On Temporal Difference Methods and Extensions

Dimitri P. Bertsekas

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

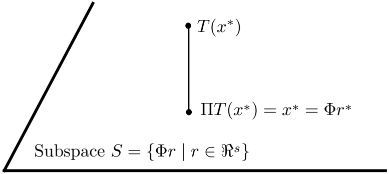
Montreal, June 2009

Focus

- Approximate solution of fixed point problem $x = T(x)$ by solving

$$x = \Pi T(x)$$

Π is projection on a subspace of basis functions (with respect to some weighted Euclidean norm). A special case of Galerkin approximation.


$T(x^*)$
$\Pi T(x^*) = x^* = \Phi r^*$
Subspace $S = \{\Phi r \mid r \in \Re^s\}$

- Traditional TD methods apply to Bellman's equation $x = T(x)$.
- Use Monte-Carlo simulation, which plays an unconventional role.
- An oversimplified view:

TD methods $\approx$ DP with subspace approximation $+$ Simulation

- A more general/extreme view:

TD methods $\approx$ Galerkin Approximation $+$ Monte-Carlo Linear Algebra

Monte-Carlo Linear Algebra

- Key idea: Compute sums $\sum_{i=1}^n a_i$ by simulation when n is large.
- Complexity advantage: Running time is independent of the number n of terms in the sum, only their “variance".
- Introduce a sampling distribution ξ and write

$$\sum_{i=1}^n a_i = \sum_{i=1}^n \xi_i \left(\frac{a_i}{\xi_i}\right) = E_\xi\{\hat a\}$$

where the random variable $\hat a$ has distribution

$$P\left\{\hat a = \frac{a_i}{\xi_i}\right\} = \xi_i, \qquad i = 1, \ldots, n$$

- We “invent" ξ to convert a “deterministic" problem to a “stochastic/simulation problem.

Summary of this Talk

- Starting point: Approximate DP/Bellman's equation/policy evaluation

$$T(x) = Ax + b, \qquad A : n \times n, \quad b \in \Re^n$$

  where A : encodes the Markov chain structure, b : cost vector.

- $x = \Pi T(x)$ is solved by TD methods [TD(λ), LSTD(λ), LSPE(λ)].
- We extend TD methods to general (nonDP) mapping T and general projection on a convex set (rather than a subspace).
- We develop as special cases new TD methods for DP with improved overhead (no matrix inversion).
- We weaken the assumptions under which old methods work (allow linearly dependent basis functions).

References

- D. P. Bertsekas, Dynamic Programming and Optimal Control, Vol. II, 2007, Chapter 6: A “living chapter."
- D. P. Bertsekas and H. Yu, “Projected Equation Methods for Approximate Solution of Large Linear Systems," Journal of Computational and Applied Mathematics, 2009.
- D. P. Bertsekas, “Projected Equations, Variational Inequalities, and Temporal Difference Methods," LIDS Report, MIT, 2009.

Outline

1. Projected Equation Approximation
   - The Approximate DP Context
   - Characteristics of the General Projected Equation Context

2. Unified Framework for Projected Equations
   - Equivalence of Projected Equations and Special Type of VI
   - Iterative Methods for VI
   - Iterative Methods for Projected Equations
   - Convergence without Full Rank Assumption

3. Simulation-Based Versions
   - Simulation Framework
   - LSTD-Type Methods
   - LSPE-Type Methods
   - Scale-Free Convergence Rate Properties

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; α: discount factor
    $A = P$ for average cost problems

Approximate Policy Evaluation

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

  $x \approx \Phi r, \qquad \Phi$ is a matrix with basis functions/features as columns

- Projected Bellman equation:

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

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

Equation Approximation - 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

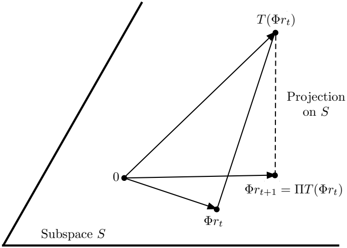
  $$T^{(\lambda)}(x) = (1-\lambda)\sum_{k=0}^{\infty}\lambda^k T^{k+1}(x) = A^{(\lambda)}x + b^{(\lambda)},$$

  where

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

Iterative Methods

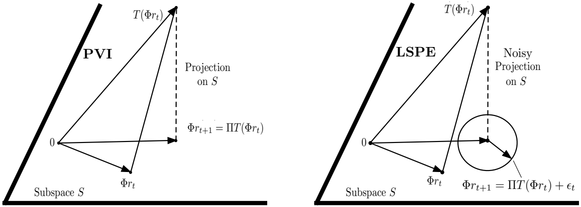
- Projected Value Iteration (PVI)
- Value Iteration => Projection => Value Iteration => Projection ....



