STATS 217: Introduction to Stochastic Processes I

Lecture 25

Last time: Stationary distributions

- Let $(X_t)_{t \geq 0}$ be a CTMC on Ω. A probability distribution π on Ω is said to be a stationary distribution if

$$\pi P^t = \pi \quad \forall t \geq 0$$

- This is equivalent to the condition that

$$\pi Q = 0.$$

- In terms of the matrix Q, the detailed balance conditions are given by

$$\pi_i q_{ij} = \pi_j q_{ji} \quad \forall i, j \in \Omega$$

Convergence theorem

Let $(X_t)_{t \geq 0}$ be an irreducible CTMC on a finite state space Ω. Then, there exists a unique stationary distribution π, and

$$\max_{x \in \Omega} \mathrm{TV}(P^t(x, \cdot), \pi) \to 0 \quad \text{as } t \to \infty.$$

Convergence theorem

Let $(X_t)_{t\geq 0}$ be an irreducible CTMC on a finite state space Ω. Then, there exists a unique stationary distribution π, and

$$\max_{x\in\Omega} \mathrm{TV}(P^t(x,\cdot),\pi) \to 0 \quad \text{as } t\to\infty.$$

(1)

(2)

We have already done the work to prove this theorem.

Existence of stationary distribution

- The first point is the existence of the stationary distribution.

Existence of stationary distribution

* assume that the chain $(X_t)_{t\geq 0}$ is described using Q.

- The first point is the existence of the stationary distribution.
- Recall the notation $\lambda_i = \sum_{j\neq i} \cancel{q_{ij}}$ q_{ij}, $\Lambda = \max_{i\in\Omega} \lambda_i$.
- Since Ω is finite, $\Lambda < \infty$.

In the case when $\Lambda < \infty$, we had a simpler way of simulating $(X_t)_{t\geq 0}$ given Q.

in this case, we were able to show that

$X_t = Y_{N(t)}$

$N(t)$ is a PPP w/ rate Λ

Y_n is a DTMC.

Existence of stationary distribution

- The first point is the existence of the stationary distribution.
- Recall the notation $\lambda_i = \sum_{j \neq i} q_j$, $\Lambda = \max_{i \in \Omega} \lambda_i$.
- Since Ω is finite, $\Lambda < \infty$. In this case, recall that we have the representation $X_t = Y_{N(t)}$, where $N(t)$ is a PPP with rate λ and Y_n is a DTMC with the transition matrix

$$U_{ij} = \frac{q_{ij}}{\Lambda} \quad \forall i \neq j \qquad\qquad U_{ii} = 1 - \frac{\lambda_i}{\Lambda}$$

Existence of stationary distribution

- The first point is the existence of the stationary distribution.
- Recall the notation $\lambda_i = \sum_{j \neq i} q_j$, $\Lambda = \max_{i \in \Omega} \lambda_i$.
- Since Ω is finite, $\Lambda < \infty$. In this case, recall that we have the representation $X_t = Y_{N(t)}$, where $N(t)$ is a PPP with rate λ and Y_n is a DTMC with the transition matrix

$$U_{ij} = \frac{q_{ij}}{\Lambda} \quad \forall i \neq j \qquad\qquad U_{ii} = 1 - \frac{\lambda_i}{\Lambda}$$

- Since $(X_t)_{t \geq 0}$ is irreducible, so is U, and hence, it has a unique stationary distribution π.

- goal: show that $\pi P_t = \pi \quad \forall \; t \geq 0$

$\iff \pi Q = 0$

Existence of stationary distribution

- We can check that $\pi Q = 0$.

Existence of stationary distribution

- We can check that $\pi Q = 0$. Indeed,

$$(\pi Q)_j = \sum_{i\in\Omega} \pi_i q_{ij} = \pi_j q_{jj} + \sum_{j\neq i} \pi_i q_{ij}$$

