EE365: Markov Chains

Markov chains

Transition Matrices

Distribution Propagation

Other Models

Markov chains

Markov chains

- a model for dynamical systems with possibly uncertain transitions
- very widely used, in many application areas
- one of a handful of core effective mathematical and computational tools
- often used to model systems that are not random; *e.g.*, language

Stochastic dynamical system representation

$$x_{t+1} = f(x_t, w_t) \qquad t = 0, 1, \ldots$$

- $x_0, w_0, w_1, \ldots$ are independent random variables
- state transitions are nondeterministic, uncertain
- combine system $x_{t+1} = f(x_t, u_t, w_t)$ with policy $u_t = \mu_t(x_t)$

Example: Inventory model with ordering policy

$$x_{t+1} = [x_t - d_t + u_t]_{[0,C]} \qquad u_t = \mu(x_t)$$

- $x_t \in \mathcal{X} = \{0, \ldots, C\}$ is inventory level at time t
- $C > 0$ is capacity
- $d_t \in \{0, 1, \ldots\}$ is demand that arrives just after time t
- $u_t \in \{0, 1, \ldots\}$ is new stock added to inventory
- $\mu : \mathcal{X} \to \{0, 1, \ldots, C\}$ is ordering policy
- $[z]_{[0,C]} = \min(\max(z, 0),\ C)$ (z clipped to interval $[0, C]$)
- assumption: $d_0, d_1, \ldots$ are independent

Example: Inventory model with ordering policy

$$x_{t+1} = [x_t - d_t + u_t]_{[0,C]} \qquad u_t = \mu(x_t)$$

- capacity $C = 6$
- demand distribution $p_i = \mathbf{Prob}(d_t = i)$

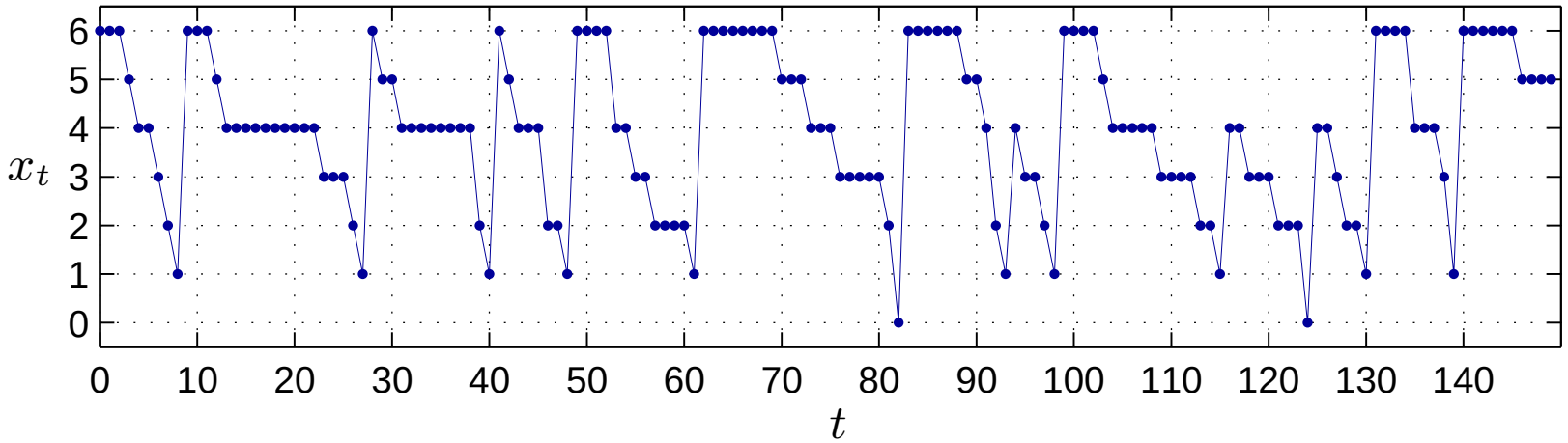
$$p = \begin{bmatrix} 0.7 & 0.2 & 0.1 \end{bmatrix}$$

- policy: refill if $x_t \leq 1$

$$\mu(x) = \begin{cases} C - x & \text{if } x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- start with full inventory $x_0 = C$

