Probabilistic Programming

Lecture #3: Markov Chains

Joost-Pieter Katoen

RWTH Lecture Series on Probabilistic Programming 2022-23

Overview

Overview

Probability theory is simple, isn't it?

In no other branch of mathematics
is it so easy to make mistakes
as in probability theory

Henk Tijms, "Understanding Probability" (2004)

Measurable space

Sample space

A *sample space* Ω of a chance experiment is a set of elementary events that have a 1-to-1 relationship to the possible outcomes of the experiment.

σ-algebra

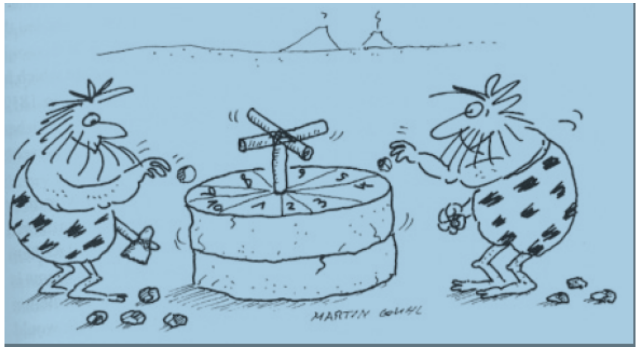
A *σ-algebra* on a set Ω is a collection $\mathcal{F} \subset 2^{\Omega}$ of subsets of Ω such that:

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \;\Rightarrow\; \Omega - A \in \mathcal{F}$ — complement
3. $(\forall i \geqslant 0.\ A_i \in \mathcal{F}) \;\Rightarrow\; \bigcup_{i \geqslant 0} A_i \in \mathcal{F}$ — countable union

Elements in $\mathcal{F}$ are called *measurable sets*, and the pair $(\Omega, \mathcal{F})$ is called a *measurable space* or *Borel space*.

For set Ω, $\mathcal{F} = \{\varnothing, \Omega\}$ yields the smallest σ-algebra; and $\mathcal{F} = 2^{\Omega}$ the largest one.

Probabilities



Probability space

Measure and probability space

Let $(\Omega, \mathcal{F})$ be a measurable space. The function $\mathcal{F} \to \mathbb{R}_{\geqslant 0}$ is a *measure* if:

1. $Pr(\varnothing) = 0$, and
2. for any countable sequence $\{A_i \in \mathcal{F}\}_{i=1,2,\ldots}$ with $A_i \cap A_j = \varnothing, i \neq j$:

$$Pr\left(\bigcup_{i \in I} A_i\right) = \sum_{i=1}^{\infty} Pr(A_i) \quad \text{, i.e., } Pr \text{ is } \sigma\text{-additive}$$

The measure Pr is a *probability measure* if $Pr(\Omega) = 1$.

Then, $(\Omega, \mathcal{F}, Pr)$ is called a *probability space*.

Conditional probabilities

Conditional probability

If $Pr(A) > 0$, $Pr(B \mid A) = \dfrac{Pr(A \cap B)}{Pr(A)}$ is the *conditional probability* of B given A.

Exercise

Show that $(\Omega, \mathcal{F}, Pr(\cdot | A))$ is a probability space for any given $A \in \mathcal{F}$.

Law of total probability

Let $A_1 + A_2 + \ldots + A_n = \Omega$ and $A_i \cap A_j = \varnothing$ if $i \neq j$. Then for any $B \in \mathcal{F}$:

$$Pr(B) = \sum_{i=1}^{n} Pr(B \mid A_i) \cdot Pr(A_i)$$

(Proof left as an exercise.)

Elementary properties

For measurable events A, B and probability measure Pr:

- $Pr(A) + Pr(\underbrace{\Omega - A}_{\neg A}) = Pr(\Omega) = 1$
- The *sum* rule:

$$Pr(A) \;=\; Pr(A \cap B) + Pr(A \cap \neg B)$$

- The *product* rule:

$$Pr(A \cap B) \;=\; Pr(A \mid B) \cdot Pr(B) \;=\; Pr(B \mid A) \cdot Pr(A)$$

  $Pr(A \cap B)$ is also written as $Pr(A, B)$ and is called the *joint* probability of A and B.

