New Exact and Approximate Policy Iteration Methods in Dynamic Programming

Dimitri P. Bertsekas

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

March 2011

Summary

- Discounted infinite horizon DP problems/Classical value and policy iteration
- Optimistic/modified policy iteration (policy evaluation is approximate, with a finite number of value iterations using the current policy)
- Convergence issues for synchronous and asynchronous versions
- Failure of asynchronous/modified policy iteration (Williams-Baird counterexample)
- A radical modification of policy iteration/evaluation: Aim to solve an optimal stopping problem instead of solving a linear system
- Convergence properties are restored/enhanced
- Optimistic policy iteration/Q-learning with cost function approximation, exploration enhancement, and approximate solution of optimal stopping problems
- Generalizations and abstractions (multi-agent aggregation, concave fixed point problems)

References

- Starting Point: D. P. Bertsekas and H. Yu, “Q-Learning and Enhanced Policy Iteration in Discounted Dynamic Programming," Report LIDS-P-2831, MIT, April 2010
- Emphasis in this talk: D. P. Bertsekas and H. Yu, “Distributed Asynchronous Policy Iteration," Proc. Allerton Conference, Sept. 2010
- Line of analysis: Theory of totally asynchronous distributed algorithms from
  - D. P. Bertsekas, "Distributed Dynamic Programming," IEEE Transactions on Aut. Control, Vol. AC-27, 1982
  - D. P. Bertsekas, "Distributed Asynchronous Computation of Fixed Points," Mathematical Programming, Vol. 27, 1983
  - D. P. Bertsekas and J. N. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, 1989

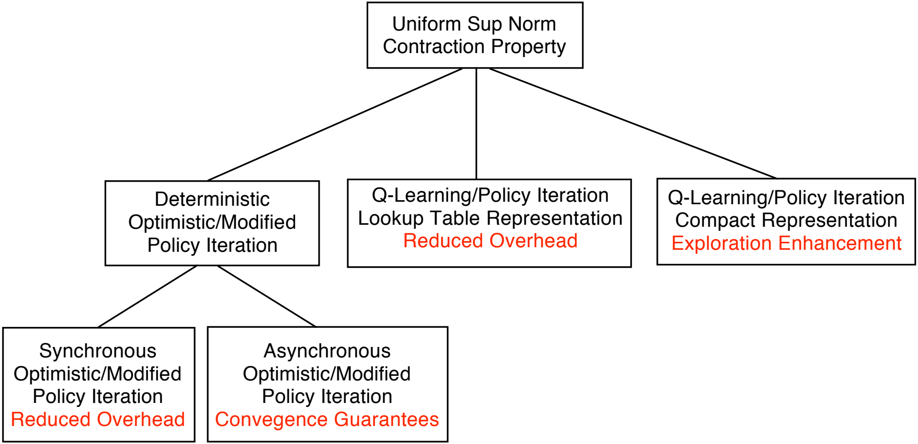
A More Abstract View of What Follows

- We want to find a fixed point J^* of a mapping $T : \Re^n \mapsto \Re^n$ of the form

$$(TJ)(i) = \min_{\mu \in \mathcal{M}_i} (T_\mu J)(i), \qquad i = 1, \ldots, n,$$

  where μ is a parameter from some set $\mathcal{M}_i$.
- We update J in two ways:
  - Iterate with T: $J \mapsto TJ$ (cf. value iteration/DP), OR
  - Pick a μ and iterate with T_μ: $J \mapsto T_\mu J$ (cf. policy evaluation/DP)
- Difficulty: T_μ has different fixed point than T ... so iterations with T_μ aim at a target other than J^*
- Our key idea (abstractly): Embed both T and T_μ within another (uniform) contraction mapping F_μ that has the same fixed point for all μ
- The uniform contraction mapping F_μ operates on the larger space of Q-factors
- In the DP context, F_μ is associated with an optimal stopping problem
- Most of what follows applies beyond DP

A High Level View of Research Directions



Outline

Dynamic Programming - Markovian Decision Problems (MDP)

