Linearly-solvable Markov decision problems

Emanuel Todorov

Presented by Jordan Frank

February 7, 2007

Contents

- Motivation
- Standard Formalism
- A class of more tractable MDPs
- Iterative solution and convergence analysis
- Alternative problem formulations
- Shortest paths as an eigenvalue problem
- Z-learning

Motivation

- How can we transform MDP problems into easier problems that can be solved efficiently via linear methods or convex optimization?
- It would seem to be a difficult problem, as the discrete and unstructured nature of traditional MDPs seems incompatible with simplifying features such as linearity and convexity.
- But alas, it turns out that there is a family of MDPs where minimization over the control space is convex and analytically tractable, where the Bellman equation can be exactly transformed into a linear equation.
- Not only that, but this new family of MDPs can yield accurate approximations to traditional MDPs.

Standard Formalism

- $\mathcal{S}$ is a finite set of states, $\mathcal{U}(i)$ is a set of admissible controls at state $i \in \mathcal{S}$, $l(i,u) \geq 0$ is a cost for being in state i and choosing control $u \in \mathcal{U}(i)$, and $P(u)$ is a stochastic matrix whose element $p_{ij}(u)$ is the transition probability from state i to j under control u.
- We focus on problems where a non-empty subset $\mathcal{A} \subseteq \mathcal{S}$ of states are absorbing and incur zero cost. If $\mathcal{A}$ can be reached with non-zero probability in a finite number of steps from any state, then the undiscounted infinite-horizon optimal value function is finite and is the unique solution to the Bellman equation

$$v(i) = \min_{u \in \mathcal{U}(i)} \{ l(i,u) + \sum\nolimits_j p_{ij}(u) v(j) \}$$

A class of more tractable MDPs

- $\mathbf{u} \in \mathbb{R}^{|\mathcal{S}|}$ is a control vector with dimensionality equal to the number of discrete states. Given a transition probability matrix $\bar{P}$ with elements $\bar{p}_{ij}$, controlled transition probabilities defined as $p_{ij}(\mathbf{u}) = \bar{p}_{ij} \exp(u_j)$.
- Define control cost in terms of difference between uncontrolled and controlled transition probabilities, measured using KL divergence. Control cost simplifies to

$$r(i, \mathbf{u}) = \sum_j p_{ij}(\mathbf{u}) u_j.$$

- Add an arbitrary state cost $q(i) \geq 0$ in addition to above control cost, and then define the cost function for our MDP as

$$l(i, \mathbf{u}) = q(i) + r(i, \mathbf{u}).$$

A class of more tractable MDPs

- The Bellman equation for our MDP is

$$v(i) = \min_{\mathbf{u} \in \mathcal{U}(i)} \{q(i) + \sum\nolimits_j \bar{p}_{ij} \exp(u_j)(u_j + v(j))\}$$

- Admissible controls are

$$\mathcal{U}(i) = \{\mathbf{u} \in \mathbb{R}^{|\mathcal{S}|}; \sum\nolimits_j \bar{p}_{ij} \exp(u_j) = 1; \bar{p}_{ij} = 0 \implies u_j = 0\}$$

- And so we have a constrained optimization problem which we can perform in closed form using Lagrange multipliers.

A class of more tractable MDPs

- Optimal control law is

$$u_j^* = -v(j) - \log\left(\sum_k \bar{p}_{ik} \exp(-v(k))\right).$$

- Optimally-controlled transition probabilities are

$$p_{ij}(\mathbf{u}^*(i)) = \frac{\bar{p}_{ij} \exp(-v(j))}{\sum_k \bar{p}_{ik} \exp(-v(k))}$$

- Introducing the exponential transformation $z(i) = \exp(-v(i))$ makes the minimized Bellman equation linear:

$$z(i) = \exp(-q(i)) \sum_j \bar{p}_{ij} z(j)$$

A class of more tractable MDPs

- Defining a vector $\mathbf{z}$ with elements $z(i)$, and the diagonal matrix G with elements $\exp(-q(i))$ along its main diagonal, we can formulate the minimized Bellman equation as

$$\mathbf{z} = G\bar{P}\mathbf{z}$$

  And so we have reduced our class of optimal control problems to a linear eigenvalue problem. $\mathbf{z}$ is an eigenvector of $G\bar{P}$ with eigenvalue 1.

