CS 287: Advanced Robotics
Fall 2009

Lecture 5: Control 4: Optimal control / Reinforcement learning--- function approximation in dynamic programming

Pieter Abbeel
UC Berkeley EECS

Today

- Optimal control/Reinforcement learning provide a general approach to tackle temporal decision making problems.
- Often the state space is too large to perform exact Dynamic Programming (DP) / Value Iteration (VI)

→ Today: Dynamic programming with function approximation

Great references:

Gordon, 1995, "Stable function approximation in dynamic programming"

Tsitsiklis and Van Roy, 1996, "Feature based methods for large scale dynamic programming"

Bertsekas and Tsitsiklis, "Neuro-dynamic programming," Chap. 6

Recall: Discounted infinite horizon

- Markov decision process (MDP) (S, A, P, γ, g)
  - γ: discount factor
- Policy $\pi = (\mu_0, \mu_1, \ldots)$, $\mu_k : S \to A$
- Value of a policy π: $J^\pi(x) = \mathrm{E}[\sum_{t=0}^{\infty} \gamma^t g(x(t), u(t)) | x_0 = x, \pi]$
- Goal: find $\pi^* \in \arg\min_{\pi \in \Pi} V^\pi$

Recall: Discounted infinite horizon

- **Dynamic programming (DP) aka Value iteration (VI):**

  For i=0,1, …

  For all s ∈ S

$$J^{(i+1)}(s) \leftarrow \min_{u \in A} g(s,u) + \gamma \sum_{s'} P(s'|s,u) J^{(i)}(s')$$

- **Facts:**

  $J^{(i)} \to J^*$ for $i \to \infty$

  There is an optimal stationary policy: $\pi^* = (\mu^*, \mu^*, \ldots)$ which satisfies:

$$\mu^*(s) = \arg\min_u g(s,u) + \gamma \sum_{s'} P(s'|s,u) J^*(s)$$

- **Issue in practice: Bellman's curse of dimensionality: number of states grows exponentially in the dimensionality of the state space**

DP/VI with function approximation

Pick some $S' \subseteq S$ [typically the idea is that $|S'| << |S|$].
Iterate for $i = 0, 1, 2, \ldots$:

back-ups: $\forall s \in S' : \bar{J}^{(i+1)}(s) \leftarrow \min_{u \in A} g(s,u) + \gamma \sum_{s'} P(s'|s,u) \hat{J}_{\theta^{(i)}}(s')$

projection: find some $\theta^{(i+1)}$ such that $\forall s \in S' \quad \hat{J}_{\theta^{(i+1)}}(s) \approx \bar{J}^{(i+1)}(s)$

Projection enables generalization to $s \in S \setminus S'$, which in turn enables the Bellman back-ups in the next iteration.

θ parameterizes the class of functions used for approximation of the cost-to-go function

Function approximation examples

- Piecewise linear over triangles (tetrahedrons, etc.)
- Piecewise constant over sets of states (=nearest neighbor, often called state aggregation)
- Fit a neural net to $\bar{J}$

  $\theta^{(i+1)} = \arg\min_\theta \sum_{s \in S'} \text{loss}(\bar{J}^{(i+1)}(s) - f_\theta(s))$
- Least squares fit to $\bar{J}$

  $\theta^{(i+1)} = \arg\min_\theta \sum_{s \in S'} (\bar{J}^{(i+1)}(s) - \phi(s)^\top \theta)^2$
- Bezier patches (=particular choice of convex weighting)
- [[TODO: work out examples in more detail.]]

Potential guarantees?

- If we have bounded error during function approximation in each iteration, i.e.,

$$\forall i : \| \bar{J}^{(i)} - \hat{J}_{\theta^{(i)}} \| \leq \epsilon$$

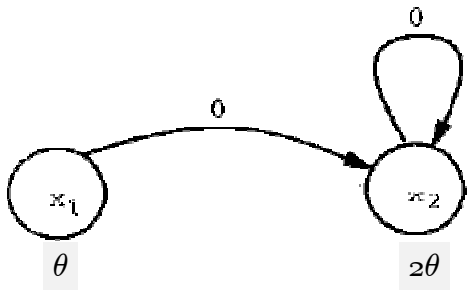
  can we provide any guarantees?

- If we have a function approximation architecture that can capture the true cost-to-go function within some approximation error, i.e.,

$$\exists \theta^* : f_{\text{loss}}(J^*, \hat{J}_{\theta^*}) \leq \epsilon$$

  can we provide any guarantees?

