6.231 Dynamic Programming and Stochastic Control
Fall 2008

6.231 DYNAMIC PROGRAMMING

LECTURE 24

LECTURE OUTLINE

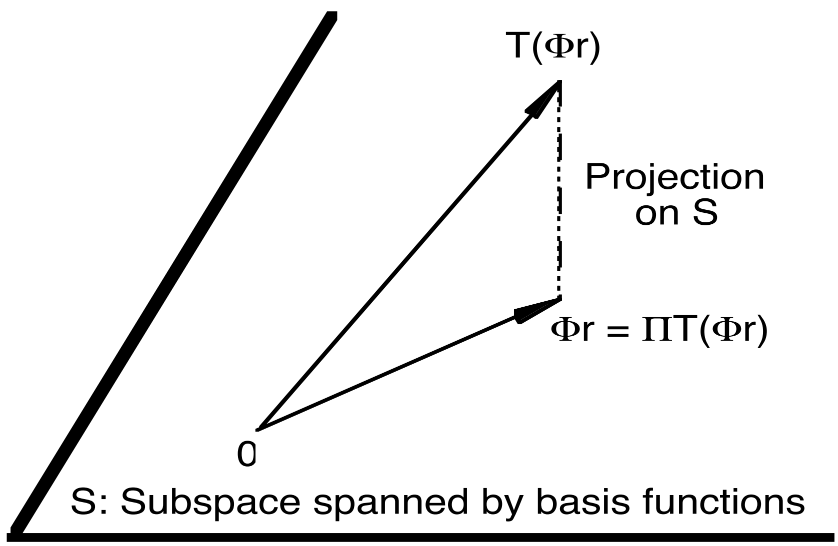
- More on projected equation methods/policy evaluation
- Stochastic shortest path problems
- Average cost problems
- Generalization - Two Markov Chain methods
- LSTD-like methods - Use to enhance exploration

REVIEW: PROJECTED BELLMAN EQUATION

- For fixed policy μ to be evaluated, the solution of Bellman's equation $J = TJ$ is approximated by the solution of

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

whose solution is in turn obtained using a simulation-based method such as LSPE(λ), LSTD(λ), or TD(λ).



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

- These ideas apply to other (linear) Bellman equations, e.g., for SSP and average cost.

- Key Issue: Construct framework where ΠT [or at least $\Pi T^{(\lambda)}$] is a contraction.

STOCHASTIC SHORTEST PATHS

- Introduce approximation subspace

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

and for a given proper policy, Bellman's equation and its projected version

$$J = TJ = g + PJ, \qquad \Phi r = \Pi T(\Phi r)$$

Also its λ-version

$$\Phi r = \Pi T^{(\lambda)}(\Phi r), \qquad T^{(\lambda)} = (1 - \lambda) \sum_{t=0}^{\infty} \lambda^t T^{t+1}$$

- **Question:** What should be the norm of projection?

- **Speculation based on discounted case:** It should be a weighted Euclidean norm with weight vector $\xi = (\xi_1, \ldots, \xi_n)$, where ξ_i should be some type of long-term occupancy probability of state i (which can be generated by simulation).

- But what does "long-term occupancy probability of a state" mean in the SSP context?

- How do we generate infinite length trajectories given that termination occurs with prob. 1?

SIMULATION TRAJECTORIES FOR SSP

- We envision simulation of trajectories up to termination, followed by restart at state i with some fixed probabilities $q_0(i) > 0$.

- Then the "long-term occupancy probability of a state" of i is proportional to

$$q(i) = \sum_{t=0}^{\infty} q_t(i), \qquad i = 1, \ldots, n,$$

where

$$q_t(i) = P(i_t = i), \qquad i = 1, \ldots, n,\ t = 0, 1, \ldots$$

- We use the projection norm

$$\|J\|_q = \sqrt{\sum_{i=1}^{n} q(i)\big(J(i)\big)^2}$$

[Note that $0 < q(i) < \infty$, but q is not a prob. distribution.]

- We can show that $\Pi T^{(\lambda)}$ is a contraction with respect to $\| \cdot \|_\xi$ (see the next slide).

CONTRACTION PROPERTY FOR SSP

- We have $q = \sum_{t=0}^{\infty} q_t$ so

$$q'P = \sum_{t=0}^{\infty} q_t'P = \sum_{t=1}^{\infty} q_t' = q' - q_0'$$

or

$$\sum_{i=1}^{n} q(i)p_{ij} = q(j) - q_0(j), \qquad \forall\ j$$

- To verify that ΠT is a contraction, we show that there exists $\beta < 1$ such that $\|Pz\|_q^2 \le \beta \|z\|_q^2$ for all $z \in \Re^n$.

