Temporal Difference Methods and Approximate Monte Carlo Linear Algebra

Dimitri P. Bertsekas

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

RL Workshop, Lille 2008

Focus

- Approximate solution of linear equations $x = T(x)$, where

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

  by solving the projected equation

$$y = \Pi T(y)$$

  Π is projection on a subspace of basis functions (with respect to some norm)
- This is the Galerkin approximation approach, but simulation plays a central and non-traditional role. We consider very large n.
- Starting point: Approximate DP/Bellman's equation/policy evaluation

$$A: \text{encodes the Markov chain structure}, \qquad b: \text{cost vector}$$

  Then $y = \Pi T(y)$ is the equation solved by TD methods [TD(λ), LSTD(λ), LSPE(λ)]
- We generalize to the case where A is arbitrary, subject only to

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

  (joint work with H. Yu - papers available from our web sites)

Benefits and Challenges of Generalization

- A higher perspective for TD methods in approximate DP
  Motivates improvements in various areas:
  - Exploration issues
  - Automatic generation of features
  - Error bounds
  - Simplified convergence analysis
- An extension to a vast new area of applications
  There are many linear systems of huge dimension in practice
- Dealing with less structure
  - Lack of contraction
  - Absence of a Markov chain
  - Ill-conditioning

Outline

DP Context/Policy Evaluation

- Markovian Decision Problems (MDP)
- n states, transition probabilities depending on control
- Policy iteration method; we focus on single policy evaluation

- Bellman's equation:

$$x = Ax + b$$

  where
  - b: cost vector
  - A has transition structure, e.g., $A = \alpha P$ for discounted problems, $A = P$ for average cost problems

Approximate Policy Evaluation

- Approximation within subspace $S = \{\Phi r \mid r \in \Re^s\}$

$$x \approx \Phi r, \qquad \Phi \text{ is a matrix with basis functions as columns}$$

- Projected Bellman equation:

$$\Phi r = \Pi(A\Phi r + b)$$

- Error bound, assuming ΠA is contraction with modulus $\alpha \in (0, 1)$

$$\|x^* - \Phi r^*\| \leq \frac{1}{1-\alpha}\|x^* - \Pi x^*\|$$

- Long history, starting with TD(λ) (Sutton, 1988)
- Least squares methods are currently more popular

Least Squares Policy Evaluation (LSTD)

- Dates to 1996 (Bradtke and Barto), with λ-extension by Boyan (2002)
- Idea: Solve a simulation-based approximation of the projected equation
  - The projected Bellman equation is written as $Cr = d$
  - LSTD solves $\hat{C}r = \hat{d}$, where

$$\hat{C} \approx C, \qquad \hat{d} \approx d$$

  are obtained using simulation
- Does not need the contraction property of DP problems
- Multistep version: LSTD(λ) which is LSTD applied to the mapping

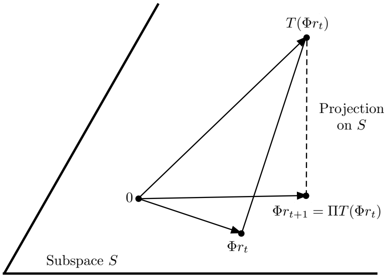
$$T^{(\lambda)}(x) = (1-\lambda)\sum_{k=0}^{\infty} \lambda^k T^{k+1}(x) = A^{(\lambda)}x + b^{(\lambda)},$$

where

$$A^{(\lambda)} = (1-\lambda)\sum_{k=0}^{\infty} \lambda^k A^{k+1}, \qquad b^{(\lambda)} = \sum_{k=0}^{\infty} \lambda^k A^k b$$

Projected Value Iteration (PVI)

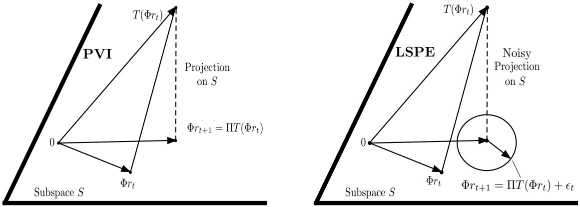
- Value Iteration => Projection => Value Iteration => Projection ....



$$\Phi r_{t+1} = \Pi T(\Phi r_t)$$

- ΠT must be a contraction - T being a contraction is not enough
- Norm matching is essential: a (Euclidean) projection norm for which T is a contraction
- There is a magical norm: the steady-state distribution norm (states are weighted by the steady-state distribution of the Markov chain)

