Quantitative Economics Workshop Paris

Dynamic Programming

John Stachurski

September 2022

Introduction

Summary of this lecture:

- Foobar
- Foobar

Introduction to Dynamic Programming

Dynamic program

an initial state X_0 is given
$t \leftarrow 0$
while $t < T$ **do**
    observe current state X_t
    choose action A_t
    receive reward R_t based on (X_t, A_t)
    state updates to X_{t+1}
    $t \leftarrow t + 1$
end

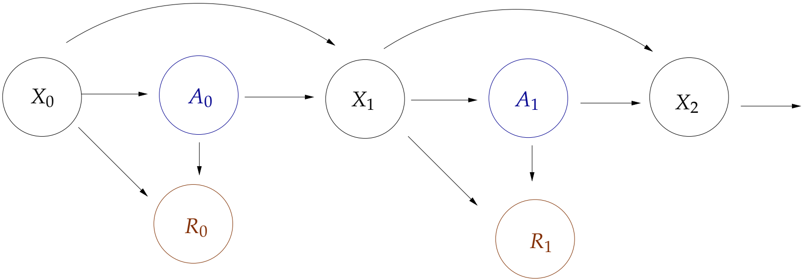

Figure: A dynamic program

Comments:

- Objective: maximize **lifetime rewards**
  - Some aggregation of $R_0, R_1, \ldots$
  - Example. $\mathbb{E}[R_0 + \beta R_1 + \beta^2 R_2 + \cdots]$ for some $\beta \in (0,1)$
- If $T < \infty$ then the problem is called a **finite horizon** problem
- Otherwise it is called an **infinite horizon** problem
- The update rule can also depend on random elements:

$$X_{t+1} = F(X_t, A_t, \xi_{t+1})$$

Example. A retailer sets prices and manages inventories to maximize profits

- X_t measures
  - current business environment
  - the size of the inventories
  - prices set by competitors, etc.
- A_t specifies current prices and orders of new stock
- R_t is current profit π_t
- Lifetime reward is

$$\mathbb{E}\left[\pi_0 + \frac{1}{1+r}\pi_1 + \left(\frac{1}{1+r}\right)^2 \pi_2 + \cdots\right] = \text{EPV}$$

Markov Decision Processes

- A class of dynamic programs
- Broad enough to encompass many economic applications
- Includes optimal stopping problems as a special case
- Clean, powerful theory
- A range of important algorithms

Also a cornerstone for

- reinforcement learning, artificial intelligence, etc.

MDPs are dynamic programs characterized by two features

1. Rewards are additively separable:

$$\text{lifetime reward } = \mathbb{E} \sum_{t \geqslant 0} \beta^t R_t$$

2. The discount rate is constant

For now we restrict attention to finite state and action spaces

- Routinely used in quantitative applications
- Avoids technical issues we can put aside for later

Notation

Let X and A be any sets

A **correspondence** Γ from X to A is a map that associates each $x \in X$ to a subset of A

- called **nonempty** if $\Gamma(x) \neq \emptyset$ for all $x \in X$

Examples.

- $\Gamma(x) = [0, x]$ is a correspondence from $\mathbb{R}$ to $\mathbb{R}$
- $\Gamma(x) = [-x, x]$ is a nonempty correspondence from $\mathbb{R}$ to $\mathbb{R}$

We study a controller who, at each integer $t \geqslant 0$

1. observes the current state X_t
2. responds with an action A_t

Her aim is to maximize expected discounted rewards

$$\mathbb{E} \sum_{t \geqslant 0} \beta^t r(X_t, A_t), \qquad X_0 = x_0 \text{ given}$$

We take as given

1. a finite set X called the **state space** and
2. a finite set A called the **action space**

The actions of the controller are limited by a **feasible correspondence** Γ

- A correspondence from X to A
- $\Gamma(x)$ is the set of actions available to the controller in state x

Given Γ, we set

$$\mathsf{G} := \{(x, a) \in \mathsf{X} \times \mathsf{A} : a \in \Gamma(x)\}$$

- called the set of **feasible state-action pairs**

Reward $r(x, a)$ is received at feasible state-action pair (x, a), A **stochastic kernel** from G to X is a map $P \colon \mathsf{G} \times \mathsf{X} \to \mathbb{R}_+$ satisfying

$$\sum_{x' \in \mathsf{X}} P(x, a, x') = 1 \quad \text{for all } (x, a) \text{ in } \mathsf{G}$$

Interpretation

- For each feasible state-action pair, $P(x, a, \cdot)$ is a distribution
- The next period state x' is selected from $P(x, a, \cdot)$

Now let's put it all together:

Given X and A, a **Markov decision process** (**MDP**) is a tuple (Γ, β, r, P) where

