Lecture Notes for "Stochastic Modeling and Computations"

M. Chertkov (lecturer), S. Belan and V. Parfeneyv (recitation instructors)
M.Sc. and Ph.D. level course at Skoltech
Moscow, March 28 - May 28, 2016
`https://sites.google.com/site/mchertkov/courses`

The course offers a soft and self-contained introduction to modern applied probability covering theory and application of stochastic models. Emphasis is placed on intuitive explanations of the theoretical concepts, such as random walks, law of large numbers, Markov processes, reversibility, sampling, etc., supplemented by practical/computational implementations of basic algorithms. In the second part of the course, the focus shifts from general concepts and algorithms per se to their applications in science and engineering with examples, aiming to illustrate the models and make the methods of solution, originating from physics, chemistry, machine learning, control and operations research, clear.

I. THEME # 2. STOCHASTIC PROCESSES

A. Lecture #4: Markov Chains [discrete space, discrete time].

1. Transition Probabilities

So far we have studied random variables and events often assuming that these are i.i.d. = independent identically distributed. However, in real world we "jump" from one random state to another so that the transition depends on the original state. We may have a memory which last more than one jump, however there is also a big family of interesting random processes which do not have long memory - only current state influences where we jump to. This is the class of random processes described by Markov Chains (MCs).

MCs can be explained in terms of directed graphs, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of vertices, $\mathcal{V} = (i)$, is asociated with the set of states, and the set of directed edges, $\mathcal{E} = (j \leftarrow i)$, correspond to possible transitions between the states. Note that we may also have "self-loops", $(i \leftarrow i)$ included in the set of edges. To make description complete we need to associate to each vertex a transition probability, $p_{j\leftarrow i} = p_{ji}$ from the state i to the state j. Since p_{ji} is the probability, $\forall (j \leftarrow i) \in \mathcal{E}: \quad p_{ji} \geq 0$, and

$$\forall i: \quad \sum_{j:(j\leftarrow i)\in\mathcal{E}} p_{ji} = 1. \tag{I.1}$$

Then, the combination of $\mathcal{G}$ and $p \doteq (p_{ji}|(j \leftarrow i) \in \mathcal{E})$ defines a MC. Mathematically we also say that the tuple (finite ordered set of elements), $(\mathcal{V}, \mathcal{E}, p)$, defines the Markov chain. We will mainly consider in the following stationary Markov chains, i.e. these with p_{ji} constant - not changing in time. However, for many of the following statements/considerations generalization to the time-dependent processes is straightforward.

MC generates a random (stochastic) dynamic process. Time flows continuously, however as a matter of convenient abstraction we consider discrete times (and sometimes, actually quite often, events do happen discreetly). One uses $t = 0, 1, 2, \cdots$ for the times when jumps occur. Then a particular random trajectory/path/sample of the system will look like

$$i_1(0), i_2(1), \cdots, i_k(t_k), \quad \text{where} \quad i_1, \cdots, i_k \in \mathcal{V}$$

We can also generate many samples (many trajectories)

$$n = 1, \cdots, N: \quad i_1^{(n)}(0), i_2^{(n)}(1), \cdots, i_k^{(n)}(t_k), \quad \text{where} \quad i_1, \cdots, i_k \in \mathcal{V}$$

where N is the number of trajectories.

How does one relates the directed graph with weights (associated to the transition probabilities) to samples? The relation, actually, has two sides. The direct one - is about how one generates the samples. The samples are generated by advancing the trajectory from the current-time state flipping coin according to the transition probability p_{ij}. The inverse one - is about reconstructing characteristics of Markov chain from samples (or may be, even on the first place, verifying if the samples where indeed generated according to (rather restrictive) MC rules.

Now let us get back to the direct problem where a MC is described in terms of $(\mathcal{V}, \mathcal{E}, p)$. However, instead of characterizing the system in terms of the trajectories/paths/samples, we can pose the question following evolution of the "state probability vector", or simply the "state vector":

$$\forall i \in \mathcal{V}, \quad \forall t = 0, \cdots: \quad \pi_i(t+1) = \sum_{j:(j\leftarrow i)\in\mathcal{E}} p_{ij}\pi_j(t). \tag{I.2}$$