Least Squares Policy Evaluation (LSPE)



- A simulation-based approximation to PVI
- Dates to 1996 (Bertsekas and Ioffe); also in the Bertsekas and Tsitsiklis (1996) book - used in a tetris application

$$\text{LSPE:}\quad \Phi r_{t+1} = \underbrace{\Pi T(\Phi r_t)}_{\text{PVI}} + \epsilon_t, \qquad \epsilon_t \text{ is simulation noise with } \epsilon_t \to 0$$

- Incremental like TD(λ) - no stepsize unlike TD(λ)
- Same complexity/same solution as LSTD
- Asymptotically “identical" to LSTD, but differs in early stages
- Allows for a favorable initial guess r_0; may be an advantage in optimistic/few samples approximate policy iteration

Advantages of Projected Equation Methods in DP

- All operations are done in low-dimension
- The high-dimensional vector x need not be stored
- The projection norm is implemented in simulation - need not be known a priori
- There is a projection norm (the distribution norm) that induces contraction of ΠA and a priori error bounds

General/NonDP Projected Equation Methods

- A does not have a transition probability structure
- No Markov chain, no contraction guarantee
- We may introduce an artificial Markov chain for sampling/projection
- With clever choice of the chain, ΠA may be a contraction
- Computable error bounds are available
- All operations are done in low-dimension
- The high-dimensional vector x need not be stored
- Methods:
  - LSTD analog (does not require ΠA to be a contraction)
  - LSPE analog (requires ΠA to be a contraction)
  - TD(λ) analog (requires ΠA to be a contraction)

Projected Equation Approximation Method (LSTD-like)

- 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

- Explicit form of projected equation $\Phi r = \Pi(A\Phi r + b)$

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

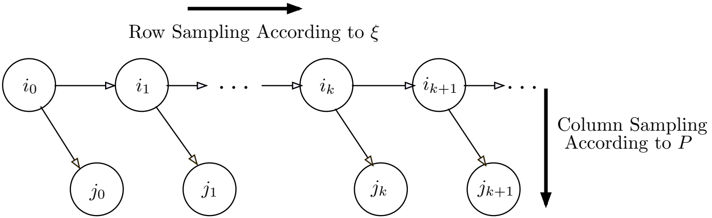
where $\phi(i)'$ denotes the ith row of the matrix Φ

- Optimality condition/equivalent form:

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

- The two expected values are approximated by simulation

Simulation Mechanism

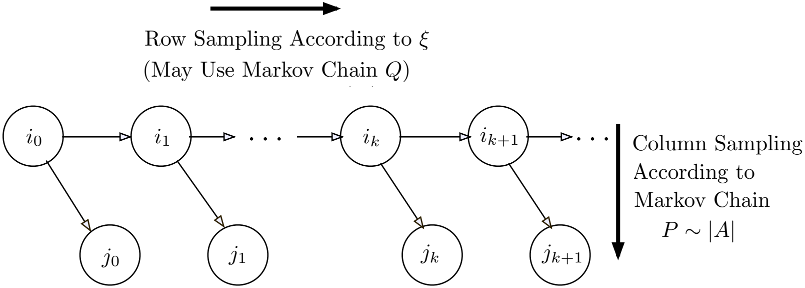


- Row sampling: Generate sequence $\{i_0, i_1, \ldots\}$ according to ξ, i.e., relative frequency of each row i is ξ_i
- Column sampling: Generate sequence $\{(i_0, j_0), (i_1, j_1), \ldots\}$ according to some transition probability matrix P with

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

  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 noise
  - Matching P with $|A|$ has an effect like in importance sampling

LSTD-Like Method

- Optimality condition/equivalent form of projected equation

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

- The two expected values are approximated by row and column sampling (batch $0 \rightarrow t$)
- At time t, we solve the linear equation

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

- Then $r_t \rightarrow r^*$

LSPE-Type Method

- Consider PVI

$$\Phi r_{t+1} = \Pi(A\Phi r_t + b), \qquad t = 0, 1, \ldots$$

- Expressing the projection as a least squares minimization, we have

$$r_{t+1} = \arg\min_{r \in \Re^s} \left\| \Phi r - (A\Phi r_t + b) \right\|_\xi^2,$$

or equivalently

$$r_{t+1} = \underbrace{\left(\sum_{i=1}^n \xi_i \, \phi(i)\phi(i)' \right)^{-1}}_{\text{Expected value}} \underbrace{\sum_{i=1}^n \xi_i \, \phi(i) \left(\sum_{j=1}^n a_{ij}\phi(j)' r_t + b_i \right)}_{\text{Expected value}}$$