Simulation

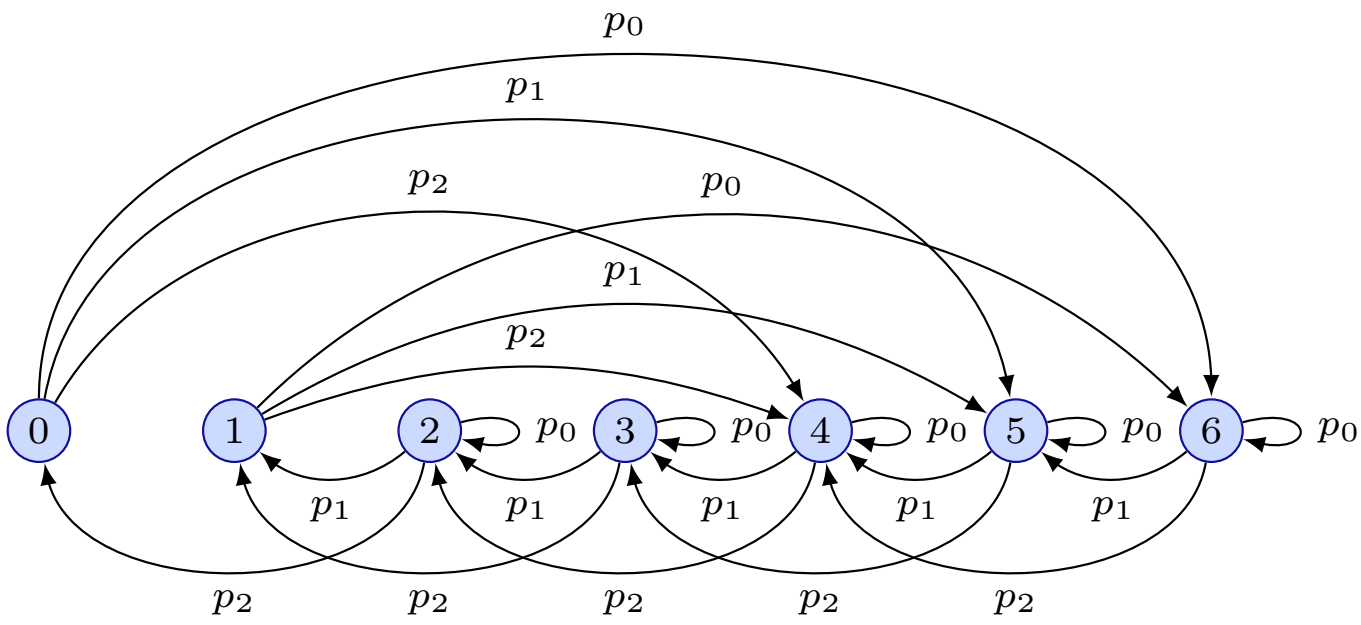


- given initial state x_0
- for $t = 0, \ldots, T-1$
  - simulate process noise $w_t := \texttt{sample}(p)$
  - update the state $x_{t+1} = f(x_t, w_t)$

Simulation

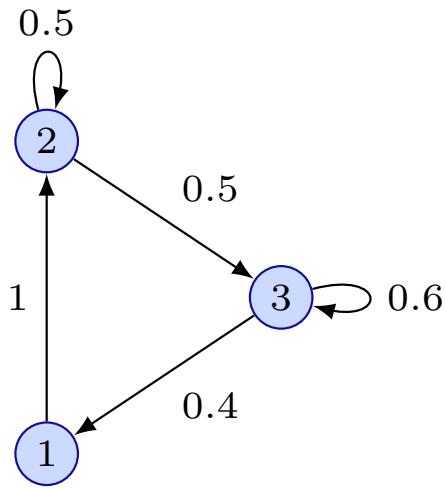
- called a *particle simulation*
- gives a sample of the joint distribution of $(x_0, \ldots, x_T)$
- useful
  - to approximately evaluate probabilities and expectations via Monte Carlo simulation
  - to get a feel for what trajectories look like (*e.g.*, for model validation)

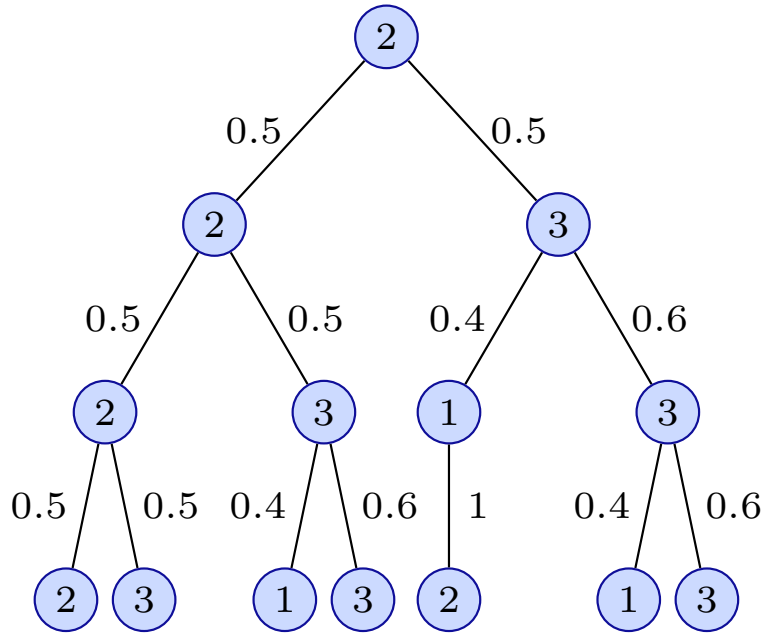
Graph representation



- called the *transition graph*
- each vertex (or node) corresponds to a state
- edge $i \to j$ is labeled with *transition probability* $\mathbf{Prob}(x_{t+1} = j \mid x_t = i)$