- System: Controlled Markov chain w/ transition probabilties $p_{ij}(u)$
- States: $i = 1, \ldots, n$
- Controls: $u \in U(i)$
- Cost per stage: $g(i, u, j)$
- Stationary policy: State to control mapping μ; apply $\mu(i)$ when at state i
- Discounted MDP: Find policy μ that minimizes the expected value of the infinite horizon cost:

$$\sum_{k=0}^{\infty} \alpha^k g\big(i_k, \mu(i_k), i_{k+1}\big)$$

where

$$i_k = \text{ state at time } k, \qquad i_{k+1} = \text{ state at time } k+1,$$

$$\alpha : \text{ discount factor } \quad 0 < \alpha < 1$$

Major DP Results

- $J^*(i)$ = Optimal cost starting from state i
- $J_\mu(i)$ = Optimal cost starting from state i using policy μ
- Bellman's equation:

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

  A system of n nonlinear equations in the unknowns $J^*(1),\dots,J^*(n)$.

- J^* is its unique solution.
- An optimal policy minimizes for each i in the RHS of Bellman's equation.
- Bellman's equation for a policy μ:

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

- It is a linear system of equations with J_μ as its unique solution.

Shorthand Notation - Fixed Point View

- Denote by T and T_μ the mappings that transform $J \in \Re^n$ to the vectors TJ and $T_\mu J$ with components

$$(TJ)(i) \overset{\text{def}}{=} \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + \alpha J(j)\big), \qquad i = 1, \ldots, n,$$

and

$$(T_\mu J)(i) \overset{\text{def}}{=} \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big)\big(g(i,\mu(i),j) + \alpha J(j)\big), \qquad i = 1, \ldots, n$$

- Bellman's equations are written as

$$J^* = TJ^*, \qquad J_\mu = T_\mu J_\mu$$

- Key structure for our purposes: T and T_μ are sup-norm contractions with common modulus α:

$$\|TJ - TJ'\|_\infty = \max_{i=1,\ldots,n} \big|(TJ)(i) - (TJ')(i)\big| \le \alpha \max_{i=1,\ldots,n} \big|J(i) - J'(i)\big| = \alpha \|J - J'\|_\infty$$

Major Methods for Finding Fixed Point of T

- Value iteration (generic fixed point method): Start with any J^0, iterate by

$$J^{t+1} = TJ^t$$

- Policy iteration (special method for T of the form $T = \min_\mu T_\mu$): Start with any J^0 and μ^0. Given J^t and μ^t, iterate by:
  - Policy evaluation: $J^{t+1} = (T_{\mu^t})^m J^t$ (m applications of T_{μ^t} on J^t; $m = \infty$ is possible)
  - Policy improvement: μ^{t+1} attains the min in TJ^{t+1} (or $T_{\mu^{t+1}} J^{t+1} = TJ^{t+1}$)

- Both methods converge to J^*:
  - Value iteration, thanks to contraction of T
  - Policy iteration, thanks to contraction and monotonicity of T and T_μ

- Typically, (optimistic/modified) policy iteration (with a reasonable choice of m) is more efficient because application of T_μ is cheaper than application of T

Convergence Issues in Optimistic/Modified Policy Iteration

- Classical convergence result assumes monotonicity of initial condition:

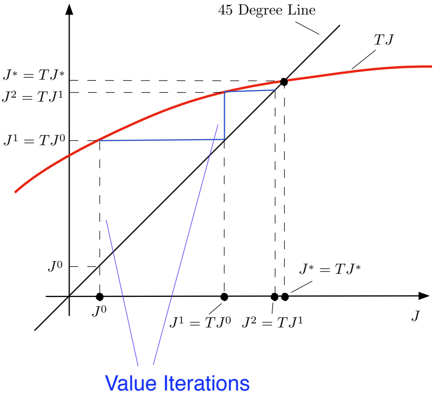
$$T_{\mu_0} J^0 \leq J^0 \tag{1}$$

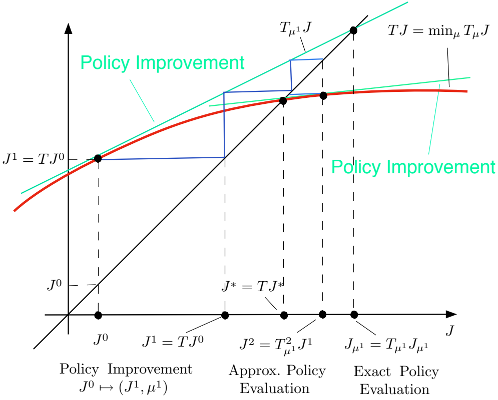
  - For a discounted MDP problem, this condition is not needed (a fortunate consequence of structure)
  - For other types of DP problems, situation unclear
  - For example: If the policy evaluations are done in Gauss-Seidel cyclic fashion (one state at a time), the situation is unclear