Discrete probability space

Discrete probability space

Pr is a *discrete* probability measure on $(\Omega, \mathcal{F})$ if

- there is a countable set $A \subseteq \Omega$ such that for $a \in A$:

$$\{ a \} \in \mathcal{F} \quad \text{and} \quad \sum_{a \in A} Pr(\{ a \}) = 1$$

- e.g., a probability measure on $(\Omega, 2^{\Omega})$

$(\Omega, \mathcal{F}, Pr)$ is then called a *discrete* probability space; otherwise, it is a *continuous probability* space.

Example

Discrete probability spaces: throwing a die, number of customers in a shop, ...

Example

Continuous probability spaces: throwing a dart on a circular board, water tank level, ...

Random variable

Measurable function

Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measurable spaces. Function $f : \Omega \to \Omega'$ is *measurable* if

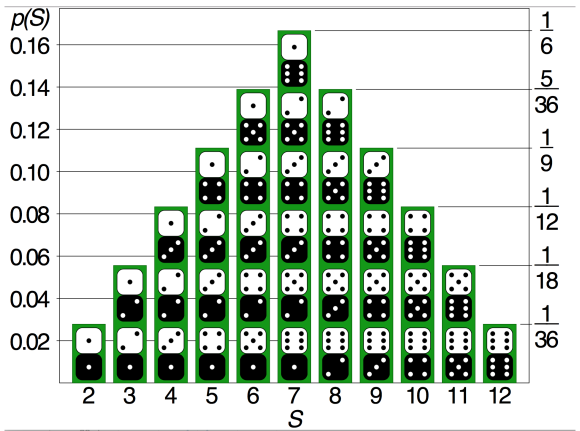
$$f^{-1}(A) \;=\; \{ a \mid f(a) \in A \} \in \mathcal{F} \quad \text{for all } A \in \mathcal{F}'$$

Random variable

Measurable function $X : \Omega \to \mathbb{R}$ is a *random variable*.

The *probability distribution* of X is $Pr_X = Pr \circ X^{-1}$ where Pr is a probability measure on $(\Omega, \mathcal{F})$.

Example: rolling a pair of fair dice



Overview

1. Probability refresher
2. Markov Chains
3. Reachability probabilities
4. State classification
5. Limiting distributions

Probability distribution

Probability distribution

A (discrete) probability distribution on countable set X is a function

$$\mu : X \to [0,1] \subseteq \mathbb{R} \quad \text{such that} \quad \sum_{x \in X} \mu(x) = 1.$$

The set $\{ x \mid \mu(x) > 0 \}$ is the support set of probability distribution μ.

Let $Dist(X)$ denote the set of all probability measures on X.

$\mu : X \to [0,1] \subseteq \mathbb{R}$ with $\sum_{x \in X} \mu(x) \leqslant 1$ is a sub-distribution on X.

Andrei Andrejewitsch Markow

Markov chains

Markov chain

A Markov chain (MC) D is a triple $(\Sigma, \sigma_I, \mathbf{P})$ with:

- Σ being a countable set of states
- $\sigma_I \in \Sigma$ the initial state, and
- $\mathbf{P} : \Sigma \to Dist(\Sigma)$ the transition probability function

where $Dist(\Sigma)$ is a (discrete) probability measure on Σ.

A state $\sigma \in \Sigma$ for which $\mathbf{P}(\sigma, \sigma) = 1$ is called absorbing.

Example

Mathematical perspective

A *Markov chain* is a time-homogeneous Markov process with discrete parameter T and discrete state space Σ.

Transition probability matrix

For MC D with *finite* state space Σ,
the function $\mathbf{P}$ is called the *transition probability matrix* of D.

Properties:

1. $\mathbf{P}$ is a (right) *stochastic* matrix, i.e., it is a square matrix, all its elements are in $[0, 1]$, and each row sum equals one.
2. $\mathbf{P}$ has an eigenvalue of one, and all its eigenvalues are at most one:

$$\lambda \text{ is an eigenvalue of } \mathbf{P} \text{ if } \mathbf{P} \cdot \mathbf{x} \;=\; \lambda \cdot \mathbf{x}$$

3. For all $n \in \mathbb{N}$, $\mathbf{P}^n = \underbrace{\mathbf{P} \cdot \ldots \cdot \mathbf{P}}_{n \text{ times}}$ is a stochastic matrix. Note: $\mathbf{P}^0 = \mathbf{I}$.

Paths

Paths

The infinite sequence $\pi = \sigma_0 \, \sigma_1 \, \ldots \in \Sigma^\omega$ is a *path* through MC D provided $\mathbf{P}(\sigma_i, \sigma_{i+1}) > 0$ for all natural i.

