Queueing Theory: Simulation of Queueing processes in continuous time

EBB074A05

Nicky D. van Foreest

2022:01:19

1 General info

This file contains the code and the results that go with the YouTube movies mentioned in the README file in this directory.

You should read the relevant section of my queueing book to understand what it going on. Most of it is very easy, but without background a bit cryptic (I believe).

I included a number of exercises to help you think about the code. Keep your answers short; you don't have to win the Nobel prize on literature.

2 Computing waiting times

Here we just follow the steps of the queueing book to construct a single server FIFO queue in continuous time.

2.1 Load standard modules

We need the standard libraries for numerical work and plotting.

Python Code

```python
import numpy as np
import matplotlib.pylab as plt
from matplotlib import style

style.use('ggplot')

np.random.seed(3)
```

2.2 Inter-arrival times

Simulate random interarrival times that are $\sim \text{Exp}(\lambda)$, with $\lambda = 3$. First I take just three jobs, so that I can print out all intermediate results and check how things work. Once I am convinced about the correctness, I run a simulation for many jobs.

Python Code

```python
labda = 3
X = np.random.exponential(scale=labda, size=3)
print(X)
```

```
[2.40084716 3.69452354 1.03129621]
```

Here is an important check (I always forget the meaning of λ when I provide it to the simulator)

Python Code

```
labda = 3
X = np.random.exponential(scale=labda, size=100)
print(X.mean())
```

```
2.5927889513302285
```

Ex 2.1. Explain that `scale=labda` sets the interarrival times to 3, but that in our queueing models, λ should correspond to the arrival rate. Why is the code below in line with what we want?

Python Code

```
labda = 3
X = np.random.exponential(scale=1/labda, size=3)
```

2.3 Arrival times

Python Code

```
A = X.cumsum()
print(A)
```

```
[1.19302331 1.82756237 2.08902011]
```

Ex 2.2. Why do we generate first random inter-arrival times, and use these to compute the arrival times? Why not directly generate random arrival times?

Check the numbers to see that the arrival time of job 0 is $A_0 > 0$. But I want time to start at time $A_0 = 0$. Here is the trick to achieve that.

Python Code

```
A = np.zeros(len(X) + 1)
A[1:] = X.cumsum()
print(A)
```

```
[0.         1.19302331 1.82756237 2.08902011]
```

This is better!

Ex 2.3. Why is the vector `A` one longer than `X`?

2.4 Service times

We have arrival times. We next need the service times of the jobs. Assume they are $\sim \text{Exp}(\mu)$ with μ somewhat larger than λ. (Recall this means that jobs can be served faster than that they arrive.)

Python Code

```
mu = 1.2 * labda
S = np.random.exponential(scale=1/mu,size=len(A))
S[0] = 0
print(S)
```

```
[0.         0.18642713 0.73061264 0.43038047]
```

Note, `S[0]` remains unused; it should corresponds to job 0, but we neglect this job 0 in the remainder.

Ex 2.4. Why do I use `size=len(A)` in the definition of `S`?

Ex 2.5. Why do we set `scale=1/mu`?

Ex 2.6. It's easy to compute the mean service time like this

Python Code

```python
S.mean()
```

Explain that we need to set `S[0]=0` to get the correct result.

2.5 Departure times

The standard recursion to compute the departure times.

Python Code

```python
D = np.zeros_like(A)

for k in range(1, len(A)):
    D[k] = max(D[k - 1], A[k]) + S[k]

print(D)
```

```
[0.         1.37945044 2.558175   2.98855547]
```

Ex 2.7. Explain now why it is practical to have $A_0 = 0$.

2.6 Sojourn times

How long do you stay in the system if you arrive at some time A_n and you depart at D_n?

Python Code

```python
J = D - A
print(J)
```

```
[0.         0.18642713 0.73061264 0.89953536]
```

2.7 Waiting times

If your sojourn time is 10, say, and your service time at the server is 3 (and there is just one server and the service discipline is FIFO), then what was your time in queue?

Python Code

```python
W = J - S
```

2.8 KPIs and plots

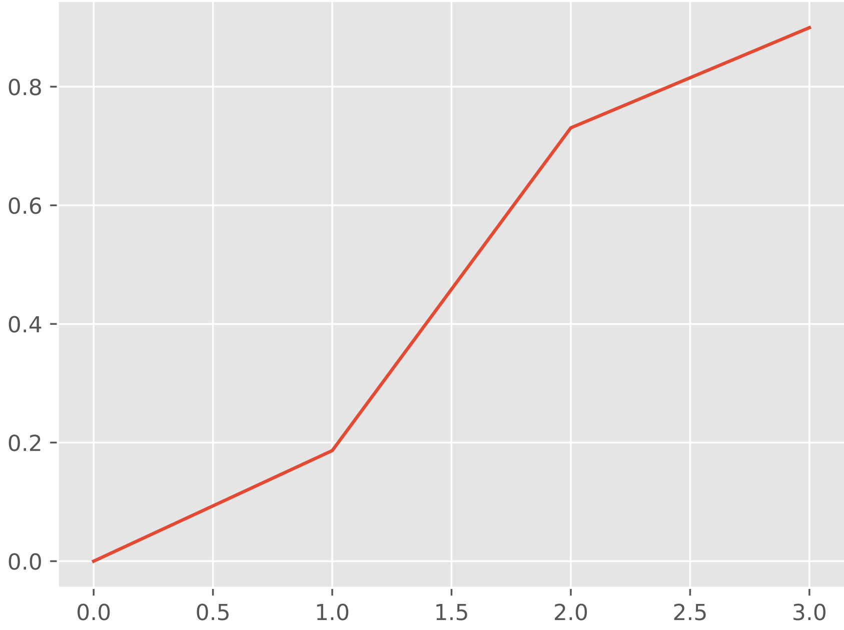
Python Code

```
print(J.mean(), J.std())
```

```
0.45414378095018604 0.37172821536485845
```

Python Code

```
plt.clf()
plt.plot(J)
plt.savefig("wait.pdf")
```



Ex 2.8. Change the simulation length to 1000 jobs. Do one run for $\mu = 3.5$ and another for 2.8. Compute the KPIs, make a plot, and include that in your assignment. Comment on what you see.