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

- Key fact: ΠT is a contraction with respect to 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. Conceptually:

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

- No stepsize unlike TD(λ)
- Allows for a favorable initial guess r_0; may be an advantage in optimistic/few samples approximate policy iteration
- Convergence rate: LSPE “tracks" LSTD, but differs in early stages

Advantages of Projected Equation Methods in DP

When using simulation:

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

General/NonDP Projected Equation Framework

- We consider general projected equations $x = \Pi T(x)$ as approximations to general (nonDP) fixed point equations $x = T(x)$.
- Also more general Euclidean projections (on a convex subset of a subspace S).
- In this talk we focus primarily on linear fixed point problems

$$T(x) = Ax + b$$

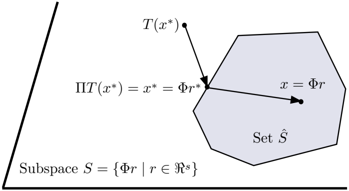
  and projection on a (full) subspace.
- Difference from DP: No Markov chain, no contraction guarantee
- Methods:
  - LSTD analog (does not require ΠA to be a contraction)
  - LSPE analog and scaled versions/extensions (requires ΠA to be a contraction)
  - TD(λ) analog (requires ΠA to be a contraction)
- Advantages maintained: All operations are done in low-dimension and the high-dimensional vector x need not be stored

Connection of Projected Equations and Variational Inequalities

- Consider

$$x = \Pi T(x)$$

where Π is the projection operation onto a closed convex subset $\hat{S}$ of the subspace S (w/ respect to weighted norm $\|\cdot\|_{\Xi}$; Ξ: positive definite).


$T(x^*)$
$\Pi T(x^*) = x^* = \Phi r^*$
$x = \Phi r$
Set $\hat{S}$
Subspace $S = \{\Phi r \mid r \in \Re^s\}$

- From the properties of projection,

$$(x^* - T(x^*))'\Xi(x - x^*) \geq 0, \qquad \forall\, x \in \hat{S}$$

- This is a variational inequality: Find $x^* \in \hat{S}$ such that

$$f(x^*)'(x - x^*) \geq 0, \qquad \forall\, x \in \hat{S},$$

where $f(x) = \Xi\big(x - T(x)\big)$

Equivalence Conclusion

- We have two equivalent problems:
  - The projected equation
  $$x = \Pi T(x)$$
  where Π is projection with respect to $\|\cdot\|_\Xi$ on convex $\hat{S} \subset S$
  - The special form VI
  $$f(\Phi r^*)'\Phi(r - r^*) \geq 0, \qquad \forall\, r \in R,$$
  where
  $$f(x) = \Xi\big(x - T(x)\big), \qquad R = \{r \mid \Phi r \in \hat{S}\}$$
- Every projected equation $x = \Pi T(x)$ is obtained as follows:
  - Start with a suitable VI
  $$f(x^*)'(x - x^*) \geq 0, \qquad \forall\, x \in X,$$
  where X is convex
  - Restrict the solution to be of the form $x = \Phi r$
- Some special cases:
  - $X = \Re^n$: VI <==> $f(x^*) = 0$ (e.g., Bellman's equation in DP)
  - $f(x) = \nabla H(x)$: VI <==> Minimize $H(x)$ over $x \in X$ (e.g., approximate LP)
  - Cooperative and zero-sum games, etc.

Iterative Methods for VI

- Consider the VI

$$f(\Phi r^*)'\Phi(r - r^*) \geq 0, \qquad \forall\, r \in R,$$

  where R is a closed convex set.

