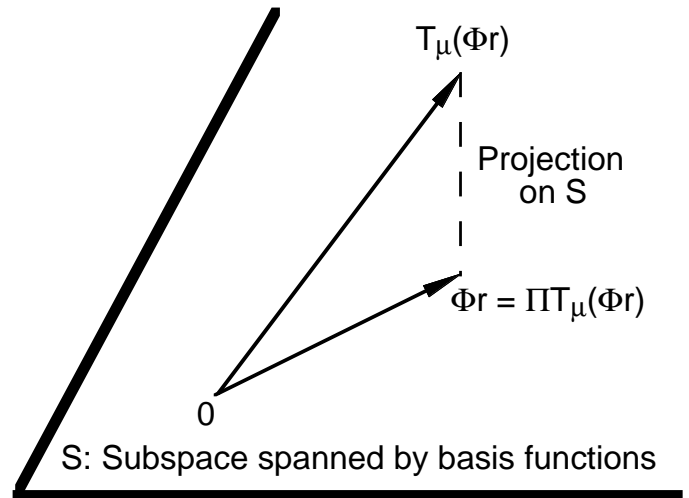


# POLICY EVALUATION

- Let's focus on policy evaluation: approximate the cost of the current policy by using a simulation method.
  - **Direct policy evaluation** - Cost samples generated by simulation, and optimization by least squares
  - **Indirect policy evaluation** - solving the projected equation  $\Phi r = \Pi T_\mu(\Phi r)$  where  $\Pi$  is projection w/ respect to a suitable weighted Euclidean norm



Direct Method: Projection of cost vector  $J_\mu$



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

- Recall that projection can be implemented by simulation and least squares

# WEIGHTED EUCLIDEAN PROJECTIONS

- Consider a weighted Euclidean norm

$$\|J\|_{\xi} = \sqrt{\sum_{i=1}^n \xi_i (J(i))^2},$$

where  $\xi$  is a vector of positive weights  $\xi_1, \dots, \xi_n$ .

- Let  $\Pi$  denote the projection operation onto

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

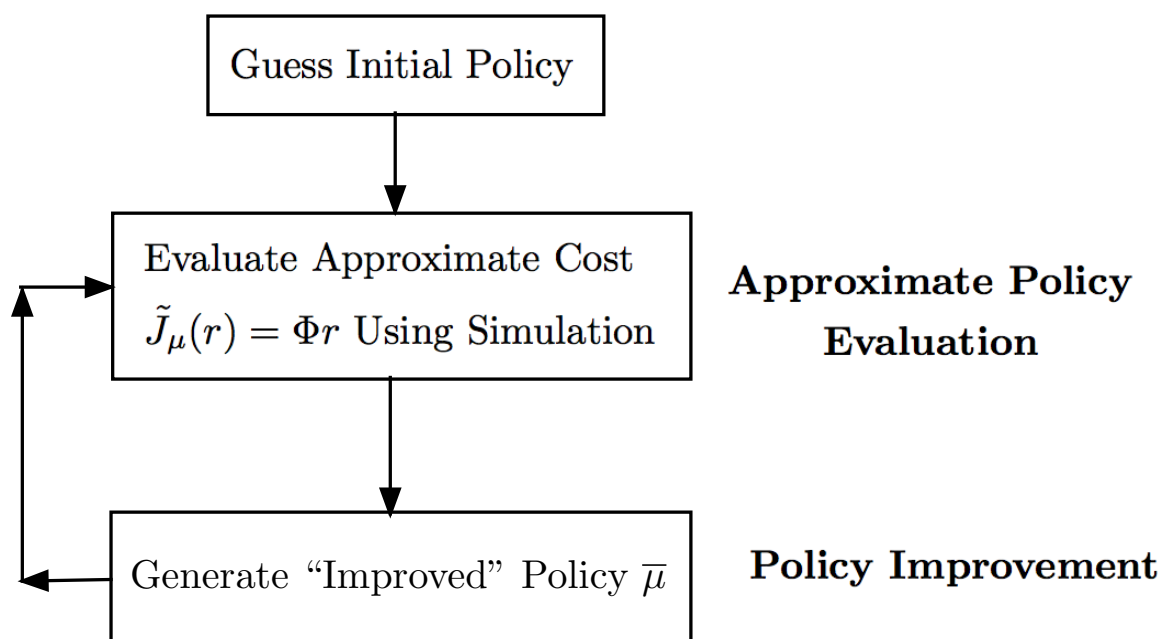
with respect to this norm, i.e., for any  $J \in \mathbb{R}^n$ ,