- For all $z \in \Re^n$, we have

$$\begin{aligned}
\|Pz\|_q^2 &= \sum_{i=1}^{n} q(i) \left(\sum_{j=1}^{n} p_{ij} z_j \right)^2 \le \sum_{i=1}^{n} q(i) \sum_{j=1}^{n} p_{ij} z_j^2 \\
&= \sum_{j=1}^{n} z_j^2 \sum_{i=1}^{n} q(i)p_{ij} = \sum_{j=1}^{n} \big(q(j) - q_0(j) \big) z_j^2 \\
&= \|z\|_q^2 - \|z\|_{q_0}^2 \le \beta \|z\|_q^2
\end{aligned}$$

where

$$\beta = 1 - \min_j \frac{q_0(j)}{q(j)}$$

PVI(λ) AND LSPE(λ) FOR SSP

- We consider PVI(λ): $\Phi r_{k+1} = \Pi T^{(\lambda)}(\Phi r_k)$, which can be written as

$$r_{k+1} = \arg\min_{r \in \Re^s} \sum_{i=1}^{n} q(i) \Bigg(\phi(i)' r - \phi(i)' r_k - \sum_{t=0}^{\infty} \lambda^t E\big\{ d_k(i_t, i_{t+1}) \mid i_0 = i \big\} \Bigg)^2$$

where $d_k(i_t, i_{t+1})$ are the TDs.

- The LSPE(λ) algorithm is a simulation-based approximation. Let $(i_{0,l}, i_{1,l}, \ldots, i_{N_l,l})$ be the lth trajectory (with $i_{N_l,l} = 0$), and let r_k be the parameter vector after k trajectories. We set

$$r_{k+1} = \arg\min_{r} \sum_{l=1}^{k+1} \sum_{t=0}^{N_l - 1} \Bigg(\phi(i_{t,l})' r - \phi(i_{t,l})' r_k - \sum_{m=t}^{N_l - 1} \lambda^{m-t} d_k(i_{m,l}, i_{m+1,l}) \Bigg)^2$$

where

$$d_k(i_{m,l}, i_{m+1,l}) = g(i_{m,l}, i_{m+1,l}) + \phi(i_{m+1,l})' r_k - \phi(i_{m,l})' r_k$$

- Can also update r_k at every transition.

AVERAGE COST PROBLEMS

- Consider a single policy to be evaluated, with single recurrent class, no transient states, and steady-state probability vector $\xi = (\xi_1, \ldots, \xi_n)$.
- The average cost, denoted by η, is independent of the initial state

$$\eta = \lim_{N \to \infty} \frac{1}{N} E\left\{ \sum_{k=0}^{N-1} g\big(x_k, x_{k+1}\big) \;\Big|\; x_0 = i \right\}, \quad \forall\, i$$

- Bellman's equation is $J = FJ$ with

$$FJ = g - \eta e + PJ$$

where e is the unit vector $e = (1, \ldots, 1)$.

- The projected equation and its λ-version are

$$\Phi r = \Pi F(\Phi r), \qquad \Phi r = \Pi F^{(\lambda)}(\Phi r)$$

- A problem here is that F is not a contraction with respect to any norm (since $e = Pe$).
- However, $\Pi F^{(\lambda)}$ turns out to be a contraction with respect to $\|\cdot\|_\xi$ assuming that e does not belong to S and $\lambda > 0$ [the case $\lambda = 0$ is exceptional, but can be handled - see the text].

LSPE(λ) FOR AVERAGE COST

- We generate an infinitely long trajectory $(i_0, i_1, \ldots)$.
- We estimate the average cost η separately: Following each transition (i_k, i_{k+1}), we set

$$\eta_k = \frac{1}{k+1} \sum_{t=0}^{k} g(i_t, i_{t+1})$$

- Also following (i_k, i_{k+1}), we update r_k by

$$r_{k+1} = \arg\min_{r \in \Re^s} \sum_{t=0}^{k} \left(\phi(i_t)'r - \phi(i_t)'r_k - \sum_{m=t}^{k} \lambda^{m-t} d_k(m) \right)^2$$

where $d_k(m)$ are the TDs

$$d_k(m) = g(i_m, i_{m+1}) - \eta_m + \phi(i_{m+1})'r_k - \phi(i_m)'r_k$$

- Note that the TDs include the estimate η_m. Since η_m converges to η, for large m it can be viewed as a constant and lumped into the one-stage cost.

GENERALIZATION/UNIFICATION

- Consider approximate solution of $x = T(x)$, where

$$T(x) = Ax + b, \qquad A \text{ is } n \times n, \quad b \in \Re^n$$

by solving the projected equation $y = \Pi T(y)$, where Π is projection on a subspace of basis functions (with respect to some Euclidean norm).