- Approximate the two expected values by row and column sampling

$$r_{t+1} = \left(\sum_{k=0}^t \phi(i_k)\phi(i_k)' \right)^{-1} \sum_{k=0}^t \phi(i_k) \left(\frac{a_{i_k j_k}}{p_{i_k j_k}} \phi(j_k)' r_t + b_{i_k} \right)$$

- If ΠA is a contraction with respect to some norm, $r_t \to r^*$

Row Sampling for Contraction I

Must have Row Sums of $|A| \leq 1$ to have hope of contraction of ΠA

Proposition: Let ξ be the invariant distribution of an irreducible Q such that

$$|A| \leq Q$$

Then T and ΠT are contraction mappings under any one of the following three conditions:

(1) For some scalar $\alpha \in (0, 1)$, we have $|A| \leq \alpha Q$.

(2) There exists an index $\bar{i}$ such that $|a_{\bar{i}j}| < q_{\bar{i}j}$ for all $j = 1, \ldots, n$.

(3) There exists an index $\bar{i}$ such that $\sum_{j=1}^{n} |a_{\bar{i}j}| < 1$.

Note 1: Under conditions (1) and (2), T and ΠT are contraction mappings with respect to the specific norm $\|\cdot\|_\xi$

Note 2: Applies to DP discounted and stochastic shortest path problems

Row Sampling for Contraction II

Must have Row Sums of $|A| \leq 1$

Proposition: Let ξ be the invariant distribution of a Q with no transient states. Assume

$$|A| \leq Q$$

and that $I - \Pi A$ is invertible. Then the mapping ΠT_γ, where

$$T_\gamma = (1 - \gamma)I + \gamma T,$$

is a contraction with respect to $\| \cdot \|_\xi$ for all $\gamma \in (0, 1)$.

Note 1: ΠT_γ and ΠT have the same fixed points
Note 2: ΠT need not be a contraction
Note 3: Applies to average cost problems (Yu and Bertsekas 2006)

Back to Discounted DP/Exploration

- Here $A = \alpha P$, where P corresponds to the policy evaluated and α is the discount factor
- If we take $Q = P$ for row sampling, then ΠA is a contraction
- We may also use Markov chain $Q \neq P$ for row sampling, to change ξ and induce exploration; for example use

  Policy R (off policy) prob. β, Policy P (on policy) prob. $1 - \beta$

- The LSTD-type algorithm always applies (it does not require that ΠA be a contraction)
- If ΠA can be shown to be a contraction, the LSPE(λ)- and TD(λ)-type algorithms apply. In particular, we get convergence with no bias if:
  - (1) For all $\lambda \in [0, 1)$ if $\beta \leq 1 - \alpha^2$
  - (2) For all $\beta \in [0, 1)$ if λ is sufficiently large

Application to Diagonally Dominant Systems

- Consider the solution of the system

$$Cx = d,$$

where $d \in \Re^n$ and C is an $n \times n$ matrix such

$$c_{ii} \neq 0, \qquad \sum_{j \neq i} |c_{ij}| \leq |c_{ii}|, \qquad i = 1, \ldots, n$$

- Convert to the system $x = Ax + b$, where $b_i = \frac{d_i}{c_{ii}}$ and

$$a_{ij} = \begin{cases} 0 & \text{if } i = j \\ -\frac{c_{ij}}{c_{ii}} & \text{if } i \neq j \end{cases}$$

- We have

$$\sum_{j=1}^{n} |a_{ij}| = \sum_{j \neq i} \frac{|c_{ij}|}{|c_{ii}|} \leq 1, \qquad i = 1, \ldots, n,$$

so row sums of $|A| \leq 1$

- Under the earlier conditions, ΠA is a contraction.

