1 Conditional probabilities and expectations

A conditional probability is denoted by $\mathbb{P}(A|B)$, which means *what is the probability of A happening, given that B happens.* Let's look at a few simple example. We denote by X the random variable that represents the number that you roll on a six-sided die.

1. What is the probability that you roll a 6 with a six-sided die? In formulas: $\mathbb{P}(X = 6)$.
2. What is the probability that you roll a 6, given that you roll at least a 4; $\mathbb{P}(X = 6|X \geq 4)$?
3. You can use the following formula to compute conditional probabilities:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)}. \tag{1}$$

   Check that this formula works by solving the second question again, but now with the formula.
4. Similarly to probabilities, we can also look at expectations. What is the expected number you roll with a six-sided die? In formulas: $\mathbb{E}[X]$.
5. What is the expected number that you roll, given that you roll at least a 4; $\mathbb{E}[X|X \geq 4]$?

2 The exponential distribution

The exponential distribution is defined in the following way. Suppose that X is exponentially distributed with parameter λ. Then $\mathbb{P}(X < t) = 1 - e^{-\lambda t}$.

1. Calculate $\mathbb{P}(X \geq t)$.
2. Calculate $\mathbb{P}(1 < X < 2)$.
3. Calculate the expectation of the exponential distribution with the following formula:

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)\, dt.$$

4. Use Equation (1) to prove the memoryless property of the exponential distribution:

$$\mathbb{P}(X > t + u|X > t) = \mathbb{P}(X > u).$$

More on the other side.

3 Mean queue length

We introduce $\rho = \lambda/\mu$ to make the calculations easier. In the $M|M|1$ queue we found that the probability of having i jobs in the system, in equilibrium, equals

$$p_i = (1-\rho)\rho^i.$$

1. Of course, the sum of all these probabilities should sum up to 1. Prove that $\sum_{i=0}^{\infty} p_i = 1$.
2. We can calculate the mean queue length using these probabilities;

$$\mathbb{E}[L] = \sum_{i=0}^{\infty} i p_i = \sum_{i=0}^{\infty} i(1-\rho)\rho^i.$$

   Calculate $\mathbb{E}[L]$.

4 Extension of the single-server queue

In the lecture we drew the transition diagram and calculated the equilibrium probabilities of the $M|M|1$ queue, which is a system where 1 job can be served at a time. In this set of questions, we will consider three extensions.

1. The $M|M|c$ queue is an extension of this model, where up to c jobs can get service simultaneously. Draw the transition diagram of the $M|M|c$ queue. Hint: suppose two jobs are getting service at the same time. The rate at which servers move from having 2 to 1 jobs, is equal to $2 \cdot \mu$. Calculate the equilibrium probabilities of the $M|M|c$ queue.
2. In the $M|M|1|k$ queue, only one job receives service at a time. The k in the name denotes that there are finitely many spots to wait in the queue. At any moment, there can be at most k jobs in this system. Whenever a job arrives and the system is full, it will be blocked and it will leave forever. Draw the transition diagram, calculate the equilibrium probabilities, and find the blocking probability; the probability that an arbitrary job will be blocked.
3. The $M|M|c|c$ model is a mix of the $M|M|1|k$ and the $M|M|c|c$. In this system, c jobs can receive service simultaneously, and at most c jobs can reside in the model. Can you find the transition diagram, equilibrium probabilities and blocking probability?

5 Other questions

1. For the 'random' dispatching, we showed differential equations, involving the fraction of servers that have i jobs in them at time t; $f_i(t)$. Do you understand these formulas?

$$\frac{df_0(t)}{dt} = \lambda f_0(t) + \mu f_1(t) \qquad \frac{df_i(t)}{dt} = \lambda(f_{i-1}(t) - f_i(t)) + \mu(f_{i+1}(t) - f_i(t)), i \geq 1.$$

2. For Power-of-2, we showed differential equations, involving the fraction of servers that have *at least* i jobs in them at time t; $g_i(t)$. Do you understand this formula?

$$\frac{dg_i(t)}{dt} = g_{i+1}(t) - g_i(t) + \lambda(g_{i-1}(t)^2 - g_i(t)^2), i \geq 1.$$

3. Can you think of a model where Join-the-Shortest-Queue is *not* smart to use?
4. In the presentation, you saw several load balancing algorithms like Join-the-Shortest-Queue and power-of-d. Can you think of your own load balancing algorithm?