Existence of stationary distribution

$U_{ij} = \frac{q_{ij}}{\Lambda} \quad \forall\, i \neq j$

- We can check that $\pi Q = 0$. Indeed,

$$\sum_{i \in \Omega} \pi_i q_{ij} = \pi_j q_{jj} + \sum_{j \neq i} \pi_i q_{ij}$$

$$= -\pi_j \lambda_j + \sum_{i \neq j} \pi_i U_{ij} \Lambda$$

$U_{jj} = 1 - \frac{\lambda_j}{\Lambda}$

$\lambda_j = -q_{jj}$

$\underbrace{\lambda_j = \sum_{k \neq j} q_{jk}}$

Existence of stationary distribution

$\pi U = \pi$

- We can check that $\pi Q = 0$. Indeed,

$$\begin{aligned}
\sum_{i \in \Omega} \pi_i q_{ij} &= \pi_j q_{jj} + \sum_{j \neq i} \pi_i q_{ij} \\
&= -\pi_j \lambda_j + \sum_{i \neq j} \pi_i U_{ij} \Lambda \\
&= -\pi_j \lambda_j + \Lambda \sum_{i \in \Omega} \pi_i U_{ij} - \Lambda \pi_j U_{jj} \\
&= -\pi_j \lambda_j + \Lambda \pi_j - \pi_j(\Lambda - \lambda_j) \\
&= 0.
\end{aligned}$$

o wts : if $\pi Q = 0 = \mu Q \Rightarrow \pi = \mu$.

$\pi P = \pi$ & $\mu P = \mu$. since P is irred. $\Rightarrow$ $\pi = \mu$.

Convergence

wts: $\max_x TV(P^t(x,\cdot),\pi) \xrightarrow{t\to\infty} 0$

we know: $\tilde{P}$ is an irr + aperiodic transition matrix on finite Ω, then $\max_x TV(\tilde{P}^n(x,\cdot),\pi) \xrightarrow{n\to\infty} 0$

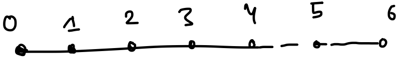
- Note that

$$\mathsf{TV}(P^{t+s}(x,\cdot),\pi) = \mathsf{TV}(\delta_x P^t P^s, \pi P^s)$$

Y_n w/ transition matrix $\mathbb{P} (= P^1)$

we know that P^1 is irr + aperiodic (Levy's dichotomy)

$\Rightarrow \max_x TV(\mathbb{P}^n(x,\cdot),\pi) \xrightarrow{n\to\infty} 0$



Recall: that for DTMC: we found $\exists$ some $r_0 > 0$ s.t.

every entry of $P^{r_0} > 0$.

then we showed

$$\max_{x} TV(P^{r_0 k}(x,\cdot), \pi) \xrightarrow{k \to \infty} 0$$

- Note that

$$P^{t+s} = P^t P^s$$

$$\begin{aligned} TV(P^{t+s}(x,\cdot),\pi) &= TV(\delta_x P^t P^s, \pi P^s) \\ &\le TV(\delta_x P^t, \pi). \end{aligned}$$

prob. distribution w/

$$\text{prob}[y] = P^{t+s}(x,y)$$

$$P^{t+s}(x,\cdot) = \delta_x P^{t+s} = \delta_x P^t P^s$$

δ_x = constant distribution at x

i.e. $|\Omega|$-dimensional row vector which is 1 at x, 0 elsewhere

Convergence

- Note that

$$\begin{aligned}\mathrm{TV}(P^{t+s}(x,\cdot),\pi) &= \mathrm{TV}(\delta_x P^t P^s, \pi P^s)\\ &\leq TV(\delta_x P^t, \pi).\end{aligned}$$

- Hence, $\mathrm{TV}(P^t(x,\cdot),\pi)$ is non-increasing in t, so it suffices to show that it converges to 0 along (say) the natural numbers.

Convergence

① first gen. joint sample from the opt. coupling of

$(\hat{X}, \hat{Y}) = (\delta_x P^t, \pi)$

② $(\hat{X}P^s, \hat{Y}P^s) \sim (\delta_x P^t P^s, \pi)$

- Note that

$$\begin{aligned} TV(P^{t+s}(x,\cdot),\pi) &= TV(\delta_x P^t P^s, \pi P^s) \\ &\le TV(\delta_x P^t, \pi). \end{aligned}$$

③ $\mathbb{P}[\hat{X}P^s \neq \hat{Y}P^s] \le \mathbb{P}[\hat{X} \neq \hat{Y}]$

- Hence, $TV(P^t(x,\cdot),\pi)$ is non-increasing in t, so it suffices to show that it converges to 0 along (say) the natural numbers.
- But P^1 is an irreducible and aperiodic transition matrix with unique stationary distribution π, so that by looking at the corresponding DTMC, we have

$$TV(P^n(x,\cdot),\pi) \to 0 \quad \text{as } n \to \infty.$$

Example: M/M/1 queues # of servers

- This is a popular queuing model in which the arrival of customers is modelled by a Poisson point process with rate λ. There is a single server, and service times are independent and exponentially distributed with parameter μ.