Iterative solution and convergence analysis

- Our $\mathbf{z}$ is going to exist, and is going to be unique because the Bellman equation has a unique solution and v is a solution to the Bellman equation iff $z = \exp(-v)$ is an admissible solution to $\mathbf{z} = G\bar{P}\mathbf{z}$.
- The obvious iterative method is $\mathbf{z}_{k+1} = G\bar{P}\mathbf{z}_k$, with $z_0 = \mathbf{1}$. This is guaranteed to converge to the unique solution.

Iterative solution and convergence analysis

- To analyze convergence rate, we permute the states so that $G\bar{P}$ is in canonical form:

$$G\bar{P} = \begin{bmatrix} T_1 & T_2 \\ 0 & I \end{bmatrix}$$

  where the absorbing states are last (hence the identity matrix in the lower-right corner).

- All eigenvalues of T_1 are smaller than 1, and so $\lim_{k\to\infty} T_1^k = 0$.

- Therefore our iterative method converges exponentially as γ^k where $\gamma < 1$ is the largest eigenvalue of T_1.

Iterative solution and convergence analysis

- Larger state costs $q(i)$ and small transition probabilities among non-absorbing states can lead to smaller values of γ.
- γ does not have any reason to increase as the dimensionality of T_1 increases though, and so convergence is independent of problem size!
- Author claims that numerical simulations on randomly generated MDPs have shown that problem size does not systematically affect the number of iterations needed to reach a given convergence criterion.
- Therefore average running time scales linearly with the number of non-zero elements in $\bar{P}$.

Alternative problem formulations

- Alternative formulations given for finite-horizon problems, infinite-horizon average-cost-per-stage problems, and infinite-horizon discounted-cost problems.
- Even in the infinite-horizon discounted-cost problem, where the formulation for the minimized Bellman equation is nonlinear, it has been observed that the iterative method discussed earlier still converges rapidly.