$$\Pi J = \Phi r^*$$

where

$$r^* = \arg \min_{r \in \mathbb{R}^s} \|J - \Phi r\|_{\xi}^2$$

# PI WITH INDIRECT POLICY EVALUATION



- Given the current policy  $\mu$ :
  - We solve the projected Bellman's equation

$$\Phi r = \Pi T_\mu(\Phi r)$$

- We approximate the solution  $J_\mu$  of Bellman's equation

$$J = T_\mu J$$

with the projected equation solution  $\tilde{J}_\mu(r)$

## KEY QUESTIONS AND RESULTS

- Does the projected equation have a solution?
- Under what conditions is the mapping  $\Pi T_\mu$  a contraction, so  $\Pi T_\mu$  has unique fixed point?
- Assuming  $\Pi T_\mu$  has unique fixed point  $\Phi r^*$ , how close is  $\Phi r^*$  to  $J_\mu$ ?
- **Assumption:** The Markov chain corresponding to  $\mu$  has a **single recurrent class and no transient states**, i.e., it has steady-state probabilities that are positive

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

- **Proposition: (Norm Matching Property)**
  - (a)  $\Pi T_\mu$  is contraction of modulus  $\alpha$  with respect to the weighted Euclidean norm  $\|\cdot\|_\xi$ , where  $\xi = (\xi_1, \dots, \xi_n)$  is the steady-state probability vector.
  - (b) The unique fixed point  $\Phi r^*$  of  $\Pi T_\mu$  satisfies

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

## PRELIMINARIES: PROJECTION PROPERTIES

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

$$\|J - \bar{J}\|_\xi^2 = \|J - \Pi J\|_\xi^2 + \|\Pi J - \bar{J}\|_\xi^2$$

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

$$\|J - \bar{J}\|_\xi^2 = \|(J - \Pi J) + (\Pi J - \bar{J})\|_\xi^2$$

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

$$\|\Pi J - \Pi \bar{J}\|_\xi \leq \|J - \bar{J}\|_\xi, \quad \text{for all } J, \bar{J} \in \Re^n.$$

To see this, note that

$$\begin{aligned} \|\Pi(J - \bar{J})\|_\xi^2 &\leq \|\Pi(J - \bar{J})\|_\xi^2 + \|(I - \Pi)(J - \bar{J})\|_\xi^2 \\ &= \|J - \bar{J}\|_\xi^2 \end{aligned}$$

# PROOF OF CONTRACTION PROPERTY

- **Lemma:** If  $P$  is the transition matrix of  $\mu$ ,

$$\|Pz\|_{\xi} \leq \|z\|_{\xi}, \quad z \in \mathfrak{R}^n$$

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

$$\begin{aligned} \|Pz\|_{\xi}^2 &= \sum_{i=1}^n \xi_i \left( \sum_{j=1}^n p_{ij} z_j \right)^2 \leq \sum_{i=1}^n \xi_i \sum_{j=1}^n p_{ij} z_j^2 \\ &= \sum_{j=1}^n \sum_{i=1}^n \xi_i p_{ij} z_j^2 = \sum_{j=1}^n \xi_j z_j^2 = \|z\|_{\xi}^2, \end{aligned}$$

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  $\Pi$ , and the definition  $T_{\mu}J = g + \alpha PJ$ , we have

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

for all  $J, \bar{J} \in \mathfrak{R}^n$ . Hence  $\Pi T_{\mu}$  is a contraction of modulus  $\alpha$ .

## PROOF OF ERROR BOUND

- Let  $\Phi r^*$  be the fixed point of  $\Pi T$ . We have

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

**Proof:** We have

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

where

- The first equality uses the Pythagorean Theorem
- The second equality holds because  $J_\mu$  is the fixed point of  $T$  and  $\Phi r^*$  is the fixed point of  $\Pi T$
- The inequality uses the contraction property of  $\Pi T$ .

**Q.E.D.**

# MATRIX FORM OF PROJECTED EQUATION

- Its solution is the vector  $J = \Phi r^*$ , where  $r^*$  solves the problem

$$\min_{r \in \mathbb{R}^s} \left\| \Phi r - (g + \alpha P \Phi r^*) \right\|_{\xi}^2.$$

- Setting to 0 the gradient with respect to  $r$  of this quadratic, we obtain

$$\Phi' \Xi (\Phi r^* - (g + \alpha P \Phi r^*)) = 0,$$

where  $\Xi$  is the diagonal matrix with the steady-state probabilities  $\xi_1, \dots, \xi_n$  along the diagonal.

