Stochastic processes

Markov Chains and Queuing Theory



Andrey Markov (1856–1922)

Photo: History of the USSR Academy of Sciences, 1964 (public domain)

Laurent Vanbever

nsg.ethz.ch

ETH Zürich

5 November 2026

Part 8

Last week we introduced

Continuous-Time Markov Chains

Last week we introduced

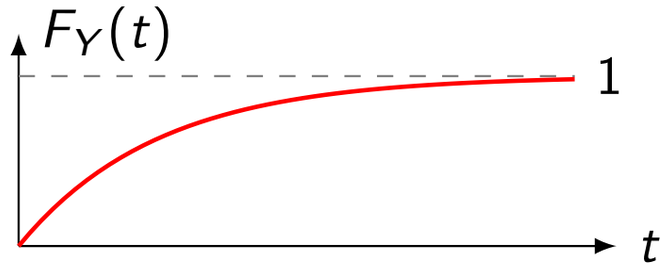
Continuous-Time Markov Chains

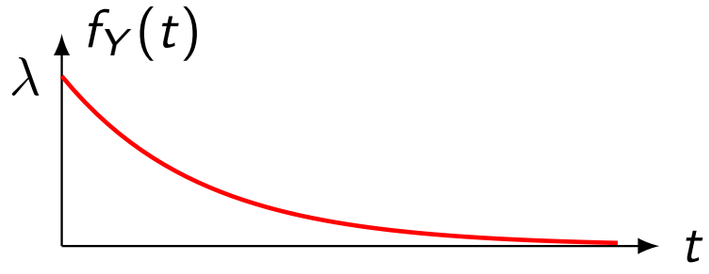
State changes happen at arbitrary times

the time spent in a state is exponentially distributed

The Exponential Distribution

- **Definition.** A random variable Y with the CDF $F_Y(t) = \Pr[Y \leq t] := 1 - e^{-\lambda t}$ for $t \geq 0$ (and 0 for $t < 0$) is **exponentially distributed** with parameter $\lambda > 0$, written $Y \sim \exp(\lambda)$.
- Its density is $f_Y(t) = \frac{\mathrm{d}}{\mathrm{d}t} F_Y(t) = \lambda e^{-\lambda t}$ for $t \geq 0$:





- λ is a **rate** (events per time unit), so $\mathbf{E}[Y] = \frac{1}{\lambda}$ and $\mathrm{Var}[Y] = \frac{1}{\lambda^2}$.

Leaving a State with Several Successors

- **Cashier problem:** two busy checkout desks, each serving at its own rate. How long until *one* of them is free?
- **Lemma 1.** Let $Y_1, \ldots, Y_k$ be independent with $Y_i \sim \exp(\lambda_i)$. Then $Y = \min\{Y_1, \ldots, Y_k\}$ is exponentially distributed with parameter $\lambda_1 + \cdots + \lambda_k$.
- **Proof** for $k = 2$ (the general case is the same):

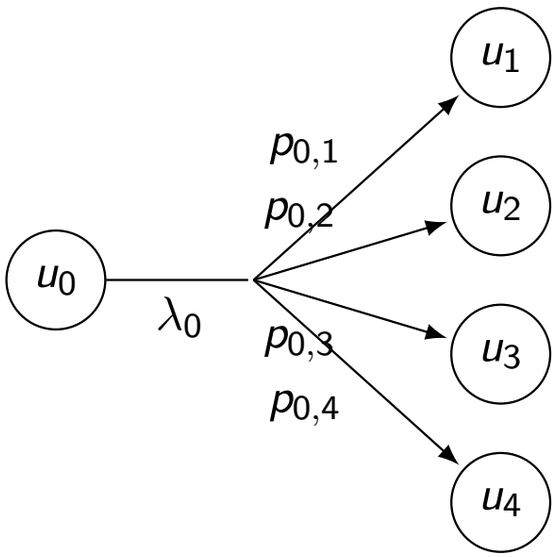
$$\begin{aligned} \Pr[Y > t] &= \Pr[Y_1 > t,\ Y_2 > t] = \Pr[Y_1 > t] \cdot \Pr[Y_2 > t] \\ &= e^{-\lambda_1 t} \cdot e^{-\lambda_2 t} = e^{-(\lambda_1+\lambda_2)t}. \end{aligned}$$ □