Automatic Generation of Powers of A as Basis Functions

- Use Φ whose ith row is

$$\phi(i)' = \big(g(i)\ (Ag)(i)\ \cdots\ (A^s g)(i)\big)$$

  where g is some vector
- Example in the MDP case: Use as features finite horizon costs
- A justification if A is a contraction and $g = b$: the fixed point of T has an expansion of the form

$$x^* = \sum_{k=0}^{\infty} A^k b$$

- While $(A^k g)(i)$ is hard to generate, it can be approximated by sampling (in effect we use noisy features)

Multistep Versions (Fixed Step and λ-Methods)

- Replace T by a multistep mapping with the same fixed points, e.g., T^k where k is fixed, or

$$T^{(\lambda)} = (1-\lambda)\sum_{k=0}^{\infty}\lambda^k T^{k+1}, \qquad A^{(\lambda)} = (1-\lambda)\sum_{k=0}^{\infty}\lambda^k A^{k+1},$$

  where $\lambda \in (0,1)$ is such that the infinite series converges

- Motivation for λ-methods, assuming that

$$\text{spectral radius of } A \ \equiv \sigma(A) \leq 1$$

- Proposition: If $I - A$ is invertible and $\sigma(A) \leq 1$, then

$$\sigma(A^{(\lambda)}) < 1, \quad \forall\, \lambda \in (0,1), \qquad \lim_{\lambda\to 1} \sigma\big(A^{(\lambda)}\big) = 0$$

- As λ increases the contraction becomes stronger
- We must have $\lambda < 1/\sigma(A)$ for a λ-method to apply. There are no restrictions for a k-step method

λ-Methods

- When the LSTD/LSPE-type methods given earlier are applied to

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

  they yield generalizations to LSTD(λ) and LSPE(λ)

- The formulas involve temporal differences, based on the expansion

$$T^{(\lambda)}(x) = x + \sum_{m=0}^{\infty} \lambda^m (A^m b + A^{m+1} x - A^m x)$$

- The entire analysis of TD(λ), LSTD(λ), and LSPE(λ) for DP generalizes subject to the following restrictions:
  - Eigenvalues of λA must be within the unit circle for LSTD analogs
  - Additional contraction assumptions for LSPE(λ) and TD(λ) [i.e., $\Pi A^{(\lambda)}$ is a contraction]

Forms of λ-Methods I

- Row and column sampling are done using the same Markov chain P. Define $w_{k,0} = 1$ and for $m \geq 1$

$$w_{k,m} = \frac{a_{i_k i_{k+1}}}{p_{i_k i_{k+1}}} \frac{a_{i_{k+1} i_{k+2}}}{p_{i_{k+1} i_{k+2}}} \cdots \frac{a_{i_{k+m-1} i_{k+m}}}{p_{i_{k+m-1} i_{k+m}}}$$

- Example: Discounted DP

$$w_{k,m} = \alpha^m, \qquad \forall\, k$$

- LSPE-type method

$$r_{t+1} = r_t + \left(\sum_{k=0}^{t} \phi(i_k)\phi(i_k)' \right)^{-1} \sum_{k=0}^{t} \phi(i_k) \sum_{m=k}^{t} \lambda^{m-k} w_{k,m-k} d_t(i_m),$$

where $d_t(i_m)$ are the temporal differences

$$d_t(i_m) = b_{i_m} + w_{m,1}\phi(i_{m+1})' r_t - \phi(i_m)' r_t, \qquad t \geq 0,\ m \geq 0$$

Forms of λ-Methods II

- Recursive/efficient update for LSPE-type method

$$r_{t+1} = r_t + B_t^{-1}\left(C_t r_t + h_t\right)$$

where

$$B_t = B_{t-1} + \phi(i_t)\phi(i_t)', \qquad C_t = C_{t-1} + z_t\big(w_{t,1}\phi(i_{t+1}) - \phi(i_t)\big)',$$

$$h_t = h_{t-1} + z_t b_{i_t}, \qquad z_t = \lambda w_{t-1,1} z_{t-1} + \phi(i_t).$$

- LSTD(λ)-type method is just

$$r_t = C_t^{-1} h_t$$

- TD(λ)-type method is

$$r_{t+1} = r_t + \gamma_t z_t d_t(i_t)$$

where γ_t is the stepsize

Convergence Result

Proposition: Assume that P is irreducible, and that λ satisfies

$$\lambda \max_{i,j} |a_{ij}|/p_{ij} < 1, \qquad \lambda \in [0,1).$$

Let r_t be generated by the LSTD(λ)-type algorithm. Then,

$$r_t \to r_\lambda^* \qquad \text{with probability 1}$$

The same is true for the LSPE(λ)-type algorithm [assuming also that $\sigma(A^{(\lambda)}) \leq 1$]