- We will generalize from DP to the case where A is arbitrary, subject only to

$$I - \Pi A: \text{ invertible}$$

- Benefits of generalization:
  - Unification/higher perspective for TD methods in approximate DP
  - An extension to a broad new area of applications, where a DP perspective may be helpful
- Challenge: Dealing with less structure
  - Lack of contraction
  - Absence of a Markov chain

LSTD-LIKE METHOD

- Let Π be projection with respect to

$$\|x\|_\xi = \sqrt{\sum_{i=1}^{n} \xi_i x_i^2}$$

where $\xi \in \Re^n$ is a probability distribution with positive components.

- If r^* is the solution of the projected equation, we have $\Phi r^* = \Pi(A\Phi r^* + b)$ or

$$r^* = \arg\min_{r \in \Re^s} \sum_{i=1}^{n} \xi_i \left(\phi(i)' r - \sum_{j=1}^{n} a_{ij} \phi(j)' r^* - b_i \right)^2$$

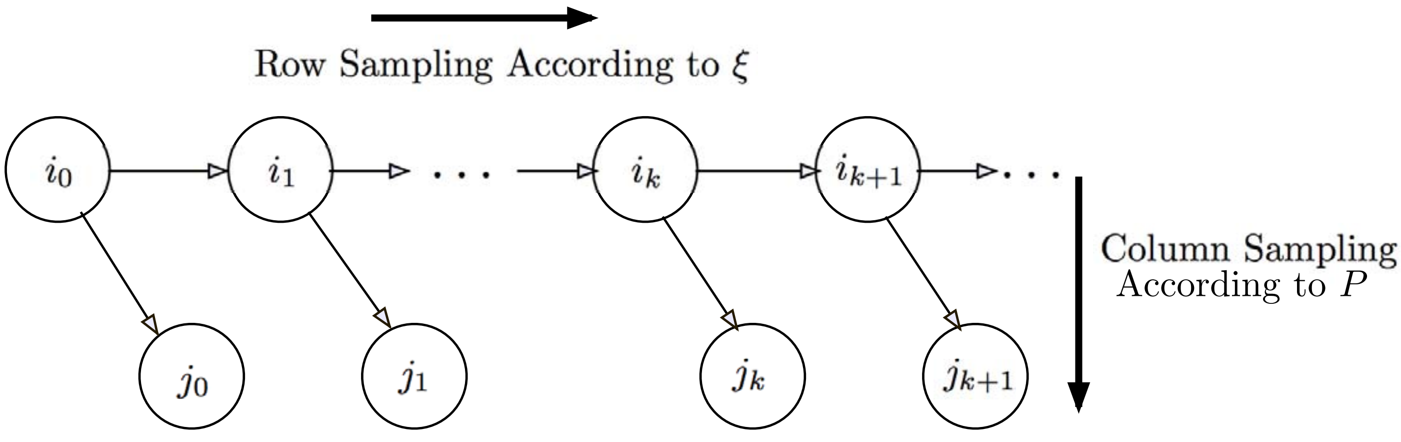
where $\phi(i)'$ denotes the ith row of the matrix Φ.

- Optimality condition/equivalent form:

$$\sum_{i=1}^{n} \xi_i \phi(i) \left(\phi(i) - \sum_{j=1}^{n} a_{ij} \phi(j) \right)' r^* = \sum_{i=1}^{n} \xi_i \phi(i) b_i$$

- The two expected values are approximated by simulation.

SIMULATION MECHANISM



- Row sampling: Generate sequence $\{i_0, i_1, \ldots\}$ according to ξ, i.e., relative frequency of each row i is ξ_i

- Column sampling: Generate $\{(i_0, j_0), (i_1, j_1), \ldots\}$ according to some transition probability matrix P with