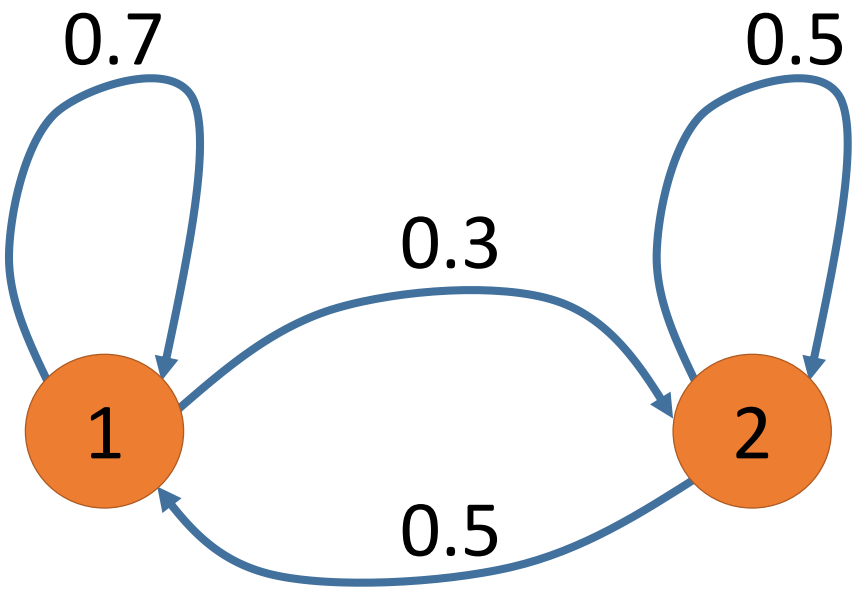

FIG. 1: Markov Chain (MC) - Example #1.

Here, $\pi(t) \doteq (\pi_i(t) \geq 0 | i \in \mathcal{V})$ is the vector built of components each representing probability for the system to be in the state i at the moment of time t. Thus, $\sum_{i\in\mathcal{V}} \pi_i = 1$. We can also rewrite Eq. (I.2) in the vector/matrix form

$$\pi(t+1) = p\pi(t), \tag{I.3}$$

where $\pi(t)$ the column/state and $p(t)$ is the transition-probability matrix, which satisfies the so-called "stochasticity" property (I.1). Sequential application of Eq. (I.3) results in

$$\pi(t+k) = p^k\pi(t), \tag{I.4}$$

and we are interested to analyze properties of p^k, characterizing the Markov chain acting for k sequential periods.

Let us first study it on the example of the simple MC shown in Fig. (1). In this case p^k is 2×2 which dependence on k is as follows

$$p^1 = \begin{pmatrix} 0.7 & 0.5 \\ 0.3 & 0.5 \end{pmatrix}, \quad p^2 = \begin{pmatrix} 0.64 & 0.6 \\ 0.36 & 0.4 \end{pmatrix}, \quad p^{10} \approx p^{100} \approx \begin{pmatrix} 0.625 & 0.625 \\ 0.375 & 0.375 \end{pmatrix}. \tag{I.5}$$

2. *Properties of Markov Chains*

The MC is **irreducible** if one can access any state from any state, formally

$$\forall i, j \in \mathcal{V}: \quad \exists k > 1, \quad \text{s.t.} \quad (p^k)_{ij} > 0. \tag{I.6}$$

Example #1 is obviously irreducible. However, if we replace $0.3 \to 0$ and $0.7 \to 1$ the MC becomes reducible – state 1 is not accessible from 2.

A state i has period k if any return to the state must occur in multiples of k. If $k = 1$ than the state is **aperiodic**. MC is **aperiodic** if all states are aperiodic. An irreducable MC only needs one aperiodic state to imply all states are aperiodic. Any MC with at least one self-loop is aperiodic. Example #1 is obviously aperiodic. However, it becomes periodic with period two if the two self-loops are removed.

A state i is said to be transient if, given that we start in state i, there is a non-zero probability that we will never return to i. State i is recurrent if it is not transient. State i is **positive-recurrent** if the expected return time (to the state) is positive (this feature is important for infinite graphs).

A state is **ergodic** if the state is aperiodic and positive-recurrent. If all states in an irreducible MC are ergodic than the MC is said to be ergodic. A MC is ergodic if there is a finite number k_* such that any state can be reached from any other state in exactly k_* steps. For the example #1 $k_* = 2$. Note, that there are other (alternative) descriptions of ergodicity. Thus most intuitive one is: the MC is ergodic if it is irreducable and aperiodic. In this course we will not dwell much on the rich mathematical formalities and details, largely considering generic ergodic MC.