- May be solved by iterative methods of the form

$$r_{k+1} = P_{D,R}\big[r_k - \gamma D^{-1}\Phi' f(\Phi r_k)\big],$$

  where γ is a positive stepsize, D is a positive definite symmetric matrix, and $P_{D,R}[\cdot]$ denotes projection on R with respect to norm $\|r\|_D = \sqrt{r'Dr}$.

- Using a classical result: Assume ΠT is a contraction and Φ has linearly independent columns. Then for γ sufficiently small, the method converges to the unique solution r^*.

- Special result: (Bertsekas and Gafni 1982) Assume ΠT is a contraction and R is polyhedral. Then for γ sufficiently small, the method converges at a linear rate to some solution r^* (even without the linear independence assumption on Φ).

Iterative Methods for Projected Linear Equations

- Assume that ΠT is a contraction with respect to $\|\cdot\|_\Xi$ and has fixed point x^*.
- For simplicity, also assume no constraint and T is linear:

$$T(x) = Ax + b$$

- The equivalent VI is $\Phi' f(\Phi r) = 0$ or

$$\Phi' f(\Phi r) = \Phi'\Xi\big(\Phi r - T(\Phi r)\big) = \Phi'\Xi(\Phi r - A\Phi r - b) = 0,$$

  or

$$Cr = d, \qquad \text{(LSTD equation in DP)}$$

  with

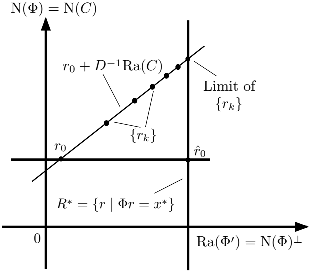
$$C = \Phi'\Xi(I - A)\Phi, \qquad d = \Phi'\Xi b$$

- The iterative method becomes

$$r_{k+1} = r_k - \gamma D^{-1}(Cr_k - d)$$

  and D just scales the direction.

Convergence Properties


$\mathrm{N}(\Phi) = \mathrm{N}(C)$
$r_0 + D^{-1}\mathrm{Ra}(C)$
Limit of $\{r_k\}$
$\{r_k\}$
r_0
$\hat r_0$
$R^* = \{r \mid \Phi r = x^*\}$
0
$\mathrm{Ra}(\Phi') = \mathrm{N}(\Phi)^\perp$

- For γ sufficiently small the iterative method

$$r_{k+1} = r_k - \gamma D^{-1}(Cr_k - d), \qquad C = \Phi'\Xi(I - A)\Phi, \quad d = \Phi'\Xi b$$

  converges at a linear rate:
  - To the unique r^* with $\Phi r^* = x^*$ if Φ has linearly independent columns.
  - To some r^* in the solution set $R^* = \{r \mid \Phi r = x^*\}$ along a linear manifold that passes through r_0 if Φ does not have linearly independent columns.
  - To the unique projection $\hat r_0$ of r_0 onto R^* if $D = I$.
- The high-dimensional sequence Φr_k converges to x^*.

Special Cases

- Projected Value Iteration/Jacobi method

$$D = \Phi'\Xi\Phi, \qquad \gamma \in (0,1],$$

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

  - Requires that Φ has full rank.
  - Important advantage: Known stepsize range for convergence.
  - For $\gamma = 1$ it becomes

    $$x_{k+1} = \Pi T(x_k)$$

    where $x_k = \Phi r_k$.
  - For approximate DP it is equivalent to projected value iteration.
  - It is scale-free: $\{x_k\}$ does not depend on Φ (only on S).

- Simple iteration ($D = I$)

$$r_{k+1} = r_k - \gamma(Cr_k - d)$$

Converges for γ sufficiently small.

- Another low-overhead choice:

D: a diagonal approximation to $\Phi'\Xi\Phi$

Converges for γ sufficiently small, and usually close to 1.