Let *Paths*(D) denotes the set of paths in D that start in its initial state σ_I. Finite paths are prefixes of (infinite) paths.

An alternative, equivalent view.

Let graph $G_D = (\Sigma, E)$ with $(\sigma, \tau) \in E$ if and only if $\mathbf{P}(\sigma, \tau) > 0$.

Then: paths are infinite traversals through the graph G_D.

Example

Cylinder sets

Cylinder set

The *cylinder set* of finite path $\hat{\pi} = \sigma_0\,\sigma_1 \ldots \sigma_n$ in MC D is defined by:

$$Cyl(\hat{\pi}) \;=\; \{\, \pi \in Paths(D) \mid \hat{\pi} \text{ is a prefix of } \pi \,\}.$$

The cylinder set spanned by finite path $\hat{\pi}$ consists of all infinite paths that have prefix $\hat{\pi}$.

Probability measure on sets of infinite paths

Probability measure

Pr is the unique *probability distribution* defined on cylinder sets by:

$$Pr(Cyl(\sigma_0 \ldots \sigma_n)) \;=\; \prod_{0 \leqslant i < n} \mathbf{P}(\sigma_i, \sigma_{i+1})$$

for $n > 0$ and $\mathbf{P}(\sigma_0) = 1$ if and only if $\sigma_0 = \sigma_I$.

By standard results in probability theory, *Pr* is a distribution on all sets of infinite paths that are countable (disjoint) unions and/or complements of cylinder sets.

Overview

Reachability

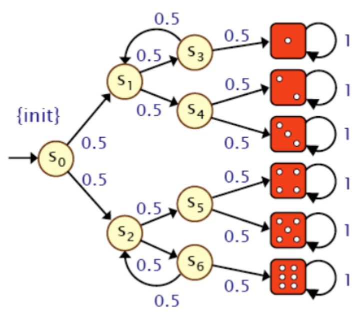
Reachability

Let MC D with (countable) state space Σ and $G \subseteq \Sigma$ the set of *goal* states. The event *eventually reaching* G is defined by:

$$\lozenge G \;=\; \{\, \sigma_0\, \sigma_1\, \sigma_2 \ldots \in \mathit{Paths}(D) \mid \exists i \in \mathbb{N}.\, \sigma_i \in G \,\}$$

The event $\lozenge G$ is *measurable*, i.e., the probability $Pr(\lozenge G)$ is well defined.

Reachability probabilities: Knuth-Yao's die



- Consider the event $\lozenge 4$
- We have:

$$Pr(\lozenge 4) = \sum_{s_0 \ldots s_n \in (\Sigma \setminus 4^*)4} \mathbf{P}(s_0 \ldots s_n)$$

- This yields:
  $\mathbf{P}(s_0 s_2 s_5 4) + \mathbf{P}(s_0 s_2 s_6 s_2 s_5 4) + \ldots\ldots$
- Or: $\sum_{k=0}^{\infty} \mathbf{P}(s_0 s_2 (s_6 s_2)^k s_5 4)$
- Or: $\frac{1}{8} \cdot \sum_{k=0}^{\infty} (\frac{1}{4})^k$
- Geometric series: $\frac{1}{8} \cdot \frac{1}{1 - \frac{1}{4}} \;=\; \frac{1}{8} \cdot \frac{4}{3} \;=\; \frac{1}{6}$

For finite Markov chains, reachability probabilities can be obtained algorithmically.

Reachability probabilities

Problem statement:

Let D be an MC with finite state space Σ, $\sigma \in \Sigma$, and $G \subseteq \Sigma$.

Aim: determine

$$Pr(\sigma \models \lozenge G) \;=\; Pr_\sigma(\lozenge G) \;=\; Pr\{\, \pi \in \mathit{Paths}(D_\sigma) \mid \pi \in \lozenge G \,\}$$

where D_σ is the MC D with initial state σ.