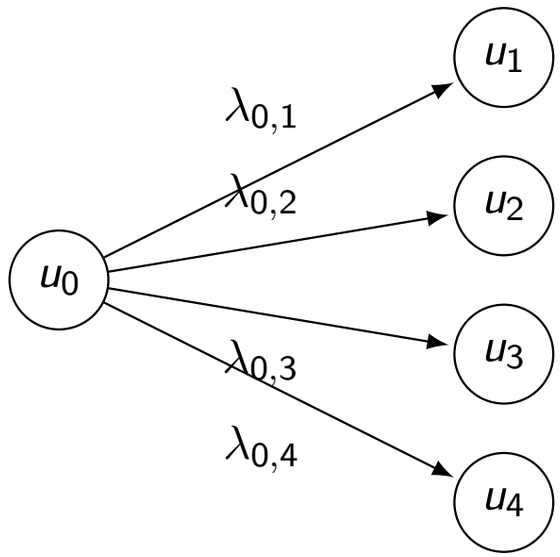
- **Lemma 2.** The probability that Y_1 is the smallest is (proof: see the script)

$$\Pr[Y_1 = \min\{Y_1, \ldots, Y_k\}] = \frac{\lambda_1}{\lambda_1 + \cdots + \lambda_k}.$$

- So a state with several successors is left at rate $\lambda = \sum_i \lambda_i$, and successor i wins with probability λ_i/λ.

Two Equivalent Views





- **Left:** stay in u_0 for an $\exp(\lambda_0)$ sojourn time, then pick the successor j with probability $p_{0,j}$.
- **Right:** every edge has its own independent clock $\exp(\lambda_{0,j})$ with $\lambda_{0,j} = p_{0,j} \cdot \lambda_0$; the first clock to ring wins.
- By the two lemmas, both describe *the same process*. In general: state i has sojourn rate λ_i, successor probabilities $p_{i,j}$ ($p_{i,i} = 0$, $\sum_j p_{i,j} = 1$) and **transition rates** $\lambda_{i,j} := \lambda_i \cdot p_{i,j}$, so $\sum_j \lambda_{i,j} = \lambda_i$.

Stationary Distribution

- In the long run the state probabilities stop changing: $\frac{\mathrm{d}}{\mathrm{d}t} q_i(t) = 0$ for all i.
- **Definition.** π is a **stationary distribution** if for all $i \in S$

$$0 = \sum_{j:j\neq i} \pi_j \cdot \lambda_{j,i} - \pi_i \cdot \lambda_i, \qquad \text{i.e.} \qquad \underbrace{\sum_{j:j\neq i} \pi_j \cdot \lambda_{j,i}}_{\text{flow into } i} = \underbrace{\pi_i \cdot \lambda_i}_{\text{flow out of } i}$$

with $\pi_i \geq 0$ and $\sum_i \pi_i = 1$.

- **What goes in must come out:** a linear system of equations, one per state, just like $\pi = \pi \cdot P$ in the discrete case.

M/M/1: Stationary Distribution (1)

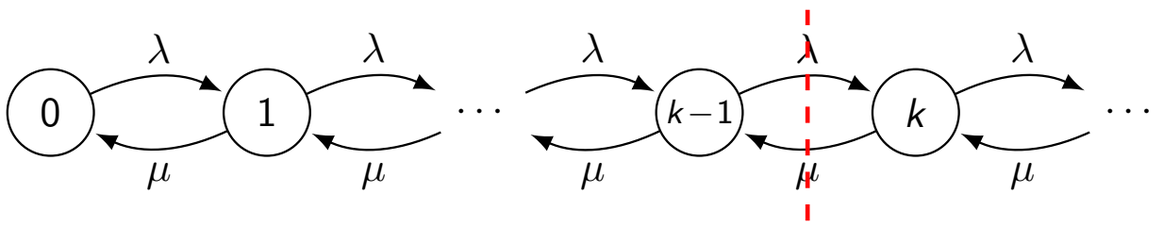
- **Theorem.** An M/M/1 queue has a stationary distribution if and only if

$$\rho := \frac{\lambda}{\mu} < 1. \qquad \text{In that case } \pi_k = (1-\rho)\,\rho^k.$$