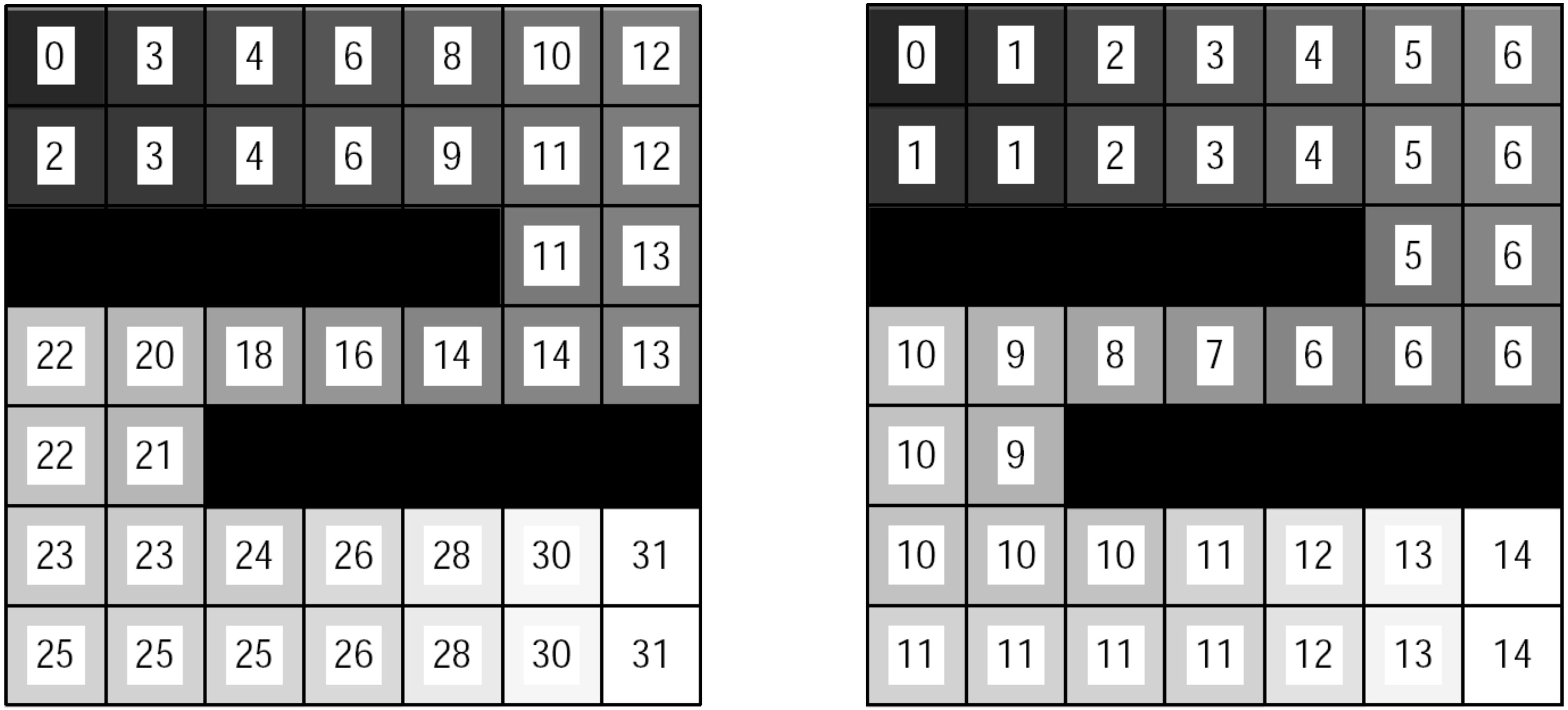
Shortest paths as an eigenvalue problem

- Define uncontrolled transition probability matrix $\bar{P}$ corresponding to a random walk on the graph.
- Choose $\rho > 0$ and define state costs $q_\rho(i) = \rho$ when $i \notin \mathcal{A}$ and $q_\rho(i) = 0$ when $i \in \mathcal{A}$.
- Let $v_\rho(i)$ denote the optimal value function defined by $\bar{P}$ and $q_\rho(i)$, and then the shortest path lengths $s(i)$ become

$$s(i) = \lim_{\rho \to \infty} \frac{v_\rho(i)}{\rho}$$

- Our control costs are bounded and so we can choose ρ arbitrarily large to give us a good approximation with one caveat, when ρ gets too large, $\exp(-\rho)$ becomes numerically indistinguishable from 0.

Shortest paths as an eigenvalue problem



(a) Solution with $\rho = 1$ (b) Solution with $\rho = 50$

Z-learning

- Assume model is unknown, and all we have access to are samples (i_k, j_k, q_k) where i_k is the current state, j_k is the next state, q_k is the state cost incurred at i_k, and k is the sample number. Then we can write the minimized Bellman equation as

$$z(i) = \exp(-q(i)) \sum_j \bar{p}_{ij} z(j) = \exp(-q(i)) E_{\bar{P}}[z(j)]$$

which suggests an obvious stochastic approximation $\hat{z}$ to the function z

$$\hat{z}(i_k) \leftarrow (1 - \alpha_k)\hat{z}(i_k) + \alpha_k \exp(-q_k)\hat{z}(j_k)$$

Z-learning

- Z-learning update rule is

$$\hat{z}(i_k) \leftarrow (1-\alpha_k)\hat{z}(i_k) + \alpha_k \exp(-q_k)\hat{z}(j_k)$$

- Compare to Q-learning approximation where l_k is now a total cost rather than a state cost, and we have a control u_k generated by some control policy

$$\hat{Q}(i_k, u_k) \leftarrow (1-\alpha_k)\hat{Q}(i_k, u_k) + \alpha_k \min_{u' \in \mathcal{U}(j_k)} \Big(l_k + \hat{Q}(j_k, u')\Big)$$

- Note that Z-learning does not require state-action values, or a maximization operator (Maybe Q-learning isn't the best algorithm to test against?).