Example: M/M/1 queues

- This is a popular queuing model in which the arrival of customers is modelled by a Poisson point process with rate λ. There is a single server, and service times are independent and exponentially distributed with parameter μ.
- Due to the memorylessness property of the exponential distribution, this can be modelled as a continuous time birth and death chain with jump rates

think of n as the # of customers in the system.

$$Q_{n,n+1} = \lambda, \quad n = 0, 1, \ldots$$
$$Q_{n,n-1} = \mu, \quad n = 1, 2, \ldots$$

{continuous time birth & death chain.}

Example: M/M/1 queues

- This is a popular queuing model in which the arrival of customers is modelled by a Poisson point process with rate λ. There is a single server, and service times are independent and exponentially distributed with parameter μ.
- Due to the memorylessness property of the exponential distribution, this can be modelled as a continuous time birth and death chain with jump rates

$$\begin{aligned} Q_{n,n+1} &= \lambda, \quad n = 0, 1, \dots \\ Q_{n,n-1} &= \mu, \quad n = 1, 2, \dots \end{aligned}$$

- Suppose instead that there are s servers, and customers are served if there is at least one server available. This is called the $M/M/s$ queueing model,

Example: M/M/1 queues

- This is a popular queuing model in which the arrival of customers is modelled by a Poisson point process with rate λ. There is a single server, and service times are independent and exponentially distributed with parameter μ.
- Due to the memorylessness property of the exponential distribution, this can be modelled as a continuous time birth and death chain with jump rates

$$Q_{n,n+1} = \lambda, \quad n = 0, 1, \ldots$$
$$Q_{n,n-1} = \mu, \quad n = 1, 2, \ldots$$

- Suppose instead that there are s servers, and customers are served if there is at least one server available. This is called the $M/M/s$ queueing model, and the jump rates are now

$$\begin{aligned} Q_{n,n+1} &= \lambda, \quad n = 0, 1, \ldots, \\ Q_{n,n-1} &= n\mu, \quad n = 1, \ldots, s, \\ Q_{n,n-1} &= s\mu, \quad n = s+1, s+2, \ldots, \end{aligned}$$

M/G/s

GI/G/S

M/M/∞

M/M/1 queues

arrival rate

service rate

- Suppose that $\lambda < \mu$ i.e., the rate of arrivals is smaller than the rate of service. Otherwise, the size of the queue explodes.

M/M/1 queues

- Suppose that $\lambda < \mu$ i.e., the rate of arrivals is smaller than the rate of service. Otherwise, the size of the queue explodes.
- When $\lambda < \mu$, we can use the detailed balance conditions

$$\pi_i Q_{ij} = \pi_j Q_{ji}$$

M/M/1 queues

- Suppose that $\lambda < \mu$ i.e., the rate of arrivals is smaller than the rate of service. Otherwise, the size of the queue explodes.
- When $\lambda < \mu$, we can use the detailed balance conditions

$$\pi_i Q_{ij} = \pi_j Q_{ji}$$

to find the stationary distribution

$$\pi_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \quad n = 0, 1, \ldots$$

Q 1: server occupancy

: server occupied $\Leftrightarrow n \geq 1$

$= 1 - \pi_0$

M/M/1 queues

- Suppose that $\lambda < \mu$ i.e., the rate of arrivals is smaller than the rate of service. Otherwise, the size of the queue explodes.
- When $\lambda < \mu$, we can use the detailed balance conditions

$$\pi_i Q_{ij} = \pi_j Q_{ji}$$

  to find the stationary distribution

$$\pi_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \quad n = 0, 1, \ldots$$

- Given this stationary distribution, one can compute many quantities of interest.

M/M/1 queues

- Suppose that $\lambda < \mu$ i.e., the rate of arrivals is smaller than the rate of service. Otherwise, the size of the queue explodes.
- When $\lambda < \mu$, we can use the detailed balance conditions

$$\pi_i Q_{ij} = \pi_j Q_{ji}$$

  to find the stationary distribution

$$\pi_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \quad n = 0, 1, \ldots$$

- Given this stationary distribution, one can compute many quantities of interest. For instance, the long-run fraction of time that the server is busy is

$$1 - \pi_0 = \frac{\lambda}{\mu}.$$

M/M/1 queues

(1) waiting time / time spent in the system.

