Dynamic Programming & Reinforcement Learning

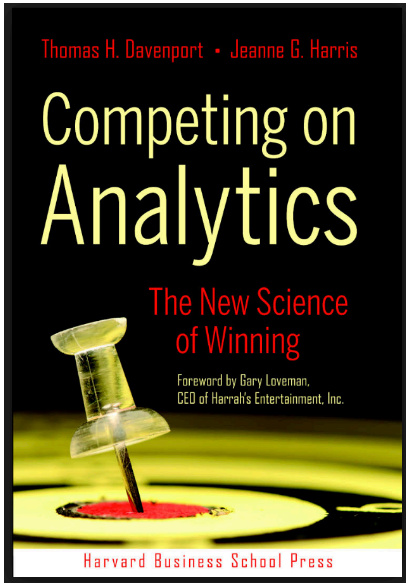
Lecture 3
Finite-horizon Revenue Management & Markov chains

Ger Koole

Revenue management

- Example of finite-horizon dynamic programming
- Illustration how (dynamic) optimization is used in practice

- Success story
  - crucial to airline & hotel industry
  - generated billions of Euros of revenue
  - used as main example in "Competing on Analytics"



Revenue management

- 1978: deregulation aviation US
- Low-cost carriers came into existence
- AA saw reduced market share
- 1985: American Airlines started *RM*
  - regular and *supersaver tariff*
  - limited number of seats and available well in advance
- Problem: how many seats to offer in different booking classes and when to close them?
- Solve a dynamic optimization model

Simple example RM

- 1 seat or room, possible prices 100 and 150
- 2 customers who arrive one-by-one, the first willing to pay 100, the second 150
- What are your consecutive prices? Poll!
- 1 seat or room, 2 “epochs”, each time with probability 0.4 a customer willing to pay 150 and one willing to pay 100
- What are your prices? Poll!
- What is the value = maximal expected reward?

Poll!

Example extended: capacity = 2, 3 epochs

Value function

state	time:	[,1]	[,2]	[,3]	[,4]
0	[1,]	0.0	0	0	0
1	[2,]	124.8	108	80	0
2	[3,]	199.2	160	80	0

Optimal policy

state	time:	[,1]	[,2]	[,3]
1	[1,]	150	150	100
2	[2,]	150	100	100

3 simulations

time	[1] 1 2 3	[1] 1 2 3	[1] 1 2 3
states	[1] 2 2 1	[1] 2 1 0	[1] 2 2 2
actions	[1] 150 100 100	[1] 150 150 150	[1] 150 100 100
demand	[1] 100 100 150	[1] 150 150 150	[1] 100 0 0
rewards	[1] 0 100 100	[1] 150 150 0	[1] 0 0 0
total reward	[1] 200	[1] 300	[1] 0

Suppose demand is 150 100 0 What is total reward under the optimal policy? Poll!

customer willing to pay 150

Customer demand model

- n possible prices/fares/rates $f_1 > \ldots > f_n$
- Every customer has a willingness-to-pay $\in \{f_1, \ldots, f_n\}$
- T short time periods
  - At most 1 request per period
  - Probability of request depends on class and time: $\lambda_t(i)$
  - $\Sigma_i \lambda_t(i) \leq 1$ for all t
  - Models "inhomogeneous Poisson process"

Dynamic programming model

- Capacity C, $\mathcal{X} = \{0,\dots,C\}$, remaining capacity
- x>0:
  - $\mathcal{A}_x = \{1,\dots,n\}$, price to use
  - $p_t(x-1|x,a) = \Sigma_{i=1}^{a} \lambda_t(i)$, $p_t(x|x,a) = 1 - \Sigma_{i=1}^{a} \lambda_t(i)$
  - $r_t(x,a) = f_a \times P(\text{customer willing to pay } f_a) = f_a \Sigma_{i=1}^{a} \lambda_t(i)$
- x=0:
  - $\mathcal{A}_0 = \{0\}$, $p_t(0|0,0) = 1$, $r_t(0,0) = 0$
- $\mathcal{T} = \{1,\dots,T+1\}$, T+1 is departure moment: $V_{T+1}(x)=0\ \forall\ x\in\mathcal{X}$
- Objective: compute $V_1(C)$
- $V_t(0) = 0 + V_{t+1}(0) = 0$, $1\le t\le T$
- $V_t(x)$ for x>0, $1\le t\le T$:

$$V_t(x) = \max_{a=1,\dots,n} \left\{ f_a \sum_{i=1}^{a} \lambda_t(i) + \sum_{i=1}^{a} \lambda_t(i) V_{t+1}(x-1) + \left(1 - \sum_{i=1}^{a} \lambda_t(i)\right) V_{t+1}(x) \right\}$$

Restrictions on policy

- Prices will typically fluctuate
  - Can go up or down
- How to avoid prices going down?
  - add state component to remember last price
- How to avoid prices going down too often?
  - add penalty for price going down

Multiple dimensions

- Many airlines have networks
- Demand for multiple resources at a time (e.g., BCN → AMS → JFK)
- State becomes multi-dim: $x = (x_1,\ldots,x_n)$
- Number of states is exponential in dimension → too big to compute
- Eg, n flights, each capacity C: $(C+1)^n$ states
  - Example $C=n=100 \rightarrow 10^{200}$ > atoms in universe^2
- Bellman's curse of dimensionality
  - approximation methods required

Forecasting and learning

- Crucial: right values for $\lambda_t(i)$
- Can be separate activity: forecasting (part of statistics)
- Can also be seen as partly unknown values that you learn while bookings arrive: partial information models
  - online learning
- Some airlines work this way: on the basis of bookings they adapt λ's

Models for time

- Finite horizon, $\mathcal{T}=\{0,\ldots,T\}$, total reward
  - Good examples: revenue management, knapsack
  - Bad examples because no clear T: shortest path, inventory mgmt
- Infinite horizon, $\mathcal{T}=\{0,1,\ldots\}$, total reward
  - Direct rewards must eventually get 0, otherwise not defined ($\pm\infty$)
  - E.g., shortest path
  - "Equivalent" to finite horizon with T big
- Infinite horizon, average reward
- Infinite horizon, discounted reward
  - both candidates for inventory mgmt
- Continuous time, infinite horizon, $\mathcal{T}=[0,\infty)$, discounted or average

This lecture: Markov chains = background for long-run average reward

Markov chains

- Time: $t \in \mathcal{T} = \{0,1,2,\ldots\}$
- States: $|\mathcal{X}| < \infty$, X_t is state at t
- No actions or rewards
- Transitions: $p(y|x)=P(X_{t+1}=y|X_t=x)$
- Initial distribution: π_0, $P(X_0=x) = \pi_0(x)$
- Goal: What is distribution at t = X_t?

MCs: distribution at t

- Reminder probability: $P(A|B)=P(AB)/P(B)$,

  $$P(A)=P(AB)+P(AB^c)=P(A|B)P(B)+P(A|B^c)P(B^c)$$

  = law of total probability
- Applied to π_1:

  $$\pi_1(y) = P(X_1=y) = \Sigma_x P(X_1=y|X_0=x)\, P(X_0=x) = \Sigma_x P(X_0=x)\, p(y|x) = \Sigma_x \pi_0(x)\, p(y|x)$$

- Recursion:

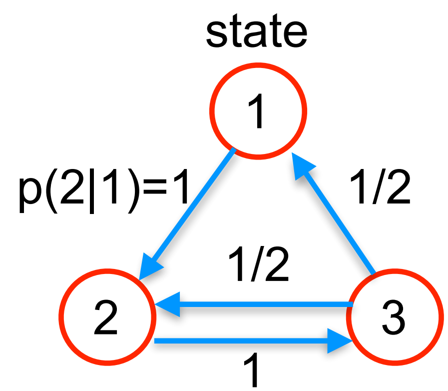
  $$\pi_{t+1}(y) = P(X_{t+1}=y) = \Sigma_x P(X_t=x)\, p(y|x) = \Sigma_x \pi_t(x)\, p(y|x)$$

- Matrix notation: $\pi_{t+1}^T = \pi_t^T P$ with P matrix with $P_{xy}=p(y|x)$

  (note: $p(y|x)$ is sometimes written as $p(x,y)$ or even p_{xy})

Example

- $\mathcal{X}$= {1,2,3}
- p(2|1)=p(3|2)=1, p(1|3)=p(2|3)=1/2



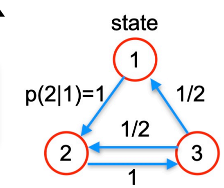
$$\pi_{t+1}(1) = \sum_x \pi_t(x)p(1 \mid x) = \pi_t(3)/2,\ \pi_{t+1}(2) = \pi_t(1) + \pi_t(3)/2,\ \pi_{t+1}(3) = \pi_t(2)$$

- What is π_4 for π_0 = (1,0,0)? Poll!
- Simulation is alternative: simulate many times, take frequencies
  - traces 1,2,3,1,2 and 1,2,3,2,3 both occur with probability 0.5

MCs: properties

- Definition: A path is chain of states $x_1x_2\ldots x_n$ for which $p(x_{k+1}|x_k)>0$
- A MC is
  - communicating: $\exists$ path between any 2 states
  - aperiodic: the gcd (greatest common divisor) of lengths of all paths from x to x = 1 $\forall$ x

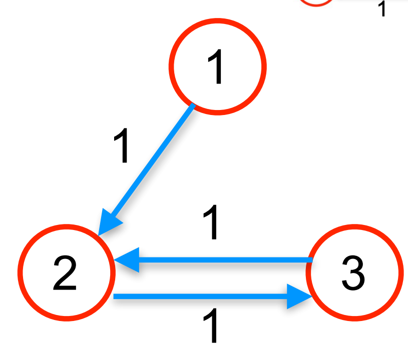
– How about the MC in the example? Poll!



– And how about this MC? Poll!

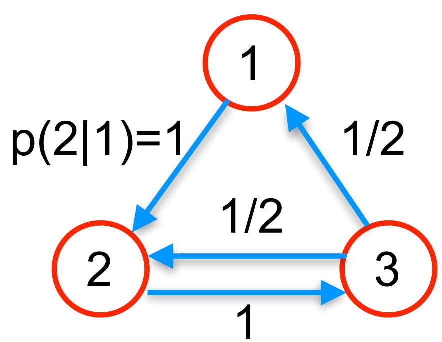
- What is π_4 for $\pi_0 = (1,0,0)$?

Poll!

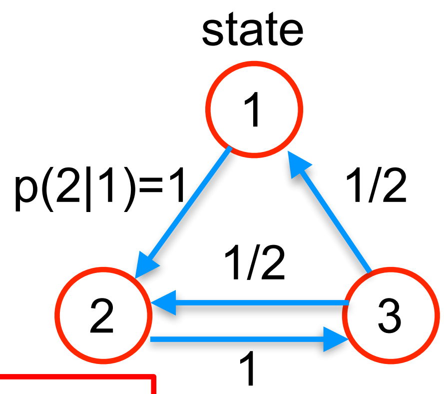


MCs: long-run behavior

- Time-average distribution : $\lim_{t\to\infty} t^{-1} \Sigma_{k=1}^{t} \pi_k$
- Limiting distribution : $\lim_{t\to\infty} \pi_t$
- Stationary distribution: solution of $\pi^T = \pi^T P$ with $\pi^T e = 1$
- **Theorem**: For aperiodic & communicating MCs the time-average, limiting and stationary distributions exist, are the same, unique and independent of π_0 (called π_*)
- What is the stationary distribution of the 1st example? Poll!