- Williams-Baird Example: Convergence fails if condition (1) does not hold, and the policy evaluations and policy improvements are (a little) less regular than Gauss-Seidel

- Williams and Baird prove that asynchronous policy iteration converges monotonically from above under condition (1)

Graphical Interpretations





Distributed Asynchronous Framework for Fixed Point Computation

- Consider solution of general fixed point problem $J = TJ$, or

$$J(i) = T_i\big(J(1), \dots, J(n)\big), \qquad i = 1, \dots, n$$

- We have a network of processors, and w/out loss of generality assume that there is a separate processor i for each component $J(i)$, $i = 1, \dots, n$

- Processor i updates $J(i)$ at a subset of times $\mathscr{T}_i \subset \{0, 1, \dots\}$
- Processor i receives (possibly outdated values) $J(j)$ from other processors $j \neq i$
- Update of processor i (no "delays")

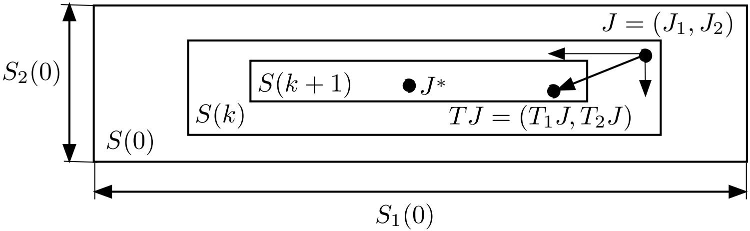
$$J^{t+1}(i) = \begin{cases} T_i\big(J^t(1), \dots, J^t(n)\big) & \text{if } t \in \mathscr{T}_i, \\ J^t(i) & \text{if } t \notin \mathscr{T}_i. \end{cases}$$

- Update of processor i [with "delays" $t - \tau_{ij}(t)$]

$$J^{t+1}(i) = \begin{cases} T_i\big(J^{\tau_{i1}(t)}(1), \dots, J^{\tau_{in}(t)}(n)\big) & \text{if } t \in \mathscr{T}_i, \\ J^t(i) & \text{if } t \notin \mathscr{T}_i. \end{cases}$$

Distributed Convergence of Fixed Point Iterations

A general theorem for “totally asynchronous" iterations, i.e., $\mathscr{T}_i$ are infinite sets and $\tau_{ij}(t) \to \infty$ as $t \to \infty$ (Bertsekas, 1983)



- Assume there is a nested sequence of sets $S(k+1) \subset S(k)$ such that
  - (Synchronous Convergence Condition) We have

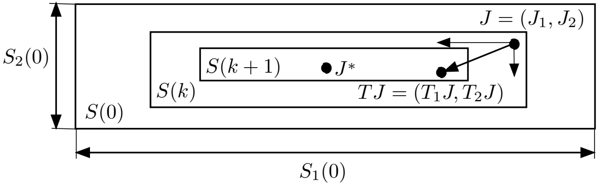
    $$TJ \in S(k+1), \qquad \forall\, J \in S(k),$$

    and the limit points of all sequences $\{J^k\}$ with $J^k \in S(k)$, for all k, are fixed points of T.
  - (Box Condition) $S(k)$ is a Cartesian product:

    $$S(k) = S_1(k) \times \cdots \times S_n(k)$$

Then, if $J^0 \in S(0)$, every limit point of $\{J^t\}$ is a fixed point of T.