- This is just the **orthogonality condition**: The error  $\Phi r^* - (g + \alpha P \Phi r^*)$  is “orthogonal” to the subspace spanned by the columns of  $\Phi$ .
- Equivalently,

$$C r^* = d,$$

where

$$C = \Phi' \Xi (I - \alpha P) \Phi, \quad d = \Phi' \Xi g.$$

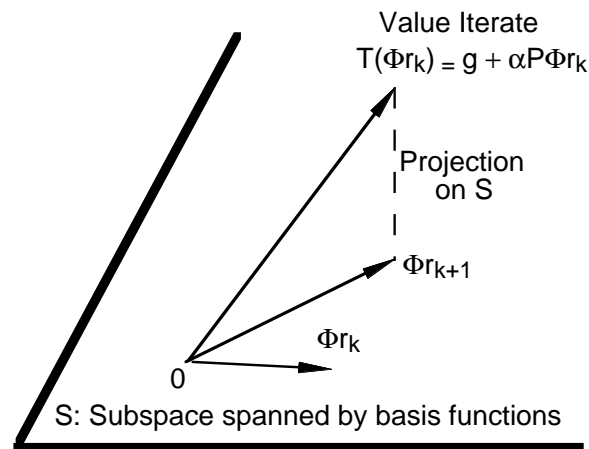


# PROJECTED EQUATION: SOLUTION METHODS

- **Matrix inversion:**  $r^* = C^{-1}d$
- **Projected Value Iteration (PVI) method:**

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

Converges to  $r^*$  because  $\Pi T$  is a contraction.



- PVI can be written as:

$$r_{k+1} = \arg \min_{r \in \mathcal{R}^s} \left\| \Phi r - (g + \alpha P \Phi r_k) \right\|_{\xi}^2$$

By setting to 0 the gradient with respect to  $r$ ,

$$\Phi' \Xi (\Phi r_{k+1} - (g + \alpha P \Phi r_k)) = 0,$$

which yields

$$r_{k+1} = r_k - (\Phi' \Xi \Phi)^{-1} (C r_k - d)$$

# SIMULATION-BASED IMPLEMENTATIONS

- **Key idea:** Calculate simulation-based approximations based on  $k$  samples

$$C_k \approx C, \quad d_k \approx d$$

- Matrix inversion  $r^* = C^{-1}d$  is approximated by

$$\hat{r}_k = C_k^{-1}d_k$$

This is the **LSTD** (Least Squares Temporal Differences) Method.

- PVI method  $r_{k+1} = r_k - (\Phi' \Xi \Phi)^{-1}(Cr_k - d)$  is approximated by

$$r_{k+1} = r_k - G_k(C_k r_k - d_k)$$

where

$$G_k \approx (\Phi' \Xi \Phi)^{-1}$$

This is the **LSPE** (Least Squares Policy Evaluation) Method.

- **Key fact:**  $C_k$ ,  $d_k$ , and  $G_k$  can be computed with low-dimensional linear algebra (of order  $s$ ; the number of basis functions).

# SIMULATION MECHANICS

- We generate an infinitely long trajectory  $(i_0, i_1, \dots)$  of the Markov chain, so states  $i$  and transitions  $(i, j)$  appear with long-term frequencies  $\xi_i$  and  $p_{ij}$ .
- After generating the transition  $(i_t, i_{t+1})$ , we compute the row  $\phi(i_t)'$  of  $\Phi$  and the cost component  $g(i_t, i_{t+1})$ .
- We form

$$C_k = \frac{1}{k+1} \sum_{t=0}^k \phi(i_t) (\phi(i_t) - \alpha \phi(i_{t+1}))' \approx \Phi' \Xi (I - \alpha P) \Phi$$

$$d_k = \frac{1}{k+1} \sum_{t=0}^k \phi(i_t) g(i_t, i_{t+1}) \approx \Phi' \Xi g$$