Characterisation of reachability probabilities

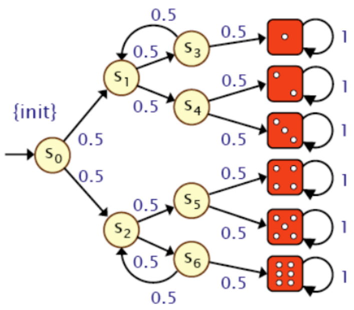
Let variable $x_\sigma = Pr(\sigma \models \lozenge G)$ for any state σ be defined by:

- if $\sigma \notin Pre^*(G)$, then $x_\sigma = 0$
- if $\sigma \in G$, then $x_\sigma = 1$
- otherwise:

$$x_\sigma \;=\; \underbrace{\sum_{\tau \in \Sigma \setminus G} \mathbf{P}(\sigma, \tau) \cdot x_\tau}_{\text{reach } G \text{ via } \tau \in \Sigma \setminus G} \;+\; \underbrace{\sum_{\gamma \in G} \mathbf{P}(\sigma, \gamma)}_{\text{reach } G \text{ in one step}}$$

$Pre^*(G)$ is the set of states in Σ from which G is reachable, i.e., $\{\, \sigma \in \Sigma \mid Pr(\sigma \models \lozenge G) > 0 \,\}$.

Reachability probabilities: Knuth-Yao's die



- Consider the event $\lozenge 4$
- The previous characterisation yields:

$$x_1 = x_2 = x_3 = x_5 = x_6 = 0 \text{ and } x_4 = 1$$
$$x_{s_1} = x_{s_3} = x_{s_4} = 0$$
$$x_{s_0} = \tfrac{1}{2} x_{s_1} + \tfrac{1}{2} x_{s_2}$$
$$x_{s_2} = \tfrac{1}{2} x_{s_5} + \tfrac{1}{2} x_{s_6}$$
$$x_{s_5} = \tfrac{1}{2} x_5 + \tfrac{1}{2} x_4$$
$$x_{s_6} = \tfrac{1}{2} x_{s_2} + \tfrac{1}{2} x_6$$

- Gaussian elimination yields:

$$x_{s_5} = \tfrac{1}{2},\ x_{s_2} = \tfrac{1}{3},\ x_{s_6} = \tfrac{1}{6}, \text{ and } \boxed{x_{s_0} = \tfrac{1}{6}}$$

Linear equation system

- Let $\Sigma_? = Pre^*(G) \setminus G$, the states that can reach G by > 0 steps
- $\mathbf{A} = \big(\mathbf{P}(\sigma, \tau)\big)_{\sigma, \tau \in \Sigma_?}$, the transition probabilities in $\Sigma_?$
- $\mathbf{b} = \big(b_\sigma\big)_{\sigma \in \Sigma_?}$, the probs to reach G in 1 step, i.e., $b_\sigma = \sum_{\gamma \in G} \mathbf{P}(\sigma, \gamma)$

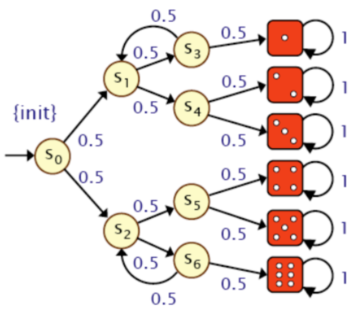
Theorem

The vector $\mathbf{x} = (x_\sigma)_{\sigma \in \Sigma_?}$ with $x_\sigma = Pr(\sigma \models \lozenge G)$ is the unique solution of the linear equation system:

$$\mathbf{x} = \mathbf{A} \cdot \mathbf{x} + \mathbf{b} \quad \text{or, equivalently} \quad (\mathbf{I} - \mathbf{A}) \cdot \mathbf{x} = \mathbf{b}$$

where $\mathbf{I}$ is the identity matrix of cardinality $|\Sigma_?| \cdot |\Sigma_?|$.

Reachability probabilities: Knuth-Yao's die



- Consider the event $\lozenge 4$
- $\Sigma_? = \{ s_0, s_2, s_5, s_6 \}$

$$\begin{pmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_{s_0} \\ x_{s_2} \\ x_{s_5} \\ x_{s_6} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}$$

- Gaussian elimination yields:

$$x_{s_5} = \tfrac{1}{2},\ x_{s_2} = \tfrac{1}{3},\ x_{s_6} = \tfrac{1}{6}, \text{ and } \boxed{x_{s_0} = \tfrac{1}{6}}$$

Computing reachability probabilities

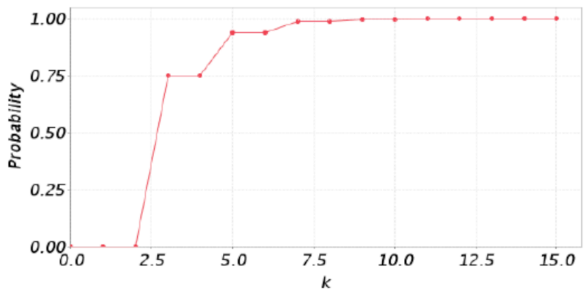
Reachability probabilities in finite MCs can be computed in polynomial time.