3. Steady State Analysis

Component-wise positive, normalized, π^*, is called stationary distribution (invariant measure) if

$$\pi^* = p\pi^* \tag{I.7}$$

An irreducible MC has a stationary distribution iff all of its states are positive recurrent. Solving Eq. (I.7) for the example # one finds

$$\pi^* = \begin{pmatrix} 0.625 \\ 0.375 \end{pmatrix}, \tag{I.8}$$

which is naturally consistent with Eq. (I.5). In general,

$$\pi^* = \frac{e}{\sum_i e_i}, \tag{I.9}$$

where e is the eigenvector with the eigenvalue 1. And how about other eigenvalues of the transition matrix?

4. Spectrum of the Transition Matrix & Speed of Convergence to the Stationary Distribution

Assume that p is diagonalizable (has $n = |p|$ linearly independent eigenvectors) then we can decompose p according to the following eigen-decomposition

$$p = U^{-1}\Sigma U \tag{I.10}$$

where $\Sigma = \text{diag}(\lambda_1, \cdots, \lambda_n)$, $1 = |\lambda_1| > \lambda_2 \geq |\lambda_3| \geq \cdots |\lambda_n|$ and U is the matrix of eigenvectors (each normalized to having an l_2 norm equal to 1) where each raw is a right eigenvector of p. Then

$$\pi^{(k)} = p^k\pi = (U^{-1}\Sigma U)^k\pi_0 = U^{-1}\Sigma^k U\pi_0. \tag{I.11}$$

Let us represent p_0 as an expansion over the normalized eigenvectors, $u_i, \cdots i = 1, \cdots, n$:

$$\pi = \sum_{i=1}^{n} a_i u_i. \tag{I.12}$$

Taking into account orthonormality of the eigenvectors one derives

$$\pi^{(k)} = \lambda_1 \left(a_1 u_1 + a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k u_2 + \cdots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^k u_n \right) \tag{I.13}$$

Since $\pi^{(k)}_{k\to\infty} \to \pi^* = u_1$, the second term on the rhs of Eq. (I.13) describes the rate of convergence of $\pi^{(k)}$ to the steady state – exponential in $\log(\lambda_1/\lambda_2$.

5. Reversible & Irreversible Markov Chains.

MC is called **reversible** if there exists π s.t.

$$\forall\{i,j\} \in \mathcal{E}: \quad p_{ji}\pi^*_i = p_{ij}\pi^*_j, \tag{I.14}$$

where $\{i,j\}$ is our notation for the undirected edge, assuming that both directed edges $(i \leftarrow j)$ and $(j \leftarrow i)$ are elements of the set $\mathcal{E}$. In physics this property is also called **Detailed Balance** (DB). If one introduces the so-called ergodicity matrix

$$Q \doteq (Q_{ji} = p_{ji}\pi_i^* | (j \leftarrow i) \in \mathcal{E}), \tag{I.15}$$

then DB translates into the statement that Q is symmetric, $Q = Q^T$, that is reversible. The MC for which the property does not hold is called **irreversible**. $Q - Q^T$ is nonzero, i.e. Q is asymmetric for reversible MC. An asymmetric component of Q is the matrix build from currents/flows (of probability). Thus for the case #1 shown in Fig. (1)

$$Q = \begin{pmatrix} 0.7 * 0.625 & 0.5 * 0.375 \\ 0.3 * 0.625 & 0.5 * 0.375 \end{pmatrix} = \begin{pmatrix} 0.4375 & 0.1875 \\ 0.1875 & 0.1875 \end{pmatrix} \tag{I.16}$$

Q is symmetric, i.e. even though $p_{12} \neq p_{21}$, there is still no flow of probability from 1 to 2 as the "population" of the two states, π_1^* and π_2^* respectively are different, so that in the result $Q_{12} - Q_{21} = 0$. In fact, one can show that in the two node situation steady state of a MC is always DB.

6. *Detailed Balance vs Global Balance. Adding cycles to accelerate mixing.*

Note that if a steady distribution, π^*, satisfy the DB condition (I.14) for a MC, $(\mathcal{V}, \mathcal{E}, p)$, it will also be a steady state of another MC, $(\mathcal{V}, \mathcal{E}, \tilde{p})$, satisfying the more general Balance (or global balance) condition

$$\sum_{j:(j\leftarrow i)\in\mathcal{E}} \tilde{p}_{ji}\pi_i^* = \sum_{j:(i\leftarrow j)\in\mathcal{E}} \tilde{p}_{ij}\pi_j^*. \tag{I.17}$$

This suggests that many different MC (many different dynamics) may result in the same steady state. Obviously DB is a particular case of the B-condition, but the opposite is not true.