Simple example



Function approximator: [1 2] * θ

Simple example

$$\bar{J}_\theta = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \theta$$

$$\begin{aligned} \bar{J}^{(1)}(x_1) &= 0 + \gamma \hat{J}_{\theta^{(0)}}(x_2) = 2\gamma\theta^{(0)} \\ \bar{J}^{(1)}(x_2) &= 0 + \gamma \hat{J}_{\theta^{(0)}}(x_2) = 2\gamma\theta^{(0)} \end{aligned}$$

Function approximation with least squares fit:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \theta^{(1)} \approx \begin{bmatrix} 2\gamma\theta^{(0)} \\ 2\gamma\theta^{(0)} \end{bmatrix}$$

Least squares fit results in:

$$\theta^{(1)} = \frac{6}{5}\gamma\theta^{(0)}$$

Repeated back-ups and function approximations result in:

$$\theta^{(i)} = \left(\frac{6}{5}\gamma\right)^i \theta^{(0)}$$

which diverges if $\gamma > \frac{5}{6}$ even though the function approximation class can represent the true value function.]

Bellman operator

- **Dynamic programming (DP) aka Value iteration (VI):**

  Set $J^{(0)} = 0$

  For i=0,1, …

  For all s ∈ S

  $$J^{(i+1)}(s) \leftarrow \min_{u \in A} \sum_{s'} P(s'|s,u) \left(g(s,a) + \gamma J^{(i)}(s') \right)$$

- **Bellman operator T**

  $T : \Re^{|S|} \rightarrow \Re^{|S|}$ is defined as:

  $$(TJ)(s) = \min_{u \in A} g(s,a) + \gamma \sum_{s'} P(s'|s,u) J(s')$$

- **Hence running value iteration can be compactly written as:**

  Set $J = 0$

  Repeat $J \leftarrow TJ$.

- **Note: T is a *non-linear* operator (b/c of the min).**

Contractions

Definition. The operator F is a α-contraction w.r.t. some norm $\|\cdot\|$ if

$$\forall X, \overline{X}: \quad \|FX - F\overline{X}\| \leq \alpha \|X - \overline{X}\|$$

Theorem 1. The sequence $X, FX, F^2X, ...$ converges for every X.

Theorem 2. F has a unique fixed point X^* which satisfies $FX^* = X^*$ and all sequences $X, FX, F^2X, ...$ converge to this unique fixed point X^*.

Proof of Theorem 1

Useful fact.
Cauchy sequences: If for $x_0, x_1, x_2, \ldots$, we have that

$$\forall \epsilon, \exists K: \|x_M - x_N\| < \epsilon \quad for \quad M, N > K$$

then we call $x_0, x_1, x_2, ...$ a Cauchy sequence.

If $x_0, x_1, x_2, \ldots$ is a Cauchy sequence, and $x_i \in \Re^n$, then there exists $x^* \in \Re^n$ such that $\lim_{i \to \infty} x_i = x^*$.

Proof.
Assume $N > M$.

$$\begin{aligned} \|F^M X - F^N X\| = & \quad \|\textstyle\sum_{i=M}^{N-1} (F^i X - F^{i+1} X)\| \\ \leq & \quad \textstyle\sum_{i=M}^{N-1} \|F^i X - F^{i+1} X\| \\ \leq & \quad \textstyle\sum_{i=M}^{N-1} \alpha^i \|X - FX\| \\ = & \quad \|X - FX\| \textstyle\sum_{i=M}^{N-1} \alpha^i \\ = & \quad \|X - FX\| \frac{\alpha^M}{1-\alpha}. \end{aligned}$$

As $\|X - FX\| \frac{\alpha^M}{1-\alpha}$ goes to zero for M going to infnity, we have that for any $\epsilon > 0$ for $\|F^M X - F^N X\| \leq \epsilon$ to hold for all $M, N > K$, it suffices to pick K large enough.

Hence $X, FX, \ldots$ is a Cauchy sequence and converges.

Proof of uniqueness of fixed point

Suppose F has two fixed points. Let's say

$$\begin{aligned} FX_1 &= X_1, \\ FX_2 &= X_2, \end{aligned}$$

this implies,

$$\|FX_1 - FX_2\| = \|X_1 - X_2\|.$$

At the same time we have from the contractive property of F

$$\|FX_1 - FX_2\| \leq \alpha \|X_1 - X_2\|.$$

Combining both gives us

$$\|X_1 - X_2\| \leq \alpha \|X_1 - X_2\|.$$

Hence,

$$X_1 = X_2.$$

Therefore, the fixed point of F is unique.