Applications of the Theorem



Major contexts where the theorem applies:

- T is a sup-norm contraction with fixed point J^* and modulus α:

$$S(k) = \{J \mid \|J - J^*\|_\infty \le \alpha^k B\}, \qquad \text{for some scalar } B$$

- T is monotone ($TJ \le TJ'$ for $J \le J'$) with fixed point J^* and for some $\underline{J}$ and $\overline{J}$ with

$$\underline{J} \le T\underline{J} \le T\overline{J} \le \overline{J},$$

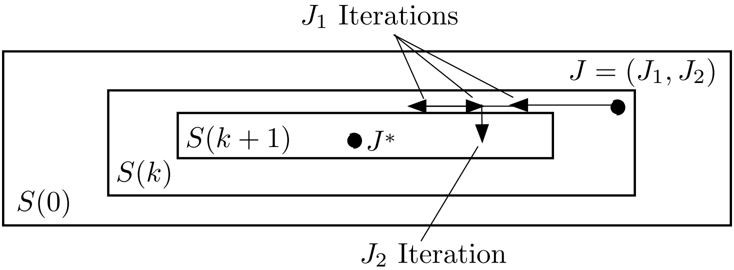
  and $\lim_{k\to\infty} T^k \underline{J} = \lim_{k\to\infty} T^k \underline{J} = J^*$, we have

$$S(k) = \{J \mid T^k \underline{J} \le J \le T^k \overline{J}\}$$

Both of these apply to various DP problems:

- 1st context applies to discounted problems
- 2nd context applies to undiscounted problems (e.g., shortest paths)

Distributed Asynchronous Convergence for Value Iteration



- **Value Iteration**: Start with any J^0. Given J^t, for all i, iterate at i by

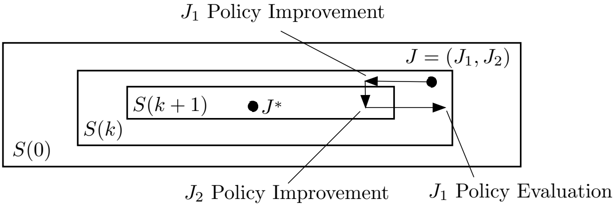
$$J^{t+1}(i) = (TJ^t)(i) \qquad \text{if } t \in \mathscr{T}_i$$

and set $J^{t+1}(i) = J^t(i)$ otherwise, where

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

- T is a sup-norm contraction with contraction modulus α
- Totally asynchronous distributed convergence (including communication delays) is obtained

Difficulties of Asynchronous Convergence for Policy Iteration



Policy iteration: Start with any J^0 and μ^0. Given J^t and μ^t, iterate as follows:

- If $t \in \mathscr{T}_i$, do a policy evaluation at i: $J^{t+1}(i) = (T_{\mu^t})^m J^t(i)$
- If $t \in \overline{\mathcal{T}}_i$, do a policy improvement at i: Set $J^{t+1}(i) = (TJ^t)(i)$ and let μ^{t+1} be the policy that attains the min in TJ^t (i.e., $T_{\mu^{t+1}} J^t = TJ^t$)

Difficulties:

- We iterate with both T and T_μ
- All these mappings are sup-norm contractions
- But they have different fixed points (J^* and J_μ)
- Policy improvement operates with a different set sequence $\{S(k)\}$ than policy evaluation

Failure of Asynchronous Policy Iteration: W-B Example

Counterexample by Williams and Baird (1993)

- Deterministic discounted MDP with 6 states arranged in a circle
- 2 controls available in half the states, 1 control available in the other half
- Policy evaluations and improvements are one state at a time, no “delays"
- A cycle of 15 iterations is constructed that repeats the initial conditions
- The order of iterated states in the cycle is “maliciously" constructed
- It is unknown whether it is possible to construct a counterexample where the order of iterated states is random

Rectifying the Difficulty - Q-Factors

- Q-factors, $Q(i,u)$ are functions of state-control pairs (i,u)
- The optimal Q-factors are given by

$$Q^*(i,u) = \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + \alpha J^*(j)\big)$$

$Q^*(i,u)$: Cost of starting at i, using u first, then use optimal policy.