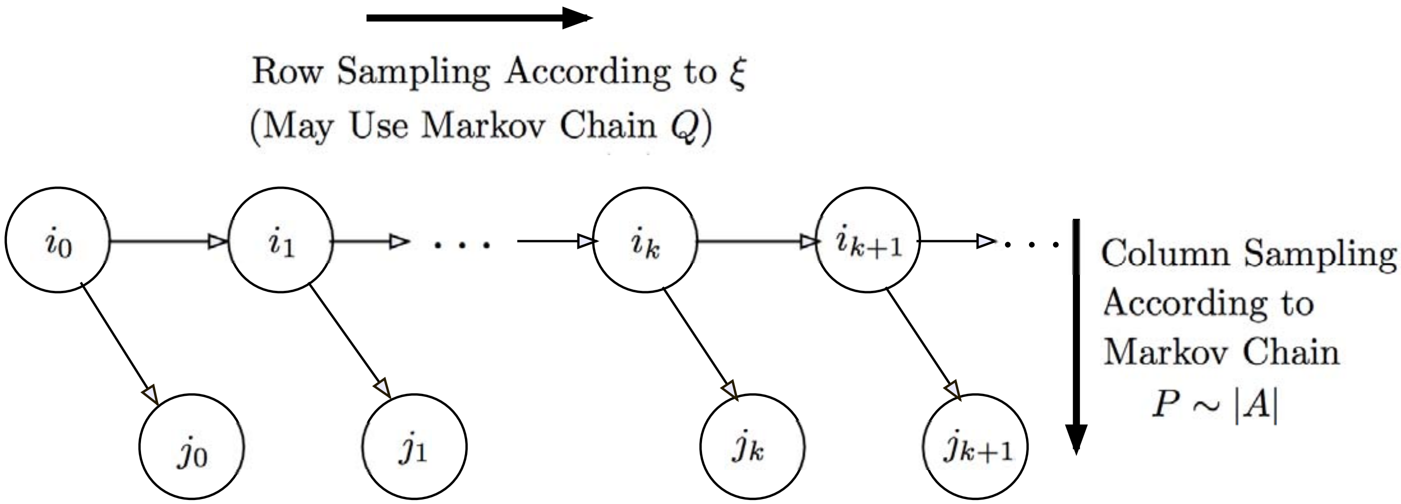
$$p_{ij} > 0 \qquad \text{if} \qquad a_{ij} \neq 0,$$

i.e., for each i, the relative frequency of (i, j) is p_{ij}

- Row sampling may be done using a Markov chain with transition matrix Q (unrelated to P)

- Row sampling may also be done without a Markov chain - just sample rows according to some known distribution ξ (e.g., a uniform)

ROW AND COLUMN SAMPLING



- Row sampling $\sim$ State Sequence Generation in DP. Affects:
  - The projection norm.
  - Whether ΠA is a contraction.
- Column sampling $\sim$ Transition Sequence Generation in DP.
  - Can be totally unrelated to row sampling. Affects the sampling/simulation error.
  - "Matching" P with $|A|$ is beneficial (has an effect like in importance sampling).
- Independent row and column sampling allows exploration at will! Resolves the exploration problem that is critical in approximate policy iteration.

LSTD-LIKE METHOD

- Optimality condition/equivalent form of projected equation

$$\sum_{i=1}^{n} \xi_i \phi(i) \left(\phi(i) - \sum_{j=1}^{n} a_{ij} \phi(j) \right)' r^* = \sum_{i=1}^{n} \xi_i \phi(i) b_i$$

- The two expected values are approximated by row and column sampling (batch $0 \to t$).
- We solve the linear equation

$$\sum_{k=0}^{t} \phi(i_k) \left(\phi(i_k) - \frac{a_{i_k j_k}}{p_{i_k j_k}} \phi(j_k) \right)' r_t = \sum_{k=0}^{t} \phi(i_k) b_{i_k}$$

- We have $r_t \to r^*$, regardless of ΠA being a contraction (by law of large numbers; see next slide).
- An LSPE-like method is also possible, but requires that ΠA is a contraction.
- Under the assumption $\sum_{j=1}^{n} |a_{ij}| \leq 1$ for all i, there are conditions that guarantee contraction of ΠA; see the paper by Bertsekas and Yu, "Projected Equation Methods for Approximate Solution of Large Linear Systems," 2008.

JUSTIFICATION W/ LAW OF LARGE NUMBERS

- We will match terms in the exact optimality condition and the simulation-based version.
- Let $\hat{\xi}_i^t$ be the relative frequency of i in row sampling up to time t.
- We have

$$\frac{1}{t+1}\sum_{k=0}^{t}\phi(i_k)\phi(i_k)' = \sum_{i=1}^{n}\hat{\xi}_i^t\phi(i)\phi(i)' \approx \sum_{i=1}^{n}\xi_i\phi(i)\phi(i)'$$

$$\frac{1}{t+1}\sum_{k=0}^{t}\phi(i_k)b_{i_k} = \sum_{i=1}^{n}\hat{\xi}_i^t\phi(i)b_i \approx \sum_{i=1}^{n}\xi_i\phi(i)b_i$$

- Let $\hat{p}_{ij}^t$ be the relative frequency of (i,j) in column sampling up to time t.

$$\begin{aligned}\frac{1}{t+1}\sum_{k=0}^{t}\frac{a_{i_kj_k}}{p_{i_kj_k}}\phi(i_k)\phi(j_k)' \\ &= \sum_{i=1}^{n}\hat{\xi}_i^t\sum_{j=1}^{n}\hat{p}_{ij}^t\frac{a_{ij}}{p_{ij}}\phi(i)\phi(j)' \\ &\approx \sum_{i=1}^{n}\xi_i\sum_{j=1}^{n}a_{ij}\phi(i)\phi(j)'\end{aligned}$$