Optimistic Policy Iteration and Q-learning in Dynamic Programming

Dimitri P. Bertsekas

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

November 2010

INFORMS, Austin, TX

Summary

- Policy iteration in infinite horizon DP
  - Maintains cost-policy pair (J^t, μ^t)
  - J^t is obtained by "policy evaluation" of μ^t (need to solve a linear system)
  - μ^{t+1} is obtained by "policy improvement" based of J^t
- Focus on "optimistic" policy iteration (also known as "modified")
  - Policy evaluation is approximate: a finite number of value iterations using μ^t
  - More efficient in practice
  - Has fragile convergence properties
  - Requires a monotonicity assumption for initial condition: $T_{\mu^0} J^0 \leq J^0$
  - Could be asynchronous: one state at a time, in any order, with "delays"
- Failure of asynchronous/optimistic policy iteration without the monotonicity condition (Williams-Baird counterexample -1993)
- A radical modification of policy evaluation: Aims to solve an optimal stopping problem instead of solving a linear system
- Convergence properties are restored/improved
- We obtain an optimistic exploration-enhanced Q-learning algorithm

References

- D. P. Bertsekas and H. Yu, “Q-Learning and Enhanced Policy Iteration in Discounted Dynamic Programming," Report LIDS-P-2831, MIT, April 2010
- D. P. Bertsekas and H. Yu, “Distributed Asynchronous Policy Iteration," Proc. Allerton Conference, Sept. 2010 (describes slightly different algorithms than these slides)

- Related lines of analysis:
    - Theory of totally asynchronous distributed algorithms from Bertsekas 1982, 1983, and Bertsekas and Tsitsiklis 1989
    - Generalized/abstract DP model: From Bertsekas 1977, and Bertsekas and Shreve 1978

Outline

Discounted MDP - Fixed Point View

- $J^*(i)$ = Optimal cost starting from state i
- $J_\mu(i)$ = Cost starting from state i using policy μ
- Denote by T and T_μ the mappings that transform $J \in \Re^n$ to the vectors TJ and $T_\mu J$ with components

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

and

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

- Bellman's equations are written as

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

- Key structure: T and T_μ are sup-norm contractions,

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

Finding Fixed Point of T: Major Methods

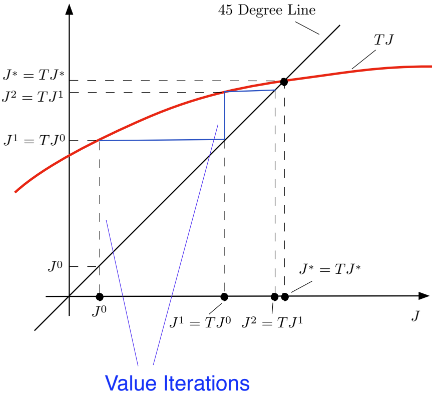
- Value iteration (generic fixed point method): Start with any J^0, iterate by

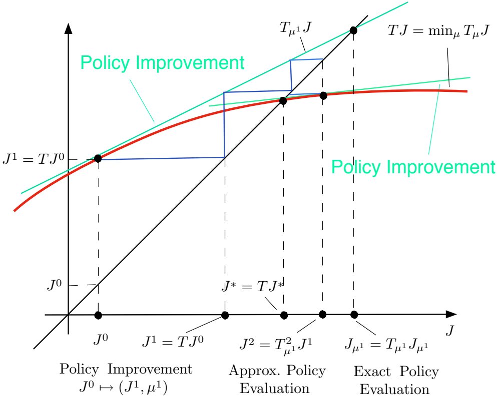
$$J^{t+1} = TJ^t$$

- 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 $T_{\mu^{t+1}} J^{t+1} = TJ^{t+1}$)
- Policy iteration is more efficient because application of T_μ is cheaper than application of T (typically, with a reasonable choice of m)
- Value iteration converges to J^*, thanks to contraction property of T
- It converges in distributed asynchronous form, thanks to sup-norm contraction and monotonicity of T
- Policy iteration converges asynchronously, thanks to sup-norn contraction and monotonicity of T and T_μ, assuming monotonicity of initial condition:

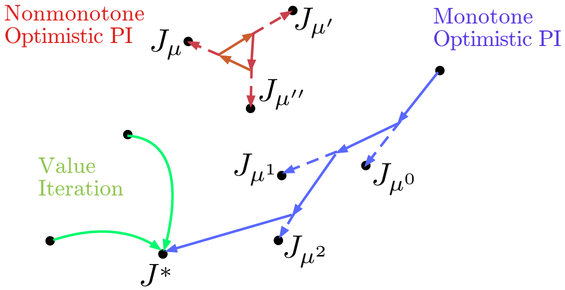
$$T_{\mu^0} J^0 \leq J^0$$

Value and Policy Iteration: Graphical Interpretations





An Abstract View of the Convergence Issue



- 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 $\mathcal{M}_i$.
- Instead of T, we iterate with a sequence of mappings T_{μ^k}, (which change when there is a policy improvement)
- Difficulty: T_μ has different fixed point than T ... so the target of the iterations keeps changing

An Abstract View of Our Approach

- Embed both T and T_μ within another mapping F_μ
- F_μ has the same fixed point for all μ from which J^* can be extracted
- F_μ is sup-norm contraction, so convergence is obtained (also in a distributed asynchronous context)
- In the DP context, F_μ is associated with an optimal stopping problem
- Because it is not crucial which μ we use, we can modify μ to effect exploration enhancement - major issue in simulation-based policy iteration
- Most of what follows applies beyond α-discounted DP

Embedding to a Uniform Sup-Norm Contraction

- 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

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

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

- Key fact: This mapping is a uniform sup-norm contraction - a common fixed point (Q^*, J^*) for all μ, where $J^*(i) = \min_{u\in U(i)} Q^*(i,u)$
- We have

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

- Fixed point iteration with this mapping converges asynchronously
- We operate with different mappings corresponding to different μ, but they all have a common fixed point

Connection to an Optimal Stopping Problem

- Consider the mapping

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

where

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

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

- For fixed J and μ the fixed point of $F_\mu(\cdot, J)$ is the optimal cost of an optimal stopping problem [transitions: $(i, u) \mapsto (j, \mu(j))$, stopping cost at j: $J(j)$]
- 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

Special Case: Optimistic Policy Iteration with Improved Convergence

- Maintain J^t, μ^t, and $V^t(i) = Q(i, \mu^t(i))$ (not necessary to maintain the entire vector Q)
  - If $t \in \mathscr{T}_i$, do a "policy evaluation" at i: Set

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

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

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

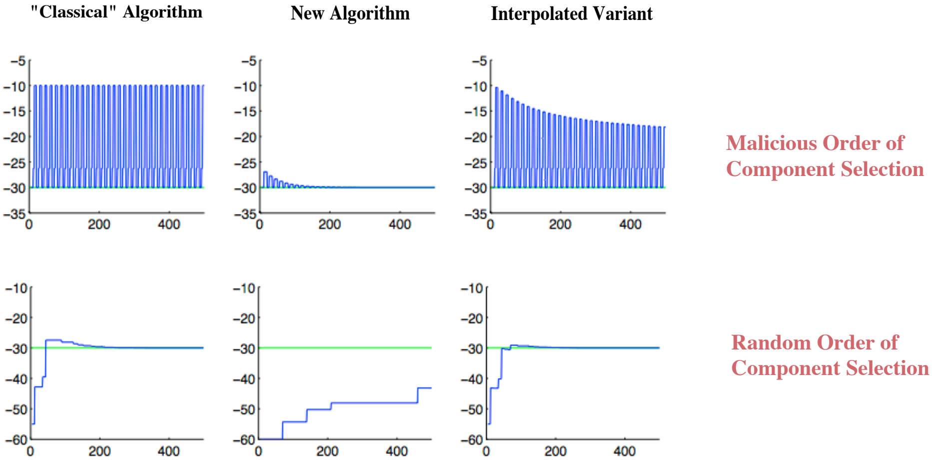
    set $\mu^{t+1}(i)$ to a u that attains the minimum.
- We restrict the increases of V^t in policy evaluations (using J^t as a "stopping" cost)
- A variant with interpolation: In place of $\min\{J^t, V^t\}$ use

$$(1 - \gamma^t)\min\{J^t, V^t\} + \gamma^t V^t$$

  when $J^t < V^t$, with $\gamma^t \downarrow 0$.