- Here r_λ^* is the solution of the projected equation

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

- Similar result for TD(λ)-type extension, under suitable (stochastic approximation-type) conditions for the stepsize

A Nonlinear Equation with Scalar Nonlinearities

- Consider the system

$$x = T(x) = Af(x) + b,$$

  where $f : \Re^n \mapsto \Re^n$ is a mapping with scalar function components of the form $f(x) = \big(f_1(x_1), \ldots, f_n(x_n)\big)$.
- Assume that each of the mappings $f_i : \Re \mapsto \Re$ is nonexpansive:

$$\big|f_i(x_i) - f_i(\bar{x}_i)\big| \leq |x_i - \bar{x}_i|, \qquad \forall\, i = 1, \ldots, n,\ \ x_i, \bar{x}_i \in \Re.$$

  Then if A is a contraction with respect to a weighted Euclidean norm, T is also a contraction
- This structure implies favorable choices of a Markov chain for simulation

Optimal Stopping

- Let $T(x) = \alpha Pf(x) + b$, where P is irreducible transition probability with invariant distribution ξ, $\alpha \in (0,1)$ is a scalar discount factor, and f has components
$$f_i(x_i) = \min\{c_i, x_i\}, \qquad i = 1, \dots, n,$$
where c_i are some scalars.
- Then $x = T(x)$ is the Q-factor equation corresponding to a discounted optimal stopping problem
- In this case, ΠA is a contraction with respect to $\|\cdot\|_\xi$ [Tsitsiklis and Van Roy (1999), who gave a Q-learning algorithm with linear function approximation]
- The LSPE algorithm has been generalized to this problem (Yu and Bertsekas 2007; also the 3rd Edition of my DP text 2007)
- There is no “good" LSTD-type algorithm for this problem (the fixed point equation to be approximated is nonlinear)

Linear Least Squares/Regresion/Bellman Error Methods

- Consider solving the problem

$$\min_{r \in \Re^s} \|A\Phi r - b\|_\xi^2$$

  to approximate the weighted least squares solution of $Ax = b$.
- Here A : $m \times n$ matrix, ξ is a known probability distribution vector, $b \in \Re^m$, and Φ is an $n \times s$ matrix of basis functions.
- The solution is

$$r^* = (\Phi' A' \Xi A \Phi)^{-1} \Phi' A' \Xi b,$$

  where Ξ is the diagonal $m \times m$ matrix having ξ along the diagonal
- To approximate the solution, we replace $\Phi' A' \Xi A \Phi$ and $\Phi' A' \Xi b$ with simulation-based estimates

Issues in Regression/Bellman Error Methods

- Need to sample two columns for each row – more noise
- Variance reduction – a form of importance sampling may be essential
- Dealing with (near) singular $\Phi' A' \Xi A \Phi$
  - Add a small multiple of the identity to $\Phi' A' \Xi A \Phi$ (like a prior in a regression setting), i.e., approximate by simulation

    $$r^* = (\Phi' A' \Xi A \Phi + \gamma I)^{-1} \Phi' A' \Xi b$$

    where γ is small positive parameter
  - Use a proximal method:

    $$r_{t+1} = (\Phi' A' \Xi A \Phi + \gamma_t I)^{-1} (\Phi' A' \Xi b + \gamma_t r_t),$$

    where γ_t is a positive parameter. This converges to the correct solution $(\Phi' A' \Xi A \Phi)^{-1} \Phi' A' \Xi b$
- Applications in inverse problems and other areas (huge dimension - e.g., $n = 10^9$, A: fully dense)

Concluding Remarks

- TD methods can be naturally extended to solve linear systems of equations
- In doing so, perspective and new methods are obtained for approximate DP
- The overall approach is very simple:
  - Start with a deterministic algorithm
  - Write it in terms of expected values
  - Approximate the expected values by simulation
- The approach applies to many linear algebra-type problems - beyond those discussed here (e.g., computing the dominant eigenvalue of a matrix, approximating the invariant distribution of a Markov chain)
- There is considerable literature and theoretical work on Monte Carlo linear algebra methods (starting with von Neumann)
- The new element here is linear function approximation and the connection with TD methods
- Exciting prospect: Application to linear algebra problems of huge dimension, far beyond the DP context