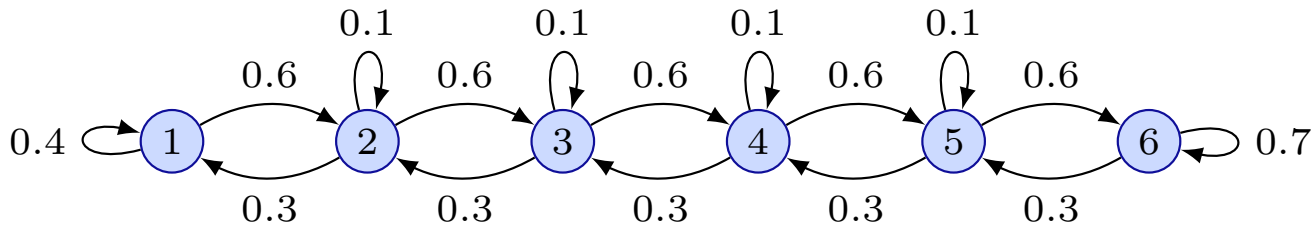
The tree





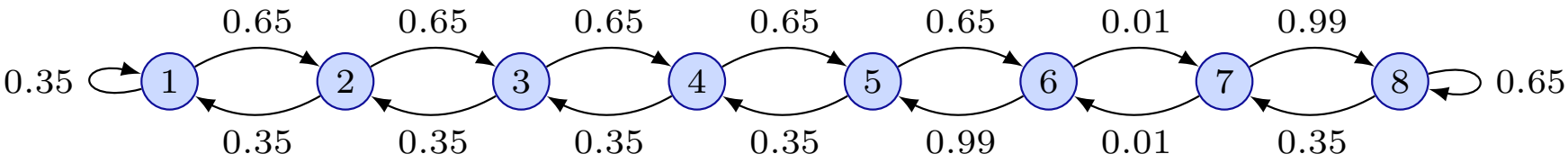
- each vertex of the tree corresponds to a path
- *e.g.*, $\mathbf{Prob}(x_3 = 1 \,|\, x_0 = 2) = 0.5 \times 0.5 \times 0.4 + 0.5 \times 0.6 \times 0.4$

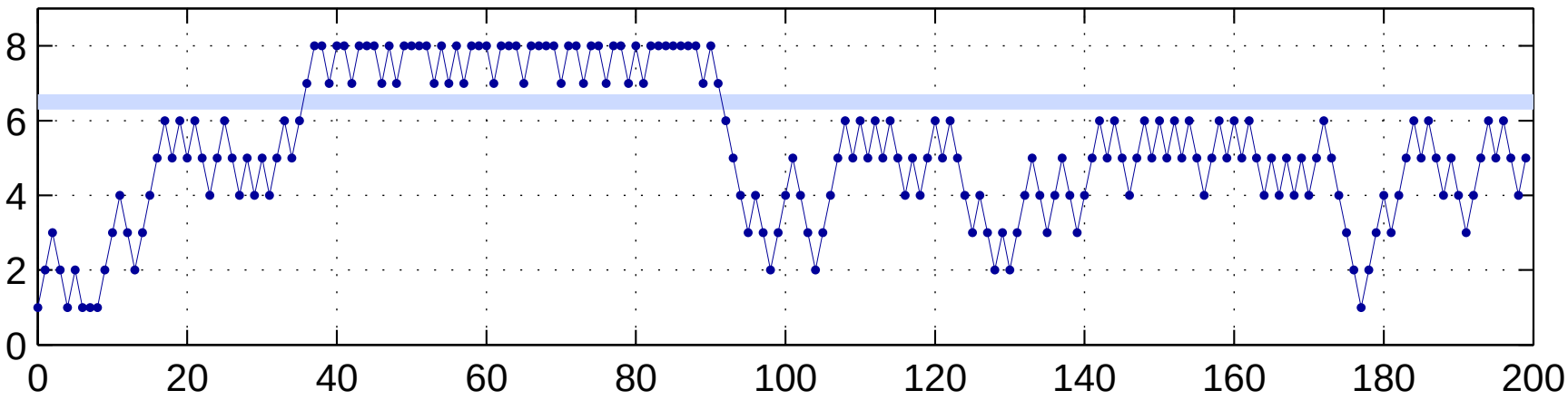
The birth-death chain



self-loops can be omitted, since they can be figured out from the outgoing probabilities