- Moreover, the expected length of the queue under the equilibrium distribution is

$$L = \sum_{n=0}^{\infty} n\pi_n = \frac{\lambda}{\mu - \lambda}.$$

$\lambda = 10$

$\mu = 15$

$$\frac{10}{15-10} = 2$$

M/M/1 queues

- Moreover, the expected length of the queue under the equilibrium distribution is
$$L = \sum_{n=0}^{\infty} n\pi_n = \frac{\lambda}{\mu - \lambda}.$$
- Another important quantity is the total time T (waiting time + time with the server) spent by a customer in the system.

M/M/1 queues

- Moreover, the expected length of the queue under the equilibrium distribution is
$$L = \sum_{n=0}^{\infty} n\pi_n = \frac{\lambda}{\mu - \lambda}.$$
- Another important quantity is the total time T (waiting time + time with the server) spent by a customer in the system.
- If there are n customers already in the system when a new customer joins the queue,

M/M/1 queues

- Moreover, the expected length of the queue under the equilibrium distribution is

$$L = \sum_{n=0}^{\infty} n\pi_n = \frac{\lambda}{\mu - \lambda}.$$

- Another important quantity is the total time T (waiting time + time with the server) spent by a customer in the system.
- If there are n customers already in the system when a new customer joins the queue, then since service times are i.i.d. exponentials with parameter μ, the total time spent by the customer is distributed as a sum of $n+1$ i.i.d. exponentials with parameter μ.

M/M/1 queues

- Then, using the law of total probability, we have

$$\mathbb{P}[T \leq t] = \mathbb{P}[T \leq t \mid n \text{ customers already in the system}] \cdot \pi_n$$

M/M/1 queues

- Then, using the law of total probability, we have

$$\begin{aligned}\mathbb{P}[T \leq t] &= \mathbb{P}[T \leq t \mid n \text{ customers already in the system}] \cdot \pi_n \\ &= 1 - \exp(-t(\mu - \lambda)),\end{aligned}$$

M/M/1 queues

- Then, using the law of total probability, we have

$$\begin{aligned}\mathbb{P}[T \le t] &= \mathbb{P}[T \le t \mid n \text{ customers already in the system}] \cdot \pi_n \\ &= 1 - \exp(-t(\mu - \lambda)),\end{aligned}$$

$\lambda = 10$
$\mu = 15$

i.e. T has exponential distribution with mean

$$W = \frac{1}{\mu - \lambda}$$

$W \sim \text{Exp}(5)$.
w. mean
$\frac{1}{15-10} = \frac{1}{5}$

μ, λ

$W, \quad L = \frac{\lambda}{\mu - \lambda}$
$\frac{1}{\mu - \lambda}$

$\boxed{L = \lambda W}$

M/M/1 queues

- Then, using the law of total probability, we have

$$\begin{aligned}\mathbb{P}[T \le t] &= \mathbb{P}[T \le t \mid n \text{ customers already in the system}] \cdot \pi_n \\ &= 1 - \exp(-t(\mu - \lambda)),\end{aligned}$$

i.e. T has exponential distribution with mean

$$W = \frac{1}{\mu - \lambda} = \frac{L}{\lambda}.$$

Little's law

- The relationship

$$L = \lambda W$$

  is called **Little's law**

Little's law

- The relationship
$$L = \lambda W$$
is called **Little's law** and is true even without the specific distributional assumptions (i.e. Poisson arrivals and exponential waiting times).

Little's law

- The relationship
$$L = \lambda W$$
is called **Little's law** and is true even without the specific distributional assumptions (i.e. Poisson arrivals and exponential waiting times). Such queues are called $GI/G/1$ queues.

Little's law

- The relationship
$$L = \lambda W$$
is called **Little's law** and is true even without the specific distributional assumptions (i.e. Poisson arrivals and exponential waiting times). Such queues are called $GI/G/1$ queues.
- Here's the intuition: Suppose each customer pays \$1 for each minute of time they spend in the system. When there are n customers in the system, the establishment is earning \$$n$ per minute, and hence, the establishment is earning an average of \$$L$ per minute.

Little's law

- The relationship
$$L = \lambda W$$
is called **Little's law** and is true even without the specific distributional assumptions (i.e. Poisson arrivals and exponential waiting times). Such queues are called $GI/G/1$ queues.
- Here's the intuition: Suppose each customer pays \$1 for each minute of time they spend in the system. When there are n customers in the system, the establishment is earning \$$n$ per minute, and hence, the establishment is earning an average of \$$L$ per minute.
- On the other hand, if each customer pays for their entire duration when they arrive, then the average rate of earning is $\lambda \times W$ per minute.