1. Γ is a nonempty correspondence from $\mathsf{X} \to \mathsf{A}$
2. β is a constant in $(0, 1)$
3. r is a function from G to $\mathbb{R}$
4. P is a stochastic kernel from G to X

In the foregoing,

- β is called the **discount factor**
- r is called the **reward function**

Algorithm 1: MDP dynamics: states, actions, and rewards

$t \leftarrow 0$
input X_0
while $t < \infty$ **do**
  observe X_t
  choose action A_t from $\Gamma(X_t)$
  receive reward $r(X_t, A_t)$
  draw X_{t+1} from $P(X_t, A_t, \cdot)$
  $t \leftarrow t + 1$
end

Rules:

- Choose $(A_t)_{t \geqslant 0}$ to maximize $\mathbb{E} \sum_{t \geqslant 0} \beta^t r(X_t, A_t)$
- Actions don't depend on future outcomes

The **Bellman equation** is

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x, a) + \beta \sum_{x' \in \mathsf{X}} v(x') P(x, a, x') \right\}$$

Reduces an infinite horizon problem to a two period problem

In the two period problem, the controller trades off

1. current rewards and
2. expected discounted value from future states

Current actions influence both terms

ADD inventory example

Policies

Actions will be governed by policies

- maps from states to actions
- today's action is a function of today's state!

The set of **feasible policies** is

$$\Sigma := \text{ all } \sigma \in \mathsf{A}^{\mathsf{X}} \text{ s.t. } \sigma(x) \in \Gamma(x) \text{ for all } x \in \mathsf{X}$$

Meaning of selecting σ from Σ:

- respond to state X_t with action $A_t := \sigma(X_t)$ at all t

Dynamics

What happens when we always follow $\sigma \in \Sigma$?

Now

$$X_{t+1} \sim P(X_t, \sigma(X_t), \cdot) \quad \text{at every } t$$

Thus, X_t updates according to the stochastic matrix

$$P_\sigma(x, x') := P(x, \sigma(x), x') \qquad (x, x' \in \mathsf{X})$$

The state process becomes P_σ-Markov

- Fixing a policy "closes the loop" in the state dynamics
- Solving an MDP means choosing a Markov chain!

Rewards

Under the policy σ, rewards at x are $r(x, \sigma(x))$

Let

$$r_\sigma(x) := r(x, \sigma(x)) \qquad (x \in \mathsf{X})$$

Now set

$$\mathbb{E}_x := \mathbb{E}[\,\cdot \mid X_0 = x]$$

Then the expected time t reward is

$$\mathbb{E}_x \, r(X_t, A_t) = \mathbb{E}_x \, r_\sigma(X_t) = (P_\sigma^t \, r_\sigma)(x)$$

Let $(X_t)_{t \geqslant 0}$ be P_σ-Markov with $X_0 = x$

The lifetime value of σ starting from x is

$$v_\sigma(x) := \mathbb{E}_x \sum_{t \geqslant 0} \beta^t r_\sigma(X_t)$$

Since $\beta < 1$, we have $r(\beta P_\sigma) < 1$ and hence

$$v_\sigma = \sum_{t \geqslant 0} \beta^t P_\sigma^t \, r_\sigma = (I - \beta P_\sigma)^{-1} \, r_\sigma$$

The **value function** is defined as

$$v^*(x) = \max_{\sigma \in \Sigma} v_\sigma(x) \qquad (x \in \mathsf{X})$$

Recall that the Bellman equation is

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \sum_{x' \in \mathsf{X}} v(x') P(x,a,x') \right\}$$

The **Bellman operator** for the MDP is the self-map T on $\mathbb{R}^{\mathsf{X}}$ defined by

$$(Tv)(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \sum_{x' \in \mathsf{X}} v(x') P(x,a,x') \right\}$$

Obviously

- $Tv = v$ iff v satisfies the Bellman equation
- T is order-preserving on $\mathbb{R}^{\mathsf{X}}$

Fix $v \in \mathbb{R}^{\mathsf{X}}$

A policy $\sigma \in \Sigma$ is called **v-greedy** if

$$\forall\, x \in \mathsf{X}, \;\; \sigma(x) \in \operatorname*{argmax}_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \sum_{x' \in \mathsf{X}} v(x') P(x,a,x') \right\}$$

A policy $\sigma \in \Sigma$ is called **optimal** if

$$v_\sigma = v^*$$

Thus,

$$\sigma \text{ is optimal } \iff \text{ lifetime value is maximal at each state}$$

Proposition. For the MDP described above

1. v^* is the unique fixed point of T in $\mathbb{R}^{\mathsf{X}}$
2. T is a contraction of modulus β on $\mathbb{R}^{\mathsf{X}}$ under the norm $\| \cdot \|_\infty$
3. A feasible policy is optimal if and only it is v^*-greedy
4. At least one optimal policy exists

Proof:

- similar to that for optimal stopping
- full details deferred until we study RDPs