Example: Metastability





Transition Matrices

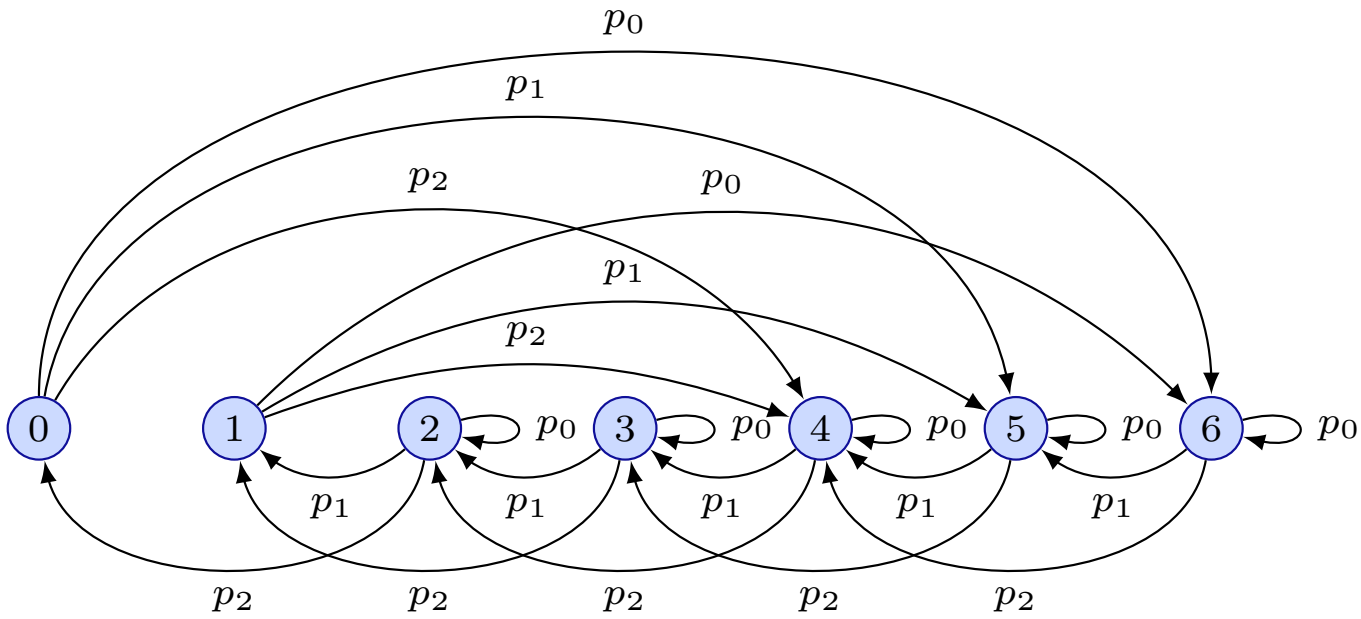
Transition matrix representation

we define the *transition matrix* $P \in \mathbb{R}^{n \times n}$

$$P_{ij} = \mathbf{Prob}(x_{t+1} = j \mid x_t = i)$$

- $P\mathbf{1} = \mathbf{1}$ and $P \geq 0$ elementwise
- a matrix with these two properties is called a *stochastic matrix*
- if P and Q are stochastic, then so is PQ

Transition matrix representation



transition matrix

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 \\ 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 \\ 0.1 & 0.2 & 0.7 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0.2 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.2 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0.2 & 0.7 & 0 \\ 0 & 0 & 0 & 0 & 0.1 & 0.2 & 0.7 \end{bmatrix}$$

Particle simulation given the transition matrix

Given

- the distribution of initial states $d \in \mathbb{R}^{1 \times n}$
- the transition matrix $P \in \mathbb{R}^{n \times n}$, with rows $p_1, \ldots, p_n$

Algorithm:

- choose initial state $x_0 := \texttt{sample}(d)$
- for $t = 0, \ldots, T-1$
  - find the distribution of the next state $d_{\text{next}} := p_{x_t}$
  - sample the next state $x_{t+1} := \texttt{sample}(d_{\text{next}})$

Definition of a Markov chain

sequence of random variables $x_t : \Omega \to \mathcal{X}$ is a *Markov chain* if, for all $s_0, s_1, \ldots$ and all t,

$$\mathbf{Prob}(x_{t+1} = s_{t+1} | x_t = s_t, \ldots, x_0 = s_0) = \mathbf{Prob}(x_{t+1} = s_{t+1} | x_t = s_t)$$

- called the *Markov property*
- means that the system is *memoryless*
- x_t is called the state at time t; $\mathcal{X}$ is called the *state space*
- if you know current state, then knowing past states doesn't give additional knowledge about next state (or any future state)