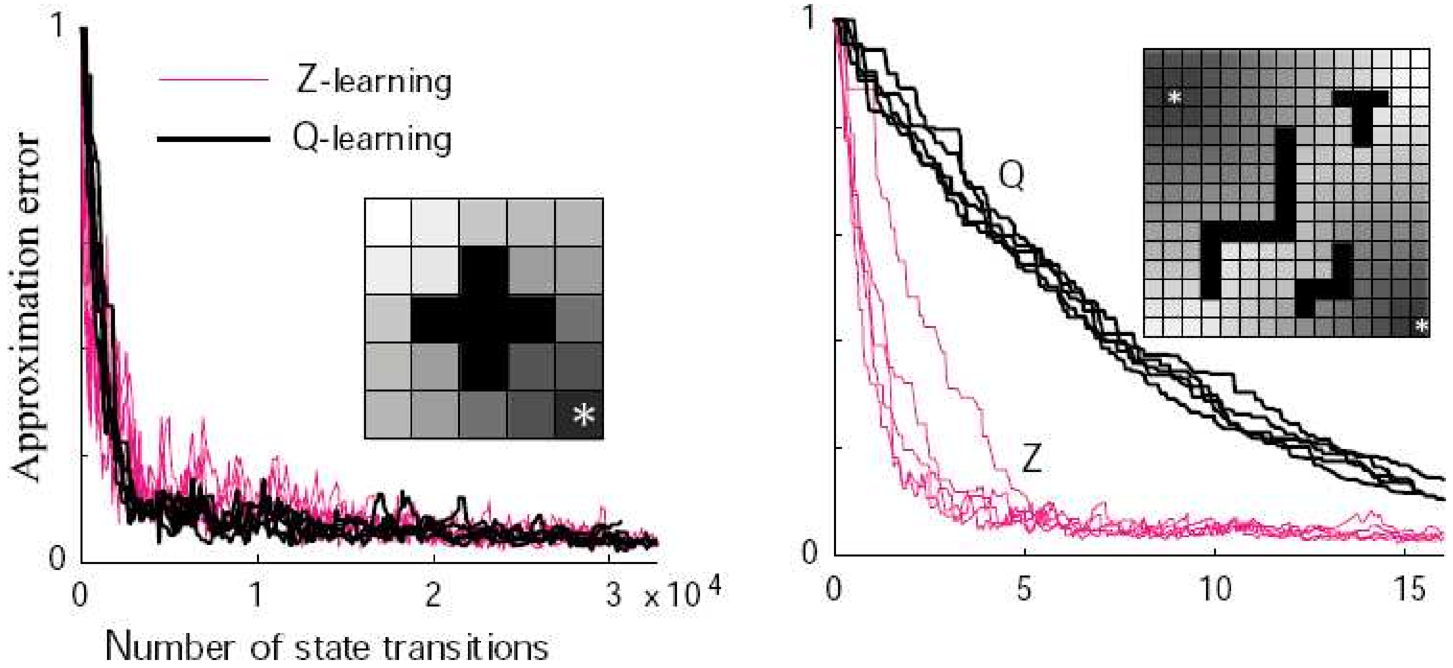
Z-learning

- To compare Q-learning and Z-learning, first construct a continuous MDP with $q(i) = 1$ and transitions to immediate neighbours.
- Find optimal transition probabilities.
- Construct discrete MDP with identical optimal value function.
- Measure approximation error as

$$\frac{\max_i |v(i) - \hat{v}(i)|}{\max_i v(i)}$$

- Compare approximation error with Z-learning and Q-learning.

Z-Learning



- Note that the performance of Q-learning can be improved by using a non-random (say ϵ-greedy) policy, and Z-learning can be improved using importance sampling.

Summary

- New class of MDPs that can be solved efficiently.
- Rate of convergence does not depend on problem size.
- We can approximate traditional MDPs to a high degree with these linearly-solvable MDPs.
- Very good approximations to shortest path problem solution in $O(n)$ time.
- Z-learning algorithm seems to outperform Q-learning on simple gridworld tasks.
- It was noted that combining Z-learning with importance sampling would improve the performance even further.