- They satisfy

$$J^*(i) = \min_{u \in U(i)} Q^*(i,u)$$

- These are Bellman's equations in expanded MDP with states (i,u), i
- The Q-factors of a policy μ are the unique solution of

$$Q_\mu(i,u) = \sum_{j=1}^{n} p_{ij}(u)\big(g(i,u,j) + \alpha Q_\mu(j,\mu(j))\big)$$

$Q_\mu(i,u)$: Cost of starting at i using u in the first step, then use μ.
Also $Q_\mu(i,\mu(i)) = J_\mu(i)$

Sup-Norm Uniform Contraction Property

- Consider Q-factors $Q(i,u)$ and costs $J(i)$. For any μ, define mapping

$$(Q,J) \quad \mapsto \quad \big(F_\mu(Q,J), M_\mu(Q,J)\big)$$

where

$$F_\mu(Q,J)(i,u) \stackrel{\text{def}}{=} \sum_{j=1}^{n} p_{ij}(u)\Big(g(i,u,j) + \alpha \min\big\{J(j), Q(j,\mu(j))\big\}\Big),$$

$$M_\mu(Q,J)(i) \stackrel{\text{def}}{=} \min_{u\in U(i)} F_\mu(Q,J)(i,u)$$

- Key fact: This mapping is a uniform sup-norm contraction - a common fixed point (Q^*,J^*) for all μ
- We have

$$\max\big\{\|F_\mu(Q,J)-Q^*\|_\infty,\ \|M_\mu(Q,J)-J^*\|_\infty\big\} \le \alpha \max\big\{\|Q-Q^*\|_\infty,\ \|J-J^*\|_\infty\big\}$$

- The mapping is convergent under asynchronous iteration
- Even though we operate with different mappings corresponding to different μ, they all have a common fixed point

Connection to an Optimal Stopping Problem

- Consider policy iteration using

$$(Q, J) \quad \mapsto \quad \big(F_\mu(Q, J), M_\mu(Q, J)\big)$$

where

$$F_\mu(Q, J)(i, u) \stackrel{\text{def}}{=} \sum_{j=1}^{n} p_{ij}(u)\big(g(i, u, j) + \alpha \min\{J(j), Q(j, \mu(j))\}\big),$$

$$M_\mu(Q, J)(i) \stackrel{\text{def}}{=} \min_{u \in U(i)} F_\mu(Q, J)(i, u)$$

- For fixed J and μ the fixed point of $F_\mu(\cdot, J)$ is the optimal cost of an optimal stopping problem [$J(j)$ is the stopping cost at j]
- Iteration with $F_\mu(\cdot, J)$ for fixed J and μ, aims to solve the stopping problem associated with J and μ
- Iteration with $M_\mu(\cdot, J)$, does a “value iteration/policy improvement" to update the stopping problem

Distributed Asynchronous Policy Iteration

Asynchronous distributed policy iteration algorithm: Maintains J^t, μ^t, and V^t, where

$$V^t(i) = Q^t\big(i, \mu^t(i)\big) \qquad \text{Q-factors of current policy}$$

- Let $\mathscr{T}_i$ (or $\overline{\mathscr{T}}_i$) be the policy evaluation (or policy improvement) times at state i.
- At time t, for all i,
  - If $t \in \mathscr{T}_i$, do a policy evaluation at i: Set

$$V^{t+1}(i) = \sum_{j=1}^{n} p_{ij}\big(\mu^t(i)\big)\Big(g\big(i, \mu^t(i), j\big) + \alpha \min\big\{J^t(j), V^t(j)\big\}\Big)$$

    and leave $J^t(i)$, $\mu^t(i)$ unchanged.
  - If $t \in \overline{\mathscr{T}}_i$, do a policy improvement at i: Set

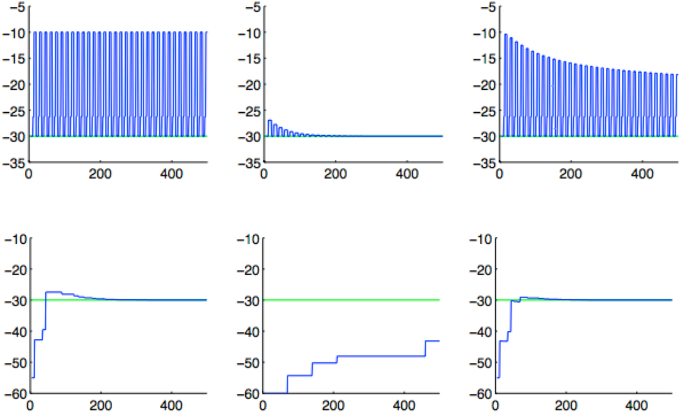
$$J^{t+1}(i) = V^{t+1}(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u)\Big(g(i, u, j) + \alpha \min\big\{J^t(j), V^t(j)\big\}\Big),$$

    and set $\mu^{t+1}(i)$ to a control that attains the minimum.
- Convergence follows by the asynchronous convergence theorem