The Bellman operator is a contraction

Definition. The infinity norm of a vector $x \in \Re^n$ is defined as

$$\|x\|_\infty = \max_i |x_i|$$

Fact. The Bellman operator T is a γ-contraction with respect to the infinity norm, i.e.,

$$\|TJ_1 - TJ_2\|_\infty \leq \gamma \|J_1 - J_2\|_\infty$$

Corollary. From any starting point, value iteration/Dynamic programming converges to a unique fixed point J^* which satisfies $J^* = TJ^*$.

Proof Bellman operator is a contraction

Lemma 1 *Sup-Norm Contraction*

$$||TJ - T\bar{J}||_\infty \leq \alpha ||J - \bar{J}||_\infty .$$

Note that for $y \in \mathbb{R}^{|S|}$, the sup-norm is defined as $||y||_\infty = \max_{s \in S} |y_s|$.

Proof Consider the following statement:

$$|\max_z f(z) - \max_z h(z)| \leq \max_z |f(z) - h(z)|.$$

To show that this is true:

$$\begin{aligned} \max_z f(z) - \max_z h(z) &= f(z^*) - \max_z h(z) \\ &\leq f(z^*) - h(z^*) \\ &\leq \max_z |f(z) - h(z)|. \end{aligned}$$

Applying this to $TJ(x)$ we have:

$$\begin{aligned} &|TJ(x) - T\bar{J}(x)| \\ &= |\max_{u \in U(x)} g(x,u) + \alpha \sum_{y \in S} P_{xy}(u) J(y) - \max_{u \in U(x)} g(x,u) + \alpha \sum_{y \in S} P_{xy}(u) \bar{J}(y)| \\ &\leq \max_{u \in U(x)} |g(x,u) + \alpha \sum_{y \in S} P_{xy}(u) J(y) - g(x,u) - \alpha \sum_{y \in S} P_{xy}(u) \bar{J}(y)| \\ &\leq \max_{u \in U(x)} \alpha |\sum_{y \in S} P_{xy}(u)(J(y) - \bar{J}(y))| \\ &\leq \alpha \max_{y \in S} |J(y) - \bar{J}(y)| \\ &= \alpha ||J - \bar{J}||_\infty . \end{aligned}$$

Value iteration variations

- Gauss-Seidel value iteration

For i=1, 2, …

for s=1,…,|S|

$$J(s) \leftarrow \min_{u \in A} g(s,u) + \gamma \sum_{s'} P(s'|s,u) J(s')$$

Compare to regular value iteration:

$$J^{(i+1)}(s) \leftarrow \min_{u \in A} g(s,u) + \gamma \sum_{s'} P(s'|s,u) J^{(i)}(s')$$

Exercise: Show that Gauss-Seidel value iteration converges to J*. [Hint: proceed by showing the combined operator which does the sequential update for all states s=1,…,|S| is a infinity norm contraction.]

Value iteration variations

- Asynchronous value iteration

Pick an infinite sequence of states,

$$s^{(0)}, s^{(1)}, s^{(2)}, \ldots$$

such that every state $s \in S$ occurs infinitely often. Define the operators $T_{s^{(k)}}$ as follows:

$$(T_{s^{(k)}} J)(s) = \begin{cases} (TJ)(s), & \text{if } s^{(k)} = s \\ J(s), & \text{otherwise} \end{cases}$$

Asynchronous value iteration initializes J and then applies, in sequence, $T_{s^{(0)}}, T_{s^{(1)}}, \ldots$.

Exercise: Show that asynchronous value iteration converges to J*.

DP/VI with function approximation

Pick some $S' \subseteq S$ [typically the idea is that $|S'| << |S|$].
Iterate for $i = 0, 1, 2, \ldots$:

back-ups: $\forall s \in S' : \bar{J}^{(i+1)}(s) \leftarrow \min_{u \in A} \sum_{s'} P(s'|s,u) \left(g(s,u) + \gamma \hat{J}_{\theta^{(i)}}(s') \right)$

projection: find some $\theta^{(i+1)}$ such that $\forall s \in S' \quad \hat{J}_{\theta^{(i+1)}}(s) \approx \bar{J}^{(i+1)}(s)$