The joint distribution

the joint distribution of $x_0, x_1, \ldots, x_t$ factorizes according to

$$\mathbf{Prob}\big(x_0 = a, x_1 = b, x_2 = c, \ldots, x_t = q\big) = d_a P_{ab} P_{bc} \cdots P_{pq}$$

because the chain rule gives

$$\begin{aligned}
&\mathbf{Prob}(x_t = s_t, x_{t-1} = s_{t-1}, \ldots, x_0 = s_0) \\
&= \mathbf{Prob}(x_t = s_t \,|\, x_{t-1} = s_{t-1}, \ldots, x_0 = s_0)\, \mathbf{Prob}(x_{t-1} = s_{t-1}, \ldots, x_0 = s_0)
\end{aligned}$$

abbreviating $x_t = s_t$ to just x_t, we can repeatedly apply this to give

$$\begin{aligned}
&\mathbf{Prob}(x_t, x_{t-1}, \ldots, x_0) \\
&= \mathbf{Prob}(x_t \,|\, x_{t-1}, \ldots, x_0)\, \mathbf{Prob}(x_{t-1} \,|\, x_{t-2}, \ldots, x_0) \cdots \mathbf{Prob}(x_1 \,|\, x_0)\, \mathbf{Prob}(x_0) \\
&= \mathbf{Prob}(x_t \,|\, x_{t-1})\, \mathbf{Prob}(x_{t-1} \,|\, x_{t-2}) \cdots \mathbf{Prob}(x_1 \,|\, x_0)\, \mathbf{Prob}(x_0)
\end{aligned}$$

The joint distribution

- the joint distribution completely specifies the process; for example

$$\mathbf{E}\, f(x_0, x_1, x_2, x_3) = \sum_{a,b,c,d \in \mathcal{X}} f(a, b, c, d)\, d_a P_{ab} P_{bc} P_{cd}$$

- *in principle* we can compute the probability of any event and the expected value of any function, but this requires a sum over n^T terms
- we can compute some expected values far more efficiently (more later)