Some Computational Experiments

Williams-Baird Counterexample

"Classical" Algorithm | **New Algorithm** | **Interpolated Variant**



Malicious Order of Component Selection

Random Order of Component Selection

Several Interlocking Research Directions

The idea of embedding into a stopping problem applies to several research contexts:

- Optimistic/modified policy iteration for costs: synchronous or asynchronous
- Optimistic Q-learning: Stochastic asynchronous policy iteration for Q-factors with function approximation (can modify μ at will to enhance exploration)
- Algorithmic variations that work without sup-norm contraction - assume just monotonicity
- Optimistic (synchronous and asynchronous) policy iteration for stochastic shortest path and other nondiscounted problems
- Multi-agent aggregation in DP
- General forms of mappings T and T_μ for other types of DP and nonDP problems, under sup-norm contraction assumptions. The discounted DP structure is not critical, sup-norm contraction is
- NonDP fixed point problems involving concave sup-norm contractions

Stopping Problem-Based Optimistic Q-learning

- Use for (nondistributed) policy iteration where policy evaluation is done by solving a stopping problem
- Given (Q^t, J^t) and μ^t, iterate by:
  - Policy evaluation: $Q^{t+1} = F_{\mu^t}^m(Q^t, J^t)$ (m applications of F_{μ^t} on Q^t with J^t kept fixed) - connection to a stopping problem
  - Policy improvement: $J^{t+1} = (MQ^{t+1})$ and set μ^{t+1} to the policy that attains the min
- Contraction property is uniform for all policies
- We may use randomized policies μ that induce exploration
- We may use simulation-based implementations and lookup-table or compact representations, and the TD algorithm of Tsitsiklis and VanRoy (1999) to solve the optimal stopping problems
- Error bounds are available thanks to the uniform contraction property

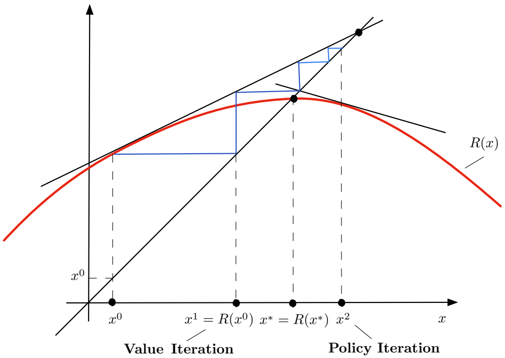
Generalized Mappings T and T_μ

- The preceding analysis uses only the contraction property of the discounted MDP (not monotonicity or the probabilistic structure)
- Abstract Mappings T and T_μ:
  - Introduce a mapping $H(i,u,J)$ and denote
  $$(TJ)(i) = \min_{u \in U(i)} H(i,u,J), \qquad (T\mu J)(i) = H\big(i,\mu(i),J\big)$$
  i.e., $TJ = \min_\mu T_\mu J$, where the min is taken separately for each component
  - Assume that for all i and $u \in U(i)$
  $$\big|H(i,u,J) - H(i,u,J')\big| \le \alpha \|J - J'\|_\infty$$
- Asynchronous "policy iteration" algorithm: At time t, for all i:
  - If $t \in \mathscr{T}_i$, do a "policy evaluation" at i: Set
  $$V^{t+1}(i) = H\big(i, \mu^t(i), \min\{J^t, V^t\}\big)$$
  and leave $J^t(i)$, $\mu^t(i)$ unchanged.
  - If $t \in \overline{\mathscr{T}}_i$, do a "policy improvement" at i: Set
  $$J^{t+1}(i) = V^{t+1}(i) = \min_{u \in U(i)} H\big(i, u, \min\{J^t, V^t\}\big)$$
  set $\mu^{t+1}(i)$ to a u that attains the minimum.