Also in the case of LSPE

$$G_k = \frac{1}{k+1} \sum_{t=0}^k \phi(i_t) \phi(i_t)' \approx \Phi' \Xi \Phi$$

- Convergence based on law of large numbers.
- $C_k$ ,  $d_k$ , and  $G_k$  can be formed incrementally. Also can be written using the formalism of temporal differences (this is just a matter of style)

## OPTIMISTIC VERSIONS

- Instead of calculating nearly exact approximations  $C_k \approx C$  and  $d_k \approx d$ , we do a less accurate approximation, based on **few simulation samples**
- Evaluate (coarsely) current policy  $\mu$ , then do a policy improvement
- This often leads to faster computation (as optimistic methods often do)
- Very complex behavior (see the subsequent discussion on oscillations)
- The matrix inversion/LSTD method has serious problems due to large simulation noise (because of limited sampling)
- LSPE tends to cope better because of its iterative nature
- A stepsize  $\gamma \in (0, 1]$  in LSPE may be useful to damp the effect of simulation noise

$$r_{k+1} = r_k - \gamma G_k(C_k r_k - d_k)$$

## MULTISTEP METHODS

- Introduce a multistep version of Bellman's equation  $J = T^{(\lambda)} J$ , where for  $\lambda \in [0, 1)$ ,

$$T^{(\lambda)} = (1 - \lambda) \sum_{\ell=0}^{\infty} \lambda^{\ell} T^{\ell+1}$$

Geometrically weighted sum of powers of  $T$ .

- Note that  $T^{\ell}$  is a contraction with modulus  $\alpha^{\ell}$ , with respect to the weighted Euclidean norm  $\|\cdot\|_{\xi}$ , where  $\xi$  is the steady-state probability vector of the Markov chain.

- Hence  $T^{(\lambda)}$  is a contraction with modulus

$$\alpha_{\lambda} = (1 - \lambda) \sum_{\ell=0}^{\infty} \alpha^{\ell+1} \lambda^{\ell} = \frac{\alpha(1 - \lambda)}{1 - \alpha\lambda}$$

Note that  $\alpha_{\lambda} \rightarrow 0$  as  $\lambda \rightarrow 1$

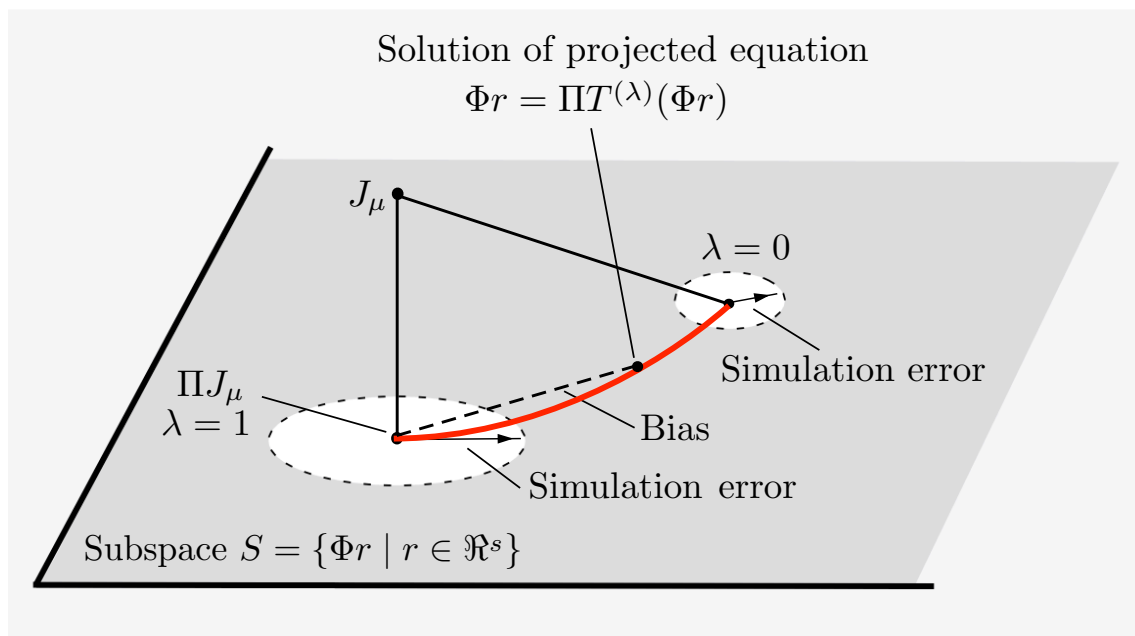
- $T^t$  and  $T^{(\lambda)}$  have the same fixed point  $J_{\mu}$  and

$$\|J_{\mu} - \Phi r_{\lambda}^*\|_{\xi} \leq \frac{1}{\sqrt{1 - \alpha_{\lambda}^2}} \|J_{\mu} - \Pi J_{\mu}\|_{\xi}$$

where  $\Phi r_{\lambda}^*$  is the fixed point of  $\Pi T^{(\lambda)}$ .