Distribution Propagation

Distribution propagation

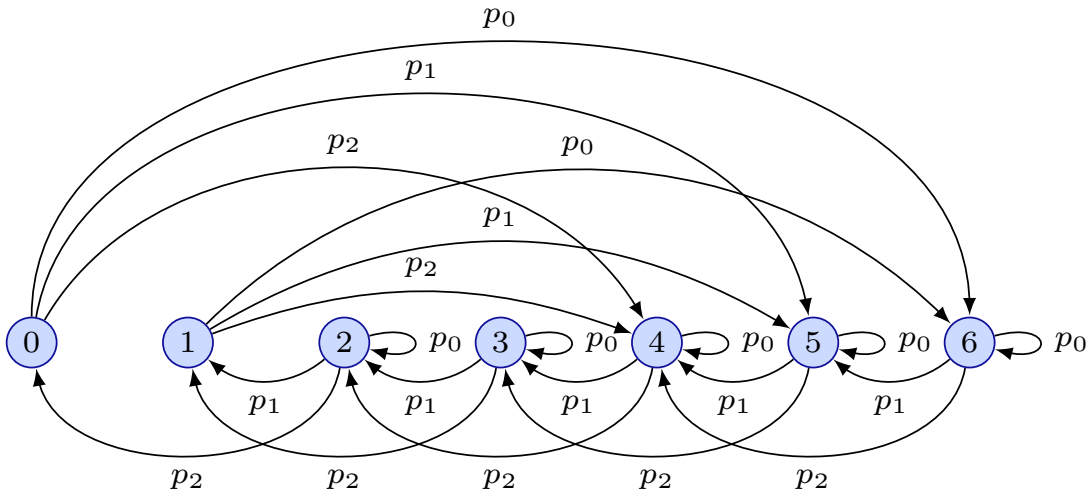
$$\pi_{t+1} = \pi_t P$$

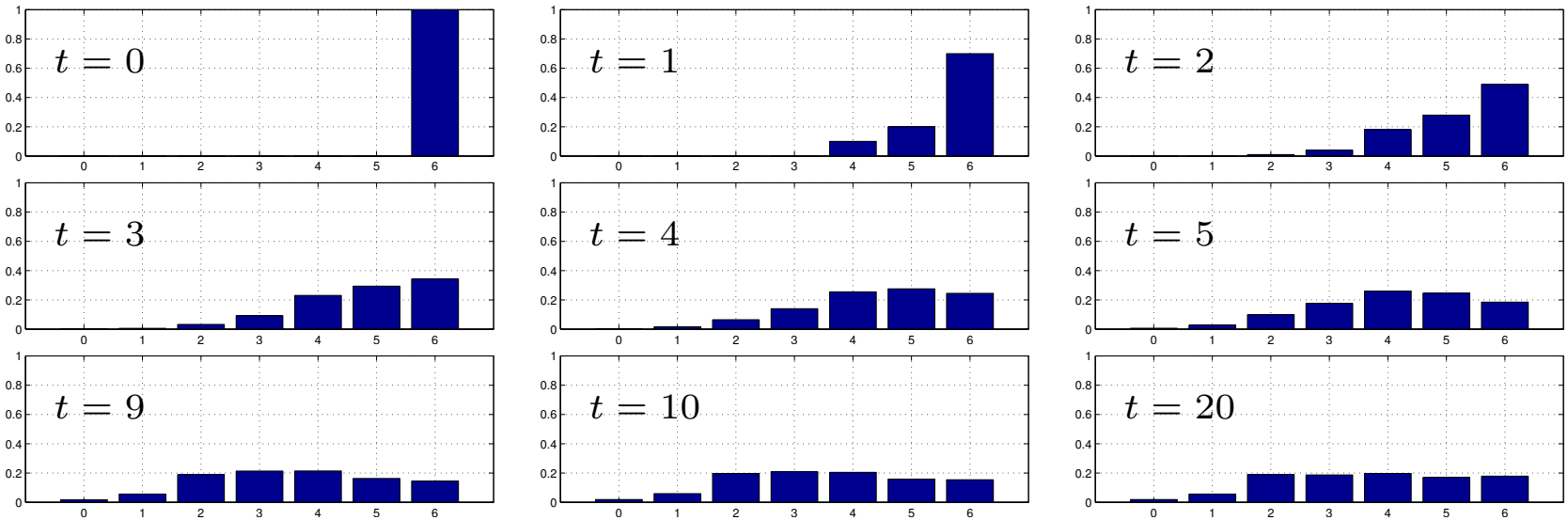
- here π_t denotes distribution of x_t, so

$$(\pi_t)_i = \mathbf{Prob}(x_t = i)$$

- can compute marginals $\pi_1, \ldots, \pi_T$ in Tn^2 operations
- a useful type of simulation ...