The difference between DB- and B- can be nicely interpreted in terms of flows (think water) in the state space. From the hydrodynamic point of view reversible MCMC corresponds to an irrotational probability flows, while irreversibility relates to nonzero rotational part, e.g. correspondent to vortices contained in the flow. Putting it formally, in the irreversible case antisymmetric part of the ergodic flow matrix, $Q = (\tilde{p}_{ij}\pi_j^* | (i \leftarrow j))$, is nonzero and it actually allows the following cycle decomposition,

$$Q_{ij} - Q_{ji} = \sum_\alpha J_\alpha \left(C_{ij}^\alpha - C_{ji}^\alpha\right) \tag{I.18}$$

where index enumerates cycles on the graph of states with the adjacency matrices C^α. Then, J_α stands for the magnitude of the probability flux flowing over cycle .

One can use the cycle decomposition to modify MC such that the steady distribution stay the same (invariant). Of course, cycles should be added with care, e.g. to make sure that all the transition probabilities in the resulting $\tilde{p}$, are positive (stochasticity of the matrix will be guaranteed by construction). "Adding cycles" along with some additional tricks (e.g. lifting/repkication MC) may help to improve mixing, i.e. speed up convergence to the steady state — which is a very desirable property for sampling π^* efficiently. This and other features of MC will be discussed in details at the recitations on a three node example.

7. *Recitation. Markov Chains: Detailed Balance. Mixing time.*

B. Lecture #5. From Bernoulli Processes to Poisson Processes [discrete space, discrete & continuous time].

The two processes discussed here are the simplest dynamic random processes. Simplicity here is related to the fact that the processes are defined with the least number of characteristics. We will also focus on important properties of the processes, e.g. memorylessness, and also on working out interesting (and rather general) questions one may ask (and answer).

1. *Bernoulli Process: Definition*

Defined as a sequence of independent Bernoulli trials. At each trial

- P(success)=P(x=1)=p

- P(failure)=P(x=0)=1-p

Can be represented as a simple MC (two nodes + two self-loops). The sequence looks like $00101010001 = **S*S*S***S$. S here stands for "success".

Examples:

- Sequence of discrete updates – ups and downs (stock market)
- sequence of lottery wins
- arrivals of busses at a station checked every 1/5/? minutes

2. *Bernoulli: Number of Successes*

Number of k successes in n steps follows the binomial distribution

$$\forall k = 0, \cdots, n: \quad P(S = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \tag{I.19}$$

$$\text{mean}: \quad \mathbb{E}[S] = np \tag{I.20}$$

$$\text{variance}: \quad var(S) = \mathbb{E}[(S - \mathbb{E}[S])^2] = np(1-p) \tag{I.21}$$

3. *Bernoulli: Distribution of Inter-Arrivals*

Call T_1 the number of trials till the first success (including the success event too). The Probability Mass Function (PMF)

$$t = 1, 2, \cdots: \quad P(T_1 = t) = p(1-p)^{t-1} [\text{Geometric PMF}] \tag{I.22}$$

The answer is product (thus memoreless) of the probabilities of $(t-1)$ failures and one success.

$$\text{mean}: \quad \mathbb{E}[T_1] = \frac{1}{p} \tag{I.23}$$

$$\text{variance}: \quad var(T_1) = \mathbb{E}[(T_1 - \mathbb{E}[T_1])^2] = \frac{1-p}{p^2} \tag{I.24}$$