Example



- Limiting distribution:

```
pi0=c(1,0,0); T=10000
for(t in 1:T){
  pi1=c(1/2*pi0[3],pi0[1]+1/2*pi0[3],pi0[2])
  pi0=pi1
}
pi1
```

```
[1] 0.2 0.4 0.4
```

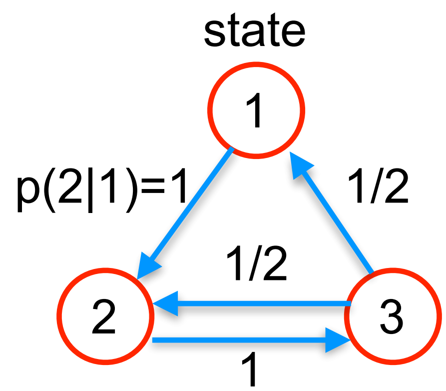
- Time-average distribution gives the same
  - Also when averaged over a simulation = frequencies

```
state=1; visits=rep(0,3)
for(t in 1:T){
  visits[state]=visits[state]+1
  if(state==3){
    state=sample(2,1)
  }else{
    state=state+1
  }
}
visits/T
```

```
[1] 0.198 0.401 0.401
```

by the law of large numbers

Example: stationary distribution



$$\pi^T = \pi^T P$$

$$\Leftrightarrow \pi(y) = \sum_x \pi(x) p(y \,|\, x) = \sum_x \pi(x) p_{xy}$$

$$\Leftrightarrow \pi(1) = 0.5\pi(3),\ \pi(2) = \pi(1) + 0.5\pi(3),\ \pi(3) = \pi(2)$$

$$\pi^T e = 1 \Leftrightarrow \pi(1) + \pi(2) + \pi(3) = 1$$

(0.2,0.4,0.4) is unique solution

(3 variables, 4 equations, but 1 is redundant)

Answers to polls

Polls

RM 2 epochs

1. What are your consecutive prices? (Single Choice) *

- [ ] 100 & 100
- [ ] 100 & 150
- [ ] 150 & 100
- [x] 150 & 150

2. What are your consecutive prices? (Single Choice) *

- [ ] 100 & 100 if still availability
- [ ] 100 & 150 if still availability
- [x] 150 & 100 if still availability
- [ ] 150 & 150 if still availability

3. What is the value? (Single Choice) *

- [ ] 80
- [x] 108
- [ ] 120
- [ ] 140

Polls

RM 3 epochs

1. Total reward for demand 150, 100 and 0? (Single Choice) *

- [ ] 100
- [x] 150
- [ ] 250
- [ ] 400

Polls

Markov chain

1. What is pi_4? (Single Choice) *

- [ ] (0,1,0)
- [x] (0,1/2,1/2)
- [ ] (1/2,0,1/2)
- [ ] (1/2,1/4,1/2)

Polls

long run

1. What is the stationary distribution of the first example? (Single Choice) *

- [ ] (1/3,1/3,1/3)
- [ ] (1/4,1/4,1/2)
- [x] (1/5,2/5,2/5)
- [ ] (1/6,2/6,1/2)

Polls

paths

1. The MC is (Single Choice) *

- [ ] periodic and communicating
- [x] aperiodic and communicating
- [ ] periodic and non-communicating
- [ ] aperiodic and non-communicating

2. The second MC is (Single Choice) *

- [ ] periodic and communicating
- [ ] aperiodic and communicating
- [x] periodic and non-communicating
- [ ] aperiodic and non-communicating

3. What is pi_4? (Single Choice) *

- [ ] (0,1,0)
- [x] (0,0,1)
- [ ] (0,1/2,1/2)
- [ ] (1/2,0,1/2)