Example: Distribution propagation

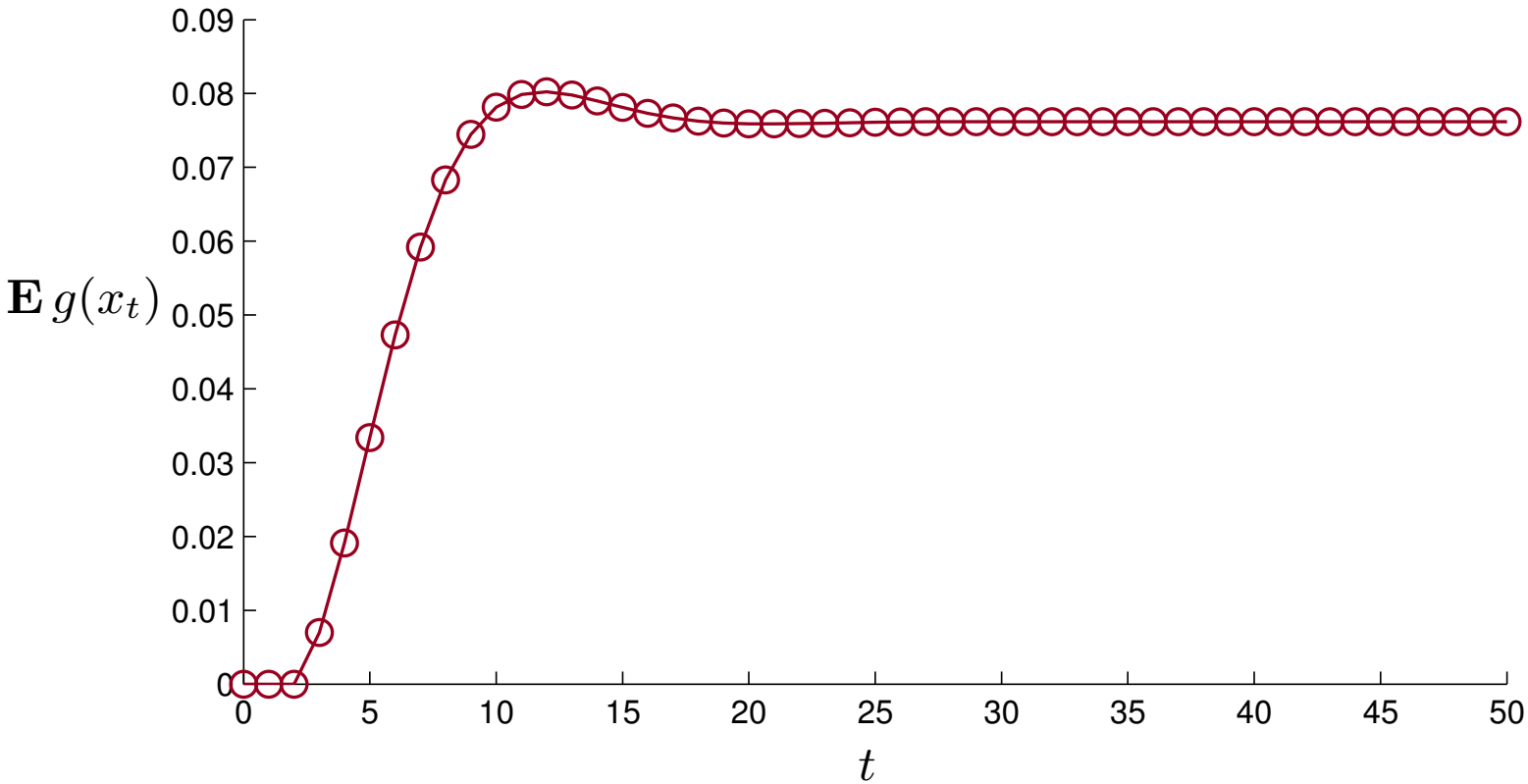




Example: Reordering

- what is the probability we reorder at time t?
- compute π_t by distribution propagation, and use

$$\mathbf{Prob}(x_t \in \{0,1\}) = \pi_t \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^{\mathsf{T}}$$



Evaluating expectation of separable function

- suppose $f : \mathcal{X}^{T+1} \to \mathbb{R}$ is separable, *i.e.*,

$$f(x_0, \ldots, x_T) = f_0(x_0) + \cdots + f_T(x_T)$$

  where $f_t : \mathcal{X} \to \mathbb{R}$

- then we have

$$\mathbf{E}\, f(x) = \pi_0 f_0 + \cdots + \pi_T f_T$$

- using distribution propagation, we can compute exactly with Tn^2 operations

Powers of the transition matrix

as an example of the use of the joint distribution, let's show

$$(P^k)_{ij} = \mathbf{Prob}\big(x_{t+k} = j \,|\, x_t = i\big)$$

this holds because

$$\mathbf{Prob}\big(x_{t+k} = s_{t+k} \,|\, x_t = s_t\big) = \frac{\mathbf{Prob}\big(x_{t+k} = s_{t+k} \text{ and } x_t = s_t\big)}{\mathbf{Prob}\big(x_t = s_t\big)}$$

$$= \frac{\displaystyle\sum_{s_0, s_1, \ldots, s_{t-1}, s_{t+1}, \ldots, s_{t+k-1}} d_{s_0} P_{s_0 s_1} P_{s_1 s_2} \cdots P_{s_{t+k-1} s_{t+k}}}{\mathbf{Prob}\big(x_t = s_t\big)}$$

$$= \frac{\left(\displaystyle\sum_{s_0, s_1, \ldots, s_{t-1}} d_{s_0} P_{s_0 s_1} \cdots P_{s_{t-1} s_t}\right)\left(\displaystyle\sum_{s_{t+1}, \ldots, s_{t+k-1}} P_{s_t s_{t+1}} \cdots P_{s_{t+k-1} s_{t+k}}\right)}{\mathbf{Prob}\big(x_t = s_t\big)}$$

Example: k step transition probabilities

two ways to compute $\mathbf{Prob}(x_t = j \mid x_0 = i)$

- direct method: sum products of probabilities over n^t state sequences
- using matrix powers: $\mathbf{Prob}(x_t = j \mid x_0 = i) = (P^t)_{ij}$
- we can compute this in
  - $O(n^3 t)$ arithmetic operations (by multiplying matrices to get P)
  - $O(n^3)$ arithmetic operations (by first diagonalizing P)

Other Models

Time-varying Markov chains

- we may have a *time-varying* Markov chain, with one transition matrix for each time

$$(P_t)_{ij} = \mathbf{Prob}(x_{t+1} = j \mid x_t = i)$$

- suppose $\mathbf{Prob}(x_t = a) \neq 0$ for all $a \in \mathcal{X}$ and t

  then the factorization property that there exists stochastic matrices $P_0, P_1, \ldots$ and a distribution d such that

$$\mathbf{Prob}\big(x_0 = a, x_1 = b, x_2 = c, \ldots, x_t = q\big) = d_a (P_0)_{ab} (P_1)_{bc} \cdots (P_{t-1})_{pq}$$

  is *equivalent* to the Markov property

- the theory of factorization properties of distributions is called *Bayesian networks* or *graphical models* (with Markov chains the simplest case)

Dynamical system representation revisited

$$x_{t+1} = f(x_t, w_t)$$

If $x_0, w_0, w_1, \ldots$ are *independent* then $x_0, x_1, \ldots$ is Markov

Dynamical system representation

To see that $x_0, x_1, \ldots$ is Markov, notice that

$$\begin{aligned}
&\mathbf{Prob}(x_{t+1} = s_{t+1} \,|\, x_0 = s_0, \ldots, x_t = s_t) \\
&= \mathbf{Prob}\big(f(s_t, w_t) = s_{t+1} \,|\, x_0 = s_0, \ldots, x_t = s_t\big) \\
&= \mathbf{Prob}\big(f(s_t, w_t) = s_{t+1} \,|\, \phi_t(x_0, w_0, \ldots, w_{t-1}) = (s_0, \ldots, s_t)\big) \\
&= \mathbf{Prob}\big(f(s_t, w_t) = s_{t+1}\big)
\end{aligned}$$