More on the memoryless property. Given n, the future sequence $x_{n+1}, x_{n+2}, \cdots$ is also a Bernoulli process and is independent of the past. Moreover, suppose we observed the process for n times and no success has occurred. Then the PMF for the remaining arrival times is also geometric

$$P(T - n = k | T > n) = p(1-p)^{k-1} \tag{I.25}$$

And how about the k^{th} arrival? Let, y_k is the number of trials until k^{th} success (inclusive). T_k (previously introduced) is $T_k = Y_k - Y_{k-1}, \quad k = 2, 3, \cdots$ for k^{th} interarrival time. Also, $y_k = T_1 + T_2 + \cdots + T_k$. Then,

$$t = k, k+1, \cdots: \quad P(y_k = t) = \left(\begin{array}{c} t-1 \\ k-1 \end{array} \right) p^k (1-p)^{t-k} [\text{Pascal PMF}] \tag{I.26}$$

$$\text{mean}: \quad \mathbb{E}[y_k] = \frac{k(1-p)}{p^2} \tag{I.27}$$

$$\text{variance}: \quad var(y_k) = \mathbb{E}[(y_k - \mathbb{E}[y_k])^2] = \frac{k(1-p)}{p^2} \tag{I.28}$$

4. *Poisson Process: Definition*

Examples:

- all examples from the Bernoulli case in continuous time
- e-mails arrivals with infrequent check

- high-energy beams collide at a high frequency (10 MHz) with a small chance of good event
- radioactive decay of a nucleus with the trial being to observe a decay within a small interval
- spin flip in a magnetic field

Two way of thinking of it. One as of a continuous version of the Bernoulli process. Another through random time intervals between successes.

Start from the first one. Intervals become infinitesimally small and we replace probabilities (of success) by probability densities (per unit time). Let $P(k,\tau)$ be the probability of k arrivals in an interval of duration τ. We assume that

- number of arrivals in disjoint time intervals are independent
- for very small δ (regularization)

$$P(k,\delta) \approx \begin{cases} 1-\lambda\delta & k=0 \\ \lambda\delta & k=1 \\ 0 & k>0 \end{cases} \tag{I.29}$$

- λ is the arrival rate of the process

on the relations between Bernoulli and Poisson (assuming $n = t/\delta$ and $p = \lambda\delta$):

Bernoulli	Possion
arrival probability in each time slot $= p$	arrival probability in each δ-interval $= \lambda\delta$
number of arrivals in t intervals	number of successes in n time slots

5. *Poisson: Inter-arrival Time*

Probability density of the first arrival, y_1:

$$p_{Y_1}(y) = \lambda \exp(-\lambda y), \quad y \geq 0 \quad \text{[exponential]}$$

Then

$$P * Y_1 \leq y) = 1 - P(0,y) = 1 - \int_0^y dy' p_{Y_1}(y') = 1 - \exp(-\lambda y)$$

Like Bernoulli, the Poisson keeps the two key properties

- **Fresh Start Property**: the time of the next arrival is independent from the past
- **Memoryless property:** suppose we observe the process for t seconds and no success occurred. Then the density of the remaining time of arrival is exponential.

By extension (taking limit), for the probability density of time of the k^{th} arrival one derives

$$p_{Y_k}(y) = \frac{\lambda^k y^{k-1} \exp(-\lambda y)}{(k-1)!}, \quad y > 0 \quad \text{(Erlang "of order" k)} \tag{I.30}$$

To conclude

	Bernoulli	Poisson
Times of Arrival	Discrete	Continuous
Arrival Rate	p/per trail	λ/unit time
PMF of Number of arrivals	Binomial	Poisson
PMF of Interarrival Time	Geometric	Exponential
PMF of k^{th} Arrival Time	Pascal	Erlang