DP Applications with Generalized T and T_μ

- DP models beyond discounted with standard policy evaluation
  - Gauss-Seidel version of optimistic/modified policy iteration for discounted problems
  - Optimistic/modified policy iteration for semi-Markov and minimax discounted problems
  - Stochastic shortest path problems
  - Q-learning versions of the above
- Multi-agent aggregation
  - Each agent updates costs at all states within a subset
  - Each agent uses detailed costs for the local states and aggregate costs for other states, as communicated by other agents

Fixed Points of Parametric Sup-Norm Contractions



- Find a fixed point of a mapping $R : \Re^n \mapsto \Re^n$, i.e. $x^* = R(x^*)$
- Special case: The components $R_i(\cdot) : \Re^n \mapsto \Re$ are concave sup-norm contractions
- A policy iteration algorithm can be used: policy evaluation corresponds to linearization (like Newton's method)

A Variant with Interpolation

- Modify the mapping F_μ with a stepsize parameter $\gamma \in [0,1)$:

  $$F_{\mu,\gamma}(Q,J) = H\big(i,u,W_\gamma(J,Q\mu)\big)$$

  where

  $$W_\gamma(J,Q\mu) = (1-\gamma)\min\{J,Q_\mu\} + \gamma Q_\mu$$

  and $Q_\mu(i) = Q\big(i,\mu(i)\big)$

- Asynchronous “policy iteration" algorithm: At time t, for all i:
  - If $t \in \mathscr{T}_i$, do a “policy evaluation" at i: Set

    $$V^{t+1}(i) = H\big(i,\mu^t(i),W_{\gamma^t}(J^t,V^t)\big)$$

    and leave $J(i)$ and $\mu(i)$ unchanged
  - If $t \in \overline{\mathscr{T}}_i$, do a “policy improvement" at i: Set

    $$J^{t+1}(i) = V^{t+1}(i) = \min_{u\in U(i)} H\big(i,u,W_{\gamma^t}(J^t,V^t)\big)$$

    set $\mu^{t+1}(i)$ to a u that attains the minimum.

- If $\gamma^t \to 0$, the algorithm converges asynchronously

Asynchronous Policy Iteration Under Monotonicity Assumptions

If H is not sup-norm contraction, we may use monotonicity properties. Assume:

(a) The mapping H is monotone in the sense that

$$H(i, u, J) \leq H(i, u, J'), \qquad \forall\, i,\, u \in U(i)$$

for all J, J' from a set of vectors $\mathcal{F}$ such that $J \leq J'$.

(b) There exist two vectors $\underline{J}$, $\overline{J} \in \mathcal{F}$ such that all $J \in F$ with $\underline{J} \leq J \leq \overline{J}$ belong to $\mathcal{F}$, and we have $\underline{J} \leq T\underline{J} \leq T\overline{J} \leq \overline{J}$. Furthermore, T has a unique fixed point J^* and

$$\lim_{k\to\infty} T^k \underline{J} = \lim_{k\to\infty} T^k \overline{J} = J^*$$

- Assuming J^0 satisfies $\underline{J} \leq J^0 \leq \overline{J}$, value iteration still converges in a distributed, totally asynchronous setting
- Asynchronous policy iteration needs to be corrected for convergence
  - Policy evaluation equation at i is changed to

    $$V^{t+1}(i) = \min\left\{J^t(i),\ H\big(i, \mu^t(i), V^t\big)\right\}$$

    (the min is outside of H)
  - Totally asynchronous convergence can be shown

Concluding Remarks

- Optimistic/modified/asynchronous policy iteration has fragile convergence properties
- We have provided a new approach and several algorithmic variants to correct the difficulties
- Key idea: Embed the problem into one that involves Q-factors/Q-learning and admits an underlying uniform sup-norm contraction
- Can be implemented by replacing the linear system used for policy evaluation with an optimal stopping problem
- Can work with cost function approximation and allows enhanced exploration
- Extensions to generalized DP models involving H:
    - They validate (distributed and nondistributed) optimistic/modified policy iteration for more general than discounted DP models (e.g., stochastic shortest path, semi-Markov, etc)
    - They provide algorithms for finding fixed points of nonDP mappings of the form $T = min_{\mu \in \mathcal{M}} T_\mu$

Thank You!