Bounded reachability probabilities

k-bounded reachability probabilities

Let MC D with (countable) state space Σ, $G \subseteq \Sigma$, and $k \in \mathbb{N}$. The event *eventually reaching G within k steps* is defined by:

$$\Diamond^{\leqslant k} G \;=\; \{ \sigma_0 \, \sigma_1 \, \sigma_2 \ldots \in Paths(D) \mid \exists i \leqslant k. \, \sigma_i \in G \}$$



k-bounded reachability for reaching 1,2, . . . , 6 in Knuth-Yao's die

Transient probabilities

Transient probabilities

The probability to move from τ to σ in exactly k steps is given by $\mathbf{P}^n(\sigma, \tau)$.

The *transient probability distribution* for k when starting in state σ is given by $\mathbf{P}^n(\sigma, \cdot)$.

Transient probability distributions depend on the (initial) state σ.

Compute the transient distribution for a fragment of Knuth-Yao's die.

Overview

First visit probabilities

For states $\sigma, \tau \in \Sigma$, let

$$f_{\sigma,\tau}^{(n)} = Pr\{ \text{ first visit to } \tau \text{ after exactly } n > 0 \text{ steps from } \sigma \; \}$$

This differs from the probability $\mathbf{P}^n(\sigma, \tau)$ to move from σ to τ in n steps.

We have:

$$\mathbf{P}^n(\sigma, \tau) \;=\; \sum_{\ell=1}^{n} f_{\sigma,\tau}^{(\ell)} \cdot \mathbf{P}^{n-\ell}(\tau, \tau)$$

The probability to reach τ from state σ equals:

$$Pr(\sigma \models \Diamond \tau) \;=\; f_{\sigma,\tau} \;=\; \sum_{n=1}^{\infty} f_{\sigma,\tau}^{(n)}$$

Return probabilities

For state $\sigma \in \Sigma$, let

$$f_\sigma^{(n)} = Pr\{ \text{ first return to } \sigma \text{ after exactly } n \text{ steps } \}$$

We have:

$$f_\sigma^{(n)} \; = \; f_{\sigma,\sigma}^{(n)} \; = \; Pr\{ \text{ first visit to } \sigma \text{ after } n \text{ steps from } \sigma \ \}.$$

The return probability to state σ equals:

$$Pr(\sigma \models \Diamond \sigma) \; = \; f_\sigma \; = \; \sum_{n=1}^{\infty} f_\sigma^{(n)}.$$

Transient and recurrent states

The return probability to σ equals: $Pr(\sigma \models \Diamond \sigma) \; = \; f_{\sigma,\sigma} \; = \; \sum_{n=1}^{\infty} f_\sigma^{(n)}.$

Transient and recurrent states

State σ is called *recurrent* if $f_\sigma = 1$, i.e., with probability one (aka: almost surely) the MC returns to σ.

State σ is called *transient* otherwise, i.e., if $f_\sigma < 1$. With a positive probability, the MC does not return to a transient state.

Example on the black board.

Null and positive recurrence

Let σ be a recurrent state, i.e., $Pr(\sigma \models \Diamond \sigma) = f_\sigma = 1$.

Mean recurrence time

The mean recurrence time of recurrent state σ equals

$$m_\sigma \; = \; \sum_{n=1}^{\infty} n \cdot f_\sigma^{(n)}$$

This is the expected number of steps between two successive visits to σ.

Null and positive recurrent states

State σ is called positive recurrent whenever $m_\sigma < \infty$. Otherwise, state σ is called null recurrent; thus, then $m_\sigma = \infty$.

Example on the black board.

Null and positive recurrence in finite MC

1. Every state in a finite MC is either positive recurrent or transient.
2. At least one state in a finite MC is positive recurrent.
3. A finite MC has no null recurrent states.
   $\Rightarrow$ null recurrence is only relevant for infinite MCs

Connected states are of the same "type"

Let σ and τ be mutually reachable states. Then:

$$\begin{aligned} \sigma \text{ is transient} \quad &\text{iff} \quad \tau \text{ is transient} \\ \sigma \text{ is null-recurrent} \quad &\text{iff} \quad \tau \text{ is null-recurrent} \\ \sigma \text{ is positive recurrent} \quad &\text{iff} \quad \tau \text{ is positive recurrent} \end{aligned}$$

Overview

Limiting distribution

We consider the long-term behaviour of Markov chains.