- **Proof.** Flow in = flow out, for state 0 and for every state $k \geq 1$:

$$\mu \cdot \pi_1 = \lambda \cdot \pi_0, \qquad \lambda \cdot \pi_{k-1} + \mu \cdot \pi_{k+1} = (\lambda + \mu) \cdot \pi_k.$$

Simpler with **cuts**: left-to-right flow = right-to-left flow across the cut between $k-1$ and k:



$$\lambda \cdot \pi_{k-1} = \mu \cdot \pi_k \quad \Rightarrow \quad \pi_k = \rho \cdot \pi_{k-1} \quad \Rightarrow \quad \pi_k = \rho^k \cdot \pi_0.$$

Little's Law

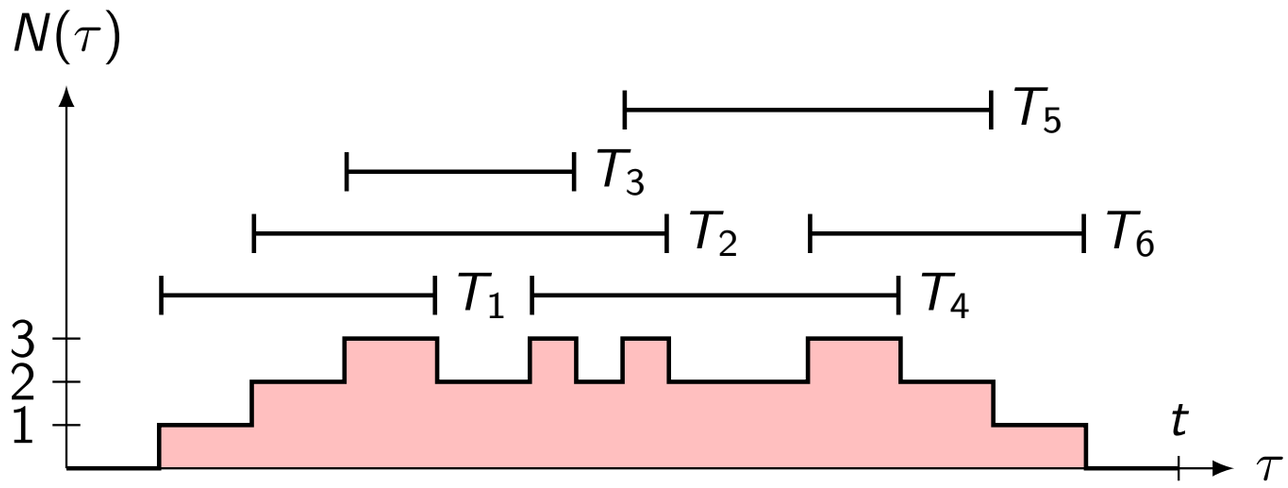
- **Definition.** Consider *any* queueing system and denote by
  - N the average number of jobs in the system,
  - λ the average arrival rate,
  - T the average **response time** of a job (waiting time + service time).

- **Theorem (Little's law, 1961).**

$$\boxed{N = \lambda \cdot T}$$

- Holds for *any* distribution of the inter-arrival and service times, *any* queueing discipline (not only FIFO), and *any* system or subsystem: an M/M/1 queue, a single server, the whole Internet.

Proof of Little's Law



- $N(\tau)$: jobs at time τ; $\alpha(t)$: arrivals until t; T_i: response time of job i. Assume $N(0) = N(t) = 0$ (simplifies, not needed).
- Measure the shaded area in **two ways**, along the curve and job by job (a strip of height 1 and length T_i per job): $\int_0^t N(\tau)\,\mathrm{d}\tau = \sum_{i=1}^{\alpha(t)} T_i$. Divide by t, expand by $\alpha(t)$:

$$\underbrace{\frac{1}{t}\int_0^t N(\tau)\,\mathrm{d}\tau}_{N} = \underbrace{\frac{\alpha(t)}{t}}_{\lambda} \cdot \underbrace{\frac{1}{\alpha(t)}\sum_{i=1}^{\alpha(t)} T_i}_{T}. \qquad \square$$

This week we finish queueing theory
and look at networks of queues

Today's plan

Thu Nov 5

1 Kendall's notation

2 Networks of queues