- New notation: projection operator Π maps from $\Re^n$ into the subset of $\Re^n$ which can be represented by the function approximator class

$$\bar{J}^{(i+1)} \leftarrow \Pi T \bar{J}^{(i)}$$

- While theoretical convergence analysis does not depend on this, the projection operator Π has to operate based upon only knowing J at the points s ∈ S' , otherwise not practically feasible for large scale problems

Composing operators

- **Definition.** An operator G is a *non-expansion* with respect to a norm || . || if

$$\|GJ_1 - GJ_2\| \leq \|J_1 - J_2\|$$

- **Fact.** If the operator F is a γ contraction with respect to a norm || . || and the operator G is a non-expansion with respect to the same norm, then the sequential application of the operators G and F is a γ-contraction, i.e., $\|GFJ_1 - GFJ_2\| \leq \gamma\|J_1 - J_2\|$
- **Corollary.** The operator ΠT is a γ-contraction w.r.t. the infinity norm if Π is a non-expansion w.r.t. the infinity norm.

Averager function approximators are non-expansions

DEFINITION: A real valued function approximation scheme is an *averager* if every fitted value is the weighted average of zero or more target values and possibly some predetermined constants. The weights involved in calculating the fitted value $\hat{Y}_i$ may depend on the sample vector X_0, but may not depend on the target values Y. More precisely, for a fixed X_0, if Y has n elements, there must exist n real numbers k_i, n^2 nonnegative real numbers β_{ij}, and n nonnegative real numbers β_i, so that for each i we have $\beta_i + \sum_j \beta_{ij} = 1$ and $\hat{Y}_i = \beta_i k_i + \sum_j \beta_{ij} Y_j$.

- Examples:
  - nearest neighbor (aka state aggregation)
  - linear interpolation over triangles (tetrahedrons, …)

Averager function approximators are non-expansions

Theorem. The mapping Π associated with any averaging method is a nonexpansion in the infinity norm.

Proof: Let J_1 and J_2 be two vectors in $\Re^n$. Consider a particular entry s of ΠJ_1 and ΠJ_2:

$$\begin{aligned} |(\Pi J_1)(s) - (\Pi J_2)(s)| &= |\beta_{s0} + \sum_{s'} \beta_{ss'} J_1(s') - \beta_{s0} + \sum_{s'} \beta_{ss'} J_2(s')| \\ &= |\sum_{s'} \beta_{ss'} (J_1(s') - J_2(s'))| \\ &\leq \max_{s'} |J_1(s') - J_2(s')| \\ &= \|J_1 - J_2\|_\infty \end{aligned}$$

This holds true for all s, hence we have

$$\|\Pi J_1 - \Pi J_2\|_\infty \leq \|J_1 - J_2\|_\infty$$

Linear regression ☹

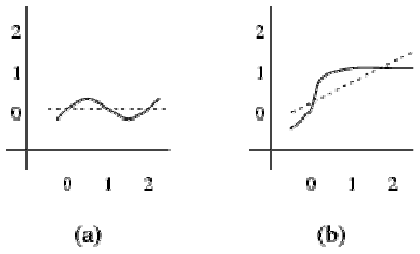


Figure 2: The mapping associated with linear regression when samples are taken at the points $x = 0, 1, 2$. In (a) we see a target value function (solid line) and its corresponding fitted value function (dotted line). In (b) we see another target function and another fitted function. The first target function has values $y = 0, 0, 0$ at the sample points; the second has values $y = 0, 1, 1$. Regression exaggerates the difference between the two functions: the largest difference between the two target functions at a sample point is 1 (at $x = 1$ and $x = 2$), but the largest difference between the two fitted functions at a sample point is $\frac{7}{6}$ (at $x = 2$).

[Example taken from Gordon, 1995.]

Guarantees for fixed point

Theorem. Let J^* be the optimal value function for a finite MDP with discount factor γ. Let the projection operator Π be a non-expansion w.r.t. the infinity norm and let $\tilde{J}$ be any fixed point of Π. Suppose $\|\tilde{J} - J^*\|_\infty \leq \epsilon$. Then ΠT converges to a value function $\bar{J}$ such that:

$$\|\bar{J} - J^*\| \leq 2\epsilon + \frac{2\gamma\epsilon}{1-\gamma}$$

- I.e., if we pick a non-expansion function approximator which can approximate J* well, then we obtain a good value function estimate.
- To apply to discretization: use continuity assumptions to show that J* can be approximated well by chosen discretization scheme