Limiting distribution

Distribution $\rho = (\rho_0, \rho_1, \ldots)$ is a limiting distribution of MC D with state space Σ if

$$\rho_j \;=\; \lim_{n \to \infty} \mathbf{P}^n(\sigma_i, \sigma_j) \quad \text{for all } \sigma_i, \sigma_j \in \Sigma \text{ and } \sum_j \rho_j = 1.$$

Questions:

1. does the limiting distribution always exist? No. Not for periodic MCs.
2. if it does, is it unique? No. Yes for irreducible, aperiodic MCs.
3. if it is unique, how to compute it? Solve a linear equation system.

Example: limiting distributions do not always exist

Example: limiting distribution may exist

Two structural properties

Irreducible

A MC is irreducible if it is strongly connected, i.e., all states are mutually reachable.

Periodic

Let D be an irreducible MC and let d be the largest natural number such that for all states σ in D:

$$f_\sigma^{(n)} > 0 \quad \text{implies} \quad (\exists k \in \mathbb{N}. \, n = k \cdot d).$$

MC D is called periodic with period d if $d > 1$, and aperiodic if $d = 1$.

Consider example MCs of just before. The first one is periodic, the second aperiodic.

Unique limiting distribution

An irreducible, aperiodic MC with finite state space Σ has a unique limiting distribution which is the solution of the system of equations $\rho = \rho \cdot \mathbf{P}$ with $\sum_{\sigma \in \Sigma} \rho_\sigma = 1$.

Second example of before.

The limiting distribution of reducible MC D can be obtained by considering the bottom strongly connected components $B_1, \ldots, B_k$ of D; determine the probability to eventually ends up in B_i and multiply this with the limiting distribution in B_i (if it exists).

Limiting distribution and mean recurrence times

Markov's theorem

A finite, irreducible MC is positive recurrent, and if aperiodic then:

$$\lim_{n \to \infty} \mathbf{P}^n = \begin{pmatrix} \rho \\ \cdot \\ \cdot \\ \rho \end{pmatrix} \quad \text{where} \quad \rho = \left(\frac{1}{m_{\sigma_1}}, \ldots, \frac{1}{m_{\sigma_k}} \right) \text{ for } k = |\Sigma|.$$

Recall: m_σ is the mean recurrence time of state σ.

Stationary distribution

Stationary distribution

A probability vector $\mathbf{x}$ satisfying $\mathbf{x} = \mathbf{x} \cdot \mathbf{P}$ is called a *stationary* distribution of MC D.

$$x_\sigma = \sum_{\tau \in \Sigma} x_\tau \cdot \mathbf{P}(\tau, \sigma) \quad \text{iff} \quad \underbrace{x_\sigma \cdot (1 - \mathbf{P}(\sigma, \sigma))}_{\text{outflow of } \sigma} = \underbrace{\sum_{\tau \neq \sigma} x_\tau \cdot \mathbf{P}(\tau, \sigma)}_{\text{inflow of } \sigma}$$

An irreducible, positive recurrent MC has a unique stationary distribution satisfying $x_\sigma = \frac{1}{m_\sigma}$ for every state σ.

Limiting = stationary distribution?

Limiting = stationary distribution

For ergodic (aka: aperiodic and positive recurrent) MCs, the stationary and limiting distribution coincide.

Relevance

Markov chains play a role in providing an operational interpretation to (discrete) probabilistic programs.

Limiting distributions play a role in understanding the termination behaviour of probabilistic programs.

Stationary distributions play a role in Markov chain Monte Carlo sampling.

Summary

- Transitions in MCs are probability distributions over next states
- Cylinders are the basis to define measures over sets of infinite paths
- Reachability probabilities in finite MCs are unique solutions of linear equation systems
- Transient distributions are n-step distributions $\mathbf{P}^n(\sigma, \cdot)$
- Limiting distribution are $\lim_{n \to \infty} \mathbf{P}^n(\sigma, \cdot)$
- Probability vector $\mathbf{x}$ with $\mathbf{x} = \mathbf{x} \cdot \mathbf{P}$ is a stationary distribution

Next lecture

Thursday Oct 20, 16:30