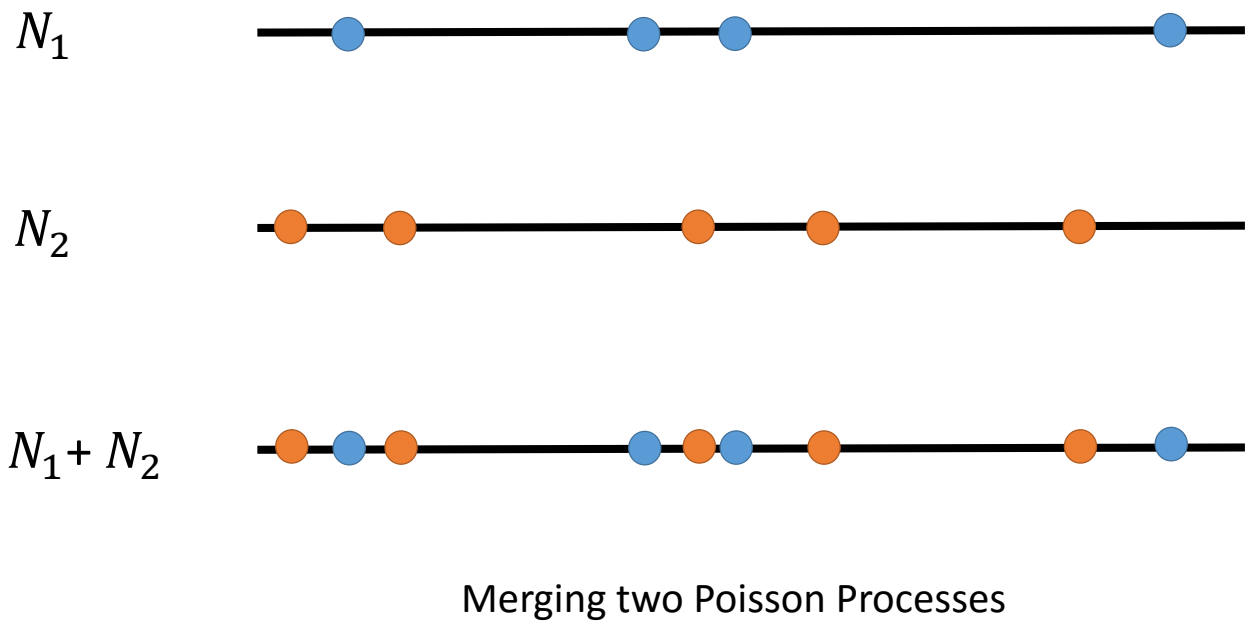

FIG. 2: Merging two Poisson processes.

6. *Poisson: Number of arrivals in a t-intervals as $n \to \infty$*

$$\binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k (1 - \frac{\lambda t}{n})^{n-k} \quad \text{[Binomial]} \tag{I.31}$$

$$= \underbrace{\frac{n!}{(n-k)!n^k}}_{\to 1} \frac{(\lambda t)^k}{k!} \underbrace{\left(1 - \frac{\lambda t}{n}\right)^n}_{\to \exp(-\lambda t)} \underbrace{\left(1 - \frac{\lambda t}{n}\right)^{-k}}_{\to 1} \quad \text{(reoder terms)} \tag{I.32}$$

$$= P(N = k) = P(k = \tau) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad \text{[Poisson}(\lambda\tau)\text{]} \tag{I.33}$$

$$\text{mean :} \quad \mathbb{E}[N] = \lambda t \tag{I.34}$$

$$\text{variance :} \quad var(N) = \mathbb{E}[(N - \mathbb{E}[N])^2] = \lambda\tau \tag{I.35}$$

7. *Merging and Splitting Processes*

Most important feature shared by Bernoulli and Poisson processes is their invariance with respect to mixing and splitting. We will show it on an example of Poisson process but the same applies to Bernoulli process.

Merging: Let $N_1(t)$ and $N_2(t)$ be two independent Poisson processes with rates λ_1 and λ_2 respectively. Let us define $N(t) = N_1(t) + N_2(t)$. This rando process is derived combining the arrivals as shown in Fig. (**??**). The claim is that $N(t)$ is the Poisson process with the rate $\lambda_1 + \lambda_2$. To see it we first note that $N(0) = N_1(0) + N_2(0) = 0$. Next, since $N_1(t)$ and $N_2(t)$ are independent and have independent increments their sum also have an independent increment. Finally, consider an interval of length τ, $(t, t + \tau]$. Then the number of arrivals in the interval are Poisson$(\lambda_1\tau)$ and Poisson$(\lambda_2\tau)$ and the two numbers are independent. Therefore the number of arrivals in the interval associated with $N(t)$ is Poisson$((\lambda_1 + \lambda_2)\tau)$ - as sum of two independent Poisson random variables. We can obviously generalize the statement to a sum of many Poisson processes. Note that in the case of Bernoulli process the story is identical provided that collision is counted as one arrival.

Splitting: Let $N(t)$ be a Poisson process with rate λ. Here, we split $N(t)$ into $N_1(t)$ and $N_2(t)$ where the splitting is decided

by coin tossing (Bernoulli process) - when an arrival occur we toss a coin and with probability p and $1-p$ add arrival to N_1 and N_2 respectively. The coin tosses are independent of each other and are independent of $N(t)$. Then, the following statements can be made

- N_1 is a Poisson process with rate λp.
- N_2 is a Poisson process with rate $\lambda(1-p)$.
- N_1 and N_2 are independent, thus Poisson.

8. Recitation. Examples of Bernoulli & Poisson Processes

+