Simulation-Based Versions

- For

$$C = \Phi'\Xi(I - A)\Phi, \qquad d = \Phi'\Xi b$$

with Ξ : diagonal, consider the projected equation

$$Cr = d,$$

and the iteration

$$r_{k+1} = r_k - \gamma D^{-1}(Cr_k - d)$$

- Use k samples to compute simulation-based approximations

$$C_k \sim C, \qquad d_k \sim d$$

- Approximate the projected equation by

$$C_k r = d_k, \qquad \text{(LSTD-type method)}$$

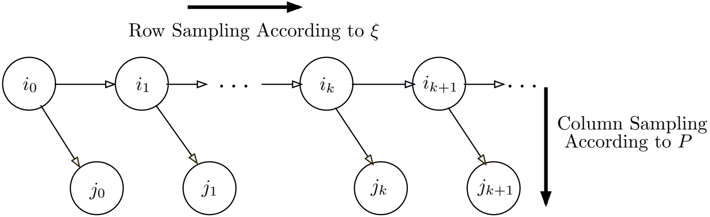
and approximate the iterative method with

$$r_{k+1} = r_k - \gamma D_k^{-1}(C_k r_k - d_k), \qquad \text{(Scaled LSPE-type method)}$$

where

$$D_k \to D > 0$$

Simulation by Row and Column Sampling


Row Sampling According to ξ
i_0
i_1
. . .
i_k
i_{k+1}
. . .
Column Sampling According to P
j_0
j_1
j_k
j_{k+1}

- Row sampling: Generate sequence $\{i_0, i_1, \ldots\}$ according to ξ (the diagonal of Ξ), i.e., relative frequency of each row i is ξ_i
- Column sampling: Generate sequence $\big\{(i_0, j_0), (i_1, j_1), \ldots\big\}$ 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)

Equation Approximation (LSTD-Type) Method

- Approximation of C and d by simulation:

$$C = \Phi'\Xi(I - A)\Phi \quad \sim \quad C_k = \frac{1}{k+1}\sum_{t=0}^{k} \phi(i_t)\left(\phi(i_t) - \frac{a_{i_t j_t}}{p_{i_t j_t}}\phi(j_t)\right)',$$

$$d = \Phi'\Xi b \quad \sim \quad d_k = \frac{1}{k+1}\sum_{t=0}^{k} \phi(i_t) b_{i_t}$$

- We have by law of large numbers $C_k \to C$, $d_k \to d$.
- Equation approximation: Solve the equation $C_k r = d_k$ in place of $Cr = d$.
- If Φ has full rank, C_k is invertible for large k.
- The method is scale-free with respect to features: The high-dimensional sequence $\Phi C_k^{-1} d_k$ does not depend on Φ (only on the subspace S).

Iterative (Scaled LSPE-Type) Method

- Simulation-based iteration

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

where

$$D_k \rightarrow D > 0$$

- Several choices for D_k:
  - Analog of projected value iteration (works with $\gamma = 1$):

$$D_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\phi(i_t)',$$

or for $\beta > 0$,

$$D_k = \frac{1}{k+1} \left(\beta I + \sum_{t=0}^{k} \phi(i_t)\phi(i_t)' \right)$$

  - Version with diagonal approximation to D_k above
  - Simple iteration $D_k = I$

Scale-Free Rate of Convergence

- The choice of D, Φ, and γ affect substantially the convergence rate of the deterministic iteration

$$r_{k+1} = r_k - \gamma D^{-1}(Cr_k - d)$$

- The choices of D_k, Φ, and γ DO NOT affect the convergence rate of the simulation-based version

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

as long as the method converges.

Justification - Two-Time Scale Proof

- Reason: The deterministic iteration

$$r_{k+1} = r_k - \gamma D^{-1}(Cr_k - d)$$

converges fast relative to the speed of the simulation.

- The simulation-based version

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

“sees D_k, C_k, and d_k as essentially constant."

- For any D_k and γ, the sequence $\{\Phi r_k\}$ “tracks" (with prob. 1) the “LDTD" sequence $\Phi C_k^{-1} d_k$ which is scale-free and does not depend on Φ.

Relation to TD(λ)

- If in the simple method ($D_k = I$) we use a single sample approximation to C_k and d_k:

$$C_k = \phi(i_k)\left(\phi(i_k) - \frac{a_{i_k j_k}}{p_{i_k j_k}}\phi(j_k)\right)', \qquad d_k = \phi(i_k) b_{i_k}$$

  we obtain TD(0) (generalized for nonDP fixed point problems).

