6.231 Dynamic Programming and Stochastic Control
Fall 2008

6.231 DYNAMIC PROGRAMMING

LECTURE 25

LECTURE OUTLINE

- Additional topics in ADP
- Nonlinear versions of the projected equation
- Extension of Q-learning for optimal stopping
- Basis function adaptation
- Gradient-based approximation in policy space

NONLINEAR EXTENSIONS OF PROJECTED EQ.

- If the mapping T is nonlinear (as for example in the case of multiple policies) the projected equation $\Phi r = \Pi T(\Phi r)$ is also nonlinear.

- Any solution r^* satisfies

$$r^* \in \arg \min_{r \in \Re^s} \big\| \Phi r - T(\Phi r^*) \big\|^2$$

or equivalently

$$\Phi' \Big(\Phi r^* - T(\Phi r^*) \Big) = 0$$

This is a nonlinear equation, which may have one or many solutions, or no solution at all.

- If ΠT is a contraction, then there is a unique solution that can be obtained (in principle) by the fixed point iteration

$$\Phi r_{k+1} = \Pi T(\Phi r_k)$$

- We have seen a nonlinear special case of projected value iteration/LSPE where ΠT is a contraction, namely optimal stopping.

- This case can be generalized.

LSPE FOR OPTIMAL STOPPING EXTENDED

- Consider a system of the form

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

where $f : \Re^n \mapsto \Re^n$ is a mapping with scalar components of the form $f(x) = \big(f_1(x_1), \ldots, f_n(x_n)\big)$.

- Assume that each $f_i : \Re \mapsto \Re$ is nonexpansive:

$$\big|f_i(x_i) - f_i(\bar{x}_i)\big| \le |x_i - \bar{x}_i|, \qquad \forall\ i,\ x_i, \bar{x}_i \in \Re$$

This guarantees that T is a contraction with respect to any weighted Euclidean norm $\|\cdot\|_\xi$ whenever A is a contraction with respect to that norm.

- Algorithms similar to LSPE [approximating $\Phi r_{k+1} = \Pi T(\Phi r_k)$] are then possible.

- Special case: In the optimal stopping problem of Section 6.4, x is the Q-factor corresponding to the continuation action, $\alpha \in (0,1)$ is a discount factor, $f_i(x_i) = \min\{c_i, x_i\}$, and $A = \alpha P$, where P is the transition matrix for continuing.

- If $\sum_{j=1}^n p_{\bar{i}j} < 1$ for some state $\bar{i}$, and $0 \le P \le Q$, where Q is an irreducible transition matrix, then $\Pi\big((1-\gamma)I + \gamma T\big)$ is a contraction with respect to $\|\cdot\|_\xi$ for all $\gamma \in (0,1)$, even with $\alpha = 1$.

BASIS FUNCTION ADAPTATION I

- An important issue in ADP is how to select basis functions.

- A possible approach is to introduce basis functions that are parametrized by a vector θ, and optimize over θ, i.e., solve the problem

$$\min_{\theta \in \Theta} F\big(\tilde{J}(\theta)\big)$$

where $\tilde{J}(\theta)$ is the solution of the projected equation.

- One example is

$$F\big(\tilde{J}(\theta)\big) = \big\|\tilde{J}(\theta) - T\big(\tilde{J}(\theta)\big)\big\|^2$$

- Another example is

$$F\big(\tilde{J}(\theta)\big) = \sum_{i \in I} |J(i) - \tilde{J}(\theta)(i)|^2,$$

where I is a subset of states, and $J(i)$, $i \in I$, are the costs of the policy at these states calculated directly by simulation.

BASIS FUNCTION ADAPTATION II

- Some algorithm may be used to minimize $F\big(\tilde{J}(\theta)\big)$ over θ.

- A challenge here is that the algorithm should use low-dimensional calculations.

- One possibility is to use a form of random search method; see the paper by Menache, Mannor, and Shimkin (Annals of Oper. Res., Vol. 134, 2005)

- Another possibility is to use a gradient method. For this it is necessary to estimate the partial derivatives of $\tilde{J}(\theta)$ with respect to the components of θ.

- It turns out that by differentiating the projected equation, these partial derivatives can be calculated using low-dimensional operations. See the paper by Menache, Mannor, and Shimkin, and a recent paper by Yu and Bertsekas (2008).

APPROXIMATION IN POLICY SPACE I

- Consider an average cost problem, where the problem data are parametrized by a vector r, i.e., a cost vector $g(r)$, transition probability matrix $P(r)$. Let $\eta(r)$ be the (scalar) average cost per stage, satisfying Bellman's equation

$$\eta(r)e + h(r) = g(r) + P(r)h(r)$$

where $h(r)$ is the corresponding differential cost vector.

- Consider minimizing $\eta(r)$ over r (here the data dependence on control is encoded in the parametrization). We can try to solve the problem by nonlinear programming/gradient descent methods.

- **Important fact:** If $\Delta\eta$ is the change in η due to a small change Δr from a given r, we have

$$\Delta\eta = \xi'(\Delta g + \Delta P h),$$

where ξ is the steady-state probability distribution/vector corresponding to $P(r)$, and all the quantities above are evaluated at r:

$$\Delta\eta = \eta(r + \Delta r) - \eta(r),$$

$$\Delta g = g(r+\Delta r) - g(r), \qquad \Delta P = P(r+\Delta r) - P(r)$$

APPROXIMATION IN POLICY SPACE II

- **Proof of the gradient formula:** We have, by "differentiating" Bellman's equation,

$$\Delta\eta(r)\cdot e+\Delta h(r) = \Delta g(r)+\Delta P(r)h(r)+P(r)\Delta h(r)$$

By left-multiplying with ξ',

$$\xi'\Delta\eta(r)\cdot e+\xi'\Delta h(r) = \xi'\Big(\Delta g(r)+\Delta P(r)h(r)\Big)+\xi' P(r)\Delta h(r)$$

Since $\xi'\Delta\eta(r)\cdot e = \Delta\eta(r)$ and $\xi' = \xi' P(r)$, this equation simplifies to

$$\Delta\eta = \xi'(\Delta g + \Delta P h)$$

- Since we don't know ξ, we cannot implement a gradient-like method for minimizing $\eta(r)$. An alternative is to use "sampled gradients", i.e., generate a simulation trajectory $(i_0, i_1, \ldots)$, and change r once in a while, in the direction of a simulation-based estimate of $\xi'(\Delta g + \Delta P h)$.

- There is much recent research on this subject, see e.g., the work of Marbach and Tsitsiklis, and Konda and Tsitsiklis, and the refs given there.