- $(x_0, \ldots, x_t) = \phi_t(x_0, w_0, \ldots, w_{t-1})$ follows from $x_{t+1} = f(x_t, w_t)$
- last line holds since $w_t \perp\!\!\!\perp (x_0, w_0, \ldots, w_{t-1})$ and $x \perp\!\!\!\perp y \implies f(x) \perp\!\!\!\perp g(y)$

$x_0, x_1, \ldots$ is Markov because the above approach also gives

$$\mathbf{Prob}(x_{t+1} = s_{t+1} \,|\, x_t = s_t) = \mathbf{Prob}\big(f(s_t, w_t) = s_{t+1}\big)$$

State augmentation

we often have models

$$x_{t+1} = f(x_t, w_t)$$

where $w_0, w_1, \ldots$ is Markov

- $w_0, w_1, \ldots$ satisfy $w_{t+1} = g(w_t, v_t)$ where $v_0, v_1, \ldots$ are *independent*
- construct an equivalent Markov chain, with state $z_t = \begin{bmatrix} x_t \\ w_t \end{bmatrix}$
- dynamics $z_{t+1} = h(z_t, v_t)$ where

$$h(z, v) = \begin{bmatrix} f(z_1, z_2) \\ g(z_2, v) \end{bmatrix}$$

Markov k

$$x_{t+1} = f(x_t, x_{t-1}, \ldots, x_{t-k+1}, w_t)$$

- called a *Markov-k* model
- assume $(x_0, x_1, \ldots, x_{k-1}), w_0, w_1, w_2, \ldots$ are independent
- initial distribution on $(x_0, x_1, \ldots, x_{k-1})$
- Markov-k property: for all $s_0, s_1, \ldots,$

$$\begin{aligned}&\mathbf{Prob}(x_{t+1} = s_{t+1} | x_t = s_t, \ldots, x_0 = s_0) \\ &\quad = \mathbf{Prob}(x_{t+1} = s_{t+1} | x_t = s_t, x_{t-1} = s_{t-1}, \ldots, x_{t-k+1} = s_{t-k+1})\end{aligned}$$

Markov k

$$x_{t+1} = f(x_t, x_{t-1}, \ldots, x_{t-k+1}, w_t)$$

- equivalent Markov system, with state $z_t \in \mathcal{Z} = \mathcal{X}^k$; notice that $|\mathcal{Z}| = |\mathcal{X}|^k$

$$z_t = \begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-k+1} \end{bmatrix}$$

- dynamics $z_{t+1} = g(z_t, w_t)$ where

$$g(z, w) = \begin{bmatrix} f\big(z_1, z_2, \ldots, z_k, w\big) \\ z_1 \\ \vdots \\ z_{k-1} \end{bmatrix}$$

Markov language models

- given text, construct a Markov k model by counting the frequency of word sequences
- applications:
  - run particle simulation to generate new text
  - suggest correct spellings when spell checking
  - optical character recognition
  - finding similar documents in information retrieval

Example: Markov language models

- example: approximately 78,000 words, 8,200 unique

He wasnt true and disappointment that wouldnt be very old school song at once it on the grayish white stone well have charlies work much to keep your wand still had a mouse into a very scratchy whiskery kiss. Then yes yes. I hope im pretty but there had never seen me what had to vacuum anymore because its all right were when flint had shot up while uncle vernon around the sun rose up to. Bill used to. Ron who to what he teaches potions one. I suppose they couldnt think about it harry hed have found out that didnt trust me all sorts of a letter which one by older that was angrier about what they would you see here albus dumbledore it was very well get a rounded bottle of everyone follows me an awful lot his face the minute. Well send you get out and see professor snape. But it must have his sleeve. I am i think. Fred and hermione and dangerous said finally. Uncle vernon opened wide open. As usual everyones celebrating all hed just have caught twice. Yehll get out. Well get the snitch yet no way up his feet.

Markov language models

- another example

Here on this paper. Seven bottles three are poison two are wine one will get us through the corridors. Even better professor flitwick could i harry said the boy miserably. Well i mean its the girls through one of them hardly daring to take rons mind off his rocker said ron loudly. Shut up said ron sounding both shocked and impressed. Id have managed it before and you should have done so because a second later hermione granger telling a downright lie to a baked potato when professor quirrell too the nervous young man made his way along the floor in front of the year the dursleys were his only chance to beat you to get past fluffy i told her you didnt say another word on the wall wiping sweat off his nerves about tomorrow. Why should he yet harry couldnt help trusting him. This is goyle said the giant turning his back legs in anger. For the food faded from the chamber theyd just chained up. Oh no it isnt the same thing. Harry who was quite clear now and gathering speed. The air became colder and colder as they jostled their way through a crowd.