Some Computational Experiments (Using the slightly different algorithms of the Allerton conference paper)

Williams-Baird Counterexample



Exploration-Enhanced Model-Based Policy Iteration

- We may replace the current policy μ with a randomized policy ν

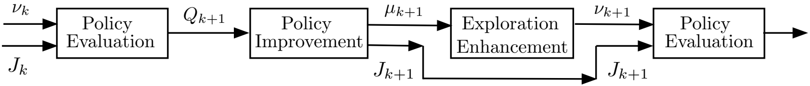
$$\{\nu(u \mid i) \mid u \in U(i)\}$$

  which provides exploration

- We use the map $Q \to F_{J,\nu}Q$, the vector of Q-factors with components

$$(F_{J,\nu}Q)(i,u) = \sum_{j=1}^{n} p_{ij}(u)\left(g(i,u,j) + \alpha \sum_{v \in U(j)} \nu(v \mid j) \min\{J(j), Q(j,v)\}\right)$$

- The randomized ν may be related to the current μ but may include unlimited amount of exploration



- The preceding uniform contraction analysis and algorithms generalize

Exploration-Enhanced Model-Free Q-Learning

- Select a state-action pair (i_k, u_k)
- Policy improvement (for k in a selected subset of times): Update J_k, μ_k according to

$$J_{k+1}(i_k) = \min_{u \in U(i_k)} Q_k(i_k, u), \quad \mu_{k+1}(j) = \arg \min_{u \in U(j)} Q_k(i_k, u), \qquad \text{for } i = i_k$$

  For $i \neq i_k$, leave $J_k(i)$ and $\mu_k(i)$ unchanged
- Policy evaluation (for all k): Select a stepsize $\gamma_{(i_k,u_k),k} \in (0, 1]$ and an exploration policy $\nu_{(i_k,u_k),k}$
  - Generate a successor state j_k according to distribution $p_{i_k j}(u_k)$, $j = 1, \ldots, n$
  - Generate a control v_k according to distribution $\nu_{(i_k,u_k),k}(v \mid j_k)$, $v \in U(j_k)$
  - Update the (i_k, u_k)th component of Q according to

$$Q_{k+1}(i_k, u_k) = \big(1 - \gamma_{(i_k,u_k),k}\big) Q_k(i_k, u_k) + \gamma_{(i_k,u_k),k}\Big(g(i_k, u_k, j_k) + \alpha \min\big\{J_k(j_k), Q_k(j_k, v_k)\big\}\Big)$$

    and leave all other components of Q_k unchanged
- Exploration policy $\nu_{(i_k,u_k),k}$ may be (arbitrarily) related to current policy μ_k
- There are versions that use cost function approximation and the stopping algorithm of Tsitsiklis and VanRoy (1999)

Generalized DP – 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

- Many DP models beyond standard discounted can be modeled this way
  - Semi-Markov and minimax discounted problems
  - Stochastic shortest path problems
  - Q-learning versions of the above
  - Multi-agent aggregation

- Assume that for all i and $u \in U(i)$

$$\big|H(i, u, J) - H(i, u, J')\big| \leq \alpha \|J - J'\|_\infty$$

- Then T and T_μ are sup-norm contractions with fixed points J^* and J_μ
- The preceding uniform contraction analysis and algorithms generalize

Concluding Remarks

- A new approach to optimistic and exploration-enhanced policy iteration
  - Replaces policy evaluation step with a stopping problem
  - Is based on a uniform sup-norm contraction ... common fixed point for all μ
  - Yields: 1) Improved convergence properties, and 2) Exploration benefit
- Several interlocking research directions
- Optimistic Q-learning (lookup table, simulation, stochastic analysis)
- Optimistic policy iteration/Q-learning with cost function approximation and enhanced exploration
- Convergence in distributed asynchronous mode (using convergence theory of distributed asynchronous algorithms)
- Generalized DP (and some nonDP) models: Fixed points of parametric minimization maps
- A nonDP context: Distributed asynchronous computation of fixed point of a concave sup-norm contraction
- Application to monotone (DP or nonDP) mappings (instead of sup-norm contractions)

THANK YOU!