- It takes the form

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

  where γ_k must be diminishing for convergence (to “average" the simulation noise), and

$$C_k r_k - d_k = \phi(i_k) \cdot \text{(the TD)}$$

- An extension with direction scaling (Choi and VanRoy, 2006)

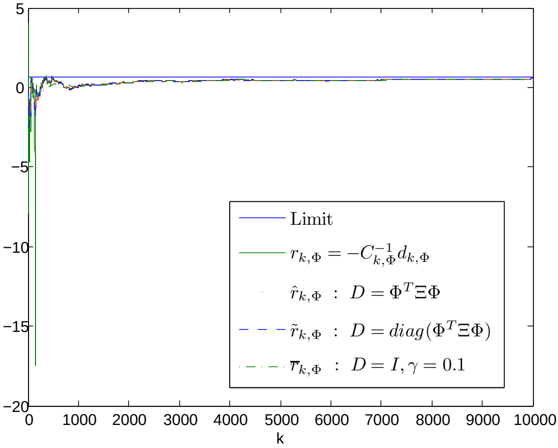
$$r_{k+1} = r_k - \gamma_k D_k^{-1} (C_k r_k - d_k)$$

- If C_k and d_k are approximations to $C^{(\lambda)}$ and $d^{(\lambda)}$, we obtain (extensions of) TD(λ).

Rate of Convergence of Low-Dimensional Sequences $\{r_k\}$

For any D_k and γ:

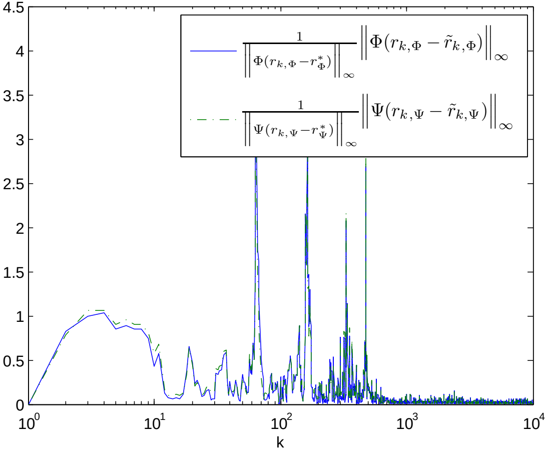
The low-dimensional scaled LSPE-type iterates track the LSTD-type iterates.



Rate of Convergence of High-Dimensional Sequences $\{\Phi r_k\}$

For any D_k, γ, and feature representation of S:

The high-dimensional scaled LSPE-type iterates track the high-dimensional LSTD-type iterates (which do not depend on feature scaling).



Concluding Remarks re NonDP Problems

- TD methods can be naturally extended to solve more general (nonDP) problems with basis function approximation.
- This leads to a Monte-Carlo Galerkin approximation methodology. A vast area of applications, e.g., operator equations, PDEs, inverse problems, boundary-value problems, regression, optimization, etc.
- The main advantage is solving (approximately) large-dimensional problems with low-order calculations.
- Unification through a connection with VIs.
- The overall approach is simple:
  - Start with a VI in high-dimension x (e.g., linear equation, fixed point problem, regression, optimization, game problem, etc)
  - Do basis function approximation $x \approx \Phi r$
  - Pick a deterministic (direct or iterative) method for the resulting low-dimension VI
  - Write it in terms of inner products/expected values
  - Approximate the expected values by simulation
- Important issues: Clever implementation, convergence analysis, efficient simulation, variance reduction, constraint sampling and/or aggregation.

Concluding Remarks re DP

- New iterative TD methods (scaled LSPE) have been obtained.
- Their rate of convergence is scale-free (does not depend on direction scaling matrix D, stepsize γ, and feature matrix Φ) – they all track the (scale-free) sequence generated by LSTD.
- With diagonal scaling the overhead per iteration is improved over LSTD/LSPE (no matrix inversion).
- For convergence and rate of convergence the full rank assumption on Φ is immaterial.
  - TD(λ) will converge to the unique projection of the starting weights r_0 on the manifold of solutions.
  - Scaled LSPE will converge to some (random) solution.