- The fixed point  $\Phi r_{\lambda}^*$  depends on  $\lambda$ .

# BIAS-VARIANCE TRADEOFF



- Error bound  $\|J_\mu - \Phi r_\lambda^*\|_\xi \leq \frac{1}{\sqrt{1-\alpha_\lambda^2}} \|J_\mu - \Pi J_\mu\|_\xi$
- As  $\lambda \uparrow 1$ , we have  $\alpha_\lambda \downarrow 0$ , so error bound (and the quality of approximation) improves as  $\lambda \uparrow 1$ . In fact

$$\lim_{\lambda \uparrow 1} \Phi r_\lambda^* = \Pi J_\mu$$

- But the simulation noise in approximating

$$T^{(\lambda)} = (1 - \lambda) \sum_{\ell=0}^{\infty} \lambda^\ell T^{\ell+1}$$

increases.

- Choice of  $\lambda$  is usually based on trial and error

# MULTISTEP PROJECTED EQ. METHODS

- The projected Bellman equation is

$$\Phi r = \Pi T^{(\lambda)}(\Phi r)$$

- In matrix form:  $C^{(\lambda)}r = d^{(\lambda)}$ , where

$$C^{(\lambda)} = \Phi' \Xi (I - \alpha P^{(\lambda)}) \Phi, \quad d^{(\lambda)} = \Phi' \Xi g^{(\lambda)},$$

with

$$P^{(\lambda)} = (1 - \lambda) \sum_{\ell=0}^{\infty} \alpha^{\ell} \lambda^{\ell} P^{\ell+1}, \quad g^{(\lambda)} = \sum_{\ell=0}^{\infty} \alpha^{\ell} \lambda^{\ell} P^{\ell} g$$

- The **LSTD( $\lambda$ ) method** is

$$(C_k^{(\lambda)})^{-1} d_k^{(\lambda)},$$

where  $C_k^{(\lambda)}$  and  $d_k^{(\lambda)}$  are simulation-based approximations of  $C^{(\lambda)}$  and  $d^{(\lambda)}$ .

- The **LSPE( $\lambda$ ) method** is

$$r_{k+1} = r_k - \gamma G_k (C_k^{(\lambda)} r_k - d_k^{(\lambda)})$$

where  $G_k$  is a simulation-based approx. to  $(\Phi' \Xi \Phi)^{-1}$

- **TD( $\lambda$ )**: An important simpler/slower iteration [similar to LSPE( $\lambda$ ) with  $G_k = I$  - see the text].

## MORE ON MULTISTEP METHODS

- The simulation process to obtain  $C_k^{(\lambda)}$  and  $d_k^{(\lambda)}$  is similar to the case  $\lambda = 0$  (single simulation trajectory  $i_0, i_1, \dots$  more complex formulas)

$$C_k^{(\lambda)} = \frac{1}{k+1} \sum_{t=0}^k \phi(i_t) \sum_{m=t}^k \alpha^{m-t} \lambda^{m-t} (\phi(i_m) - \alpha \phi(i_{m+1}))',$$

$$d_k^{(\lambda)} = \frac{1}{k+1} \sum_{t=0}^k \phi(i_t) \sum_{m=t}^k \alpha^{m-t} \lambda^{m-t} g_{i_m}$$

- In the context of approximate policy iteration, we can use optimistic versions (few samples between policy updates).
- Many different versions (see the text).
- Note the  **$\lambda$ -tradeoffs**:
  - As  $\lambda \uparrow 1$ ,  $C_k^{(\lambda)}$  and  $d_k^{(\lambda)}$  contain more “simulation noise”, so more samples are needed for a close approximation of  $r_\lambda$  (the solution of the projected equation)
  - The error bound  $\|J_\mu - \Phi r_\lambda\|_\xi$  becomes smaller
  - As  $\lambda \uparrow 1$ ,  $\Pi T^{(\lambda)}$  becomes a contraction for **arbitrary** projection norm