2.9 Server KPI: idle time

This code computes the total time the server is idle, and then computes the fraction of time the server is idle.

Python Code

```
rho = S.sum() / D[-1]
idle = (D[-1] - S.sum()) / D[-1]
print(idle)
```

Ex 2.9. Explain the code above. Some specific points:

1. Why is `S.sum()` the total busy time of the server?

2. Why do we divide by D[-1] in the computation of ρ?
3. Explain the computation of the idle variable.

The next code computes the separate idle times.

Python Code

```
idle_times = np.maximum(A[1:] - D[:-1], 0)
print(idle_times)
print(idle_times.sum())
print(D[-1] - S.sum())
```

Ex 2.10. Run this code for a simulation with 10 or so jobs (some other small number). Explain how this code works. Which line is a check on the computations?

2.10 Server KPI: busy time

We also like to know how a long the server has to work uninterruptedly. Finding the busy times is quite a bit harder than the idle times. (A busy time starts when a job arrives at an empty system and it stops when the server becomes free again.)

Ex 2.11. To help you understand the code, let's first do a numerical example. Suppose jobs $1, 4, 8$ find an empty system upon *arrival*. The simulation contains 10 jobs. Why do jobs $3, 7, 10$ leave an empty system behind upon *departure*?

With this idea, we can compute the idle times in another way (as a check on earlier work), and then we extend the approach to the busy times.

Python Code

```
import numpy as np

np.set_printoptions(suppress=True)
np.random.seed(3)

num = 10
labda = 3
X = np.random.exponential(scale=1 / labda, size=num)
A = np.zeros(len(X) + 1)
A[1:] = X.cumsum()
mu = 1.2 * labda
S = np.random.exponential(scale=1 / mu, size=len(A))
S[0] = 0
D = np.zeros_like(A)

for k in range(1, len(A)):
    D[k] = max(D[k - 1], A[k]) + S[k]


W = D - S - A # waiting times
idx = np.argwhere(np.isclose(W, 0))
idx = idx[1:]  # strip A[0]
idle_times = np.maximum(A[idx] - D[idx - 1], 0)
print(idle_times.sum())
```

Ex 2.12. What is stored in `idx`? Why do we strip `A[0]`? Why do we subtract `D[idx-1]` and not `D[idx]`? (Print out the variables to understand what they mean, e.g., `print(idx)`.)

Now put the next piece of code behind the previous code so that we can compute the busy times.

Python Code

```
busy_times = D[idx - 1][1:] - A[idx][:-1]
last_busy = D[-1] - A[idx[-1]]
print(busy_times.sum() + last_busy, S.sum())
```

Ex 2.13. Explain these lines. About the last line, explain why this acts as a check.

3 Computing Queue length

We have the waiting times, but not the number of jobs in queue. What if we would like to plot the queue length process?

A simple, but inefficient, algorithm to construct the queue length process is to walk backwards in time.

Python Code

```
import numpy as np
np.random.seed(3)

num = 10
X = np.random.exponential(scale=labda, size=num)
A = np.zeros(len(X) + 1)
A[1:] = X.cumsum()
mu = 0.8 * labda
S = np.random.exponential(scale=mu,size=len(A))
D = np.zeros_like(A)

for k in range(1, len(A)):
    D[k] = max(D[k-1], A[k]) + S[k]

L = np.zeros_like(A)
for k in range(1, len(A)):
    l = k - 1
    while D[l] > A[k]:
        l -= 1
    L[k] = k - l

print(L)
```

```
[0. 1. 1. 2. 2. 1. 1. 1. 2. 3. 3.]
```

Ex 3.1. Explain how this code works. At what points in time do we sample the queue length?

Ex 3.2. The above procedure to compute the number of jobs in the system is pretty inefficient. Why is that so?

Ex 3.3. Try to find a(more efficient algorithm to compute L. If you cannot solve this yourself, explain my code that is provided in the hint.

4 Multi-server queue

Let us now generalize the simulation to a queue that is served by multiple servers. Here is the code; see the queueing book to see how it works.

Python Code

```
import numpy as np

np.random.seed(3)

labda = 3
mu = 4
N = 3

X = np.random.exponential(scale=1 / labda, size=N + 1)
S = np.random.exponential(scale=1 / mu, size=N)

# single server queue
W = np.zeros_like(S)
for k in range(len(S)):
    W[k] = max(W[k - 1] + S[k - 1] - X[k], 0)

print(W)
print(W.mean(), W.std())

# code for multi server queue
c = np.array([1.0])
W = np.zeros_like(S)
w = np.zeros_like(c)
for k in range(len(S)):
    s = w.argmin()  # server with smallest waiting time
    W[k] = w[s]
    w[s] += S[k]  # assign arrival to this server
    w = np.maximum(w - X[k + 1] * c, 0)

print(W)
print(W.mean(), W.std())
```

```
[0.         0.14810509 0.60006325]
0.24938944623397144 0.255229135234328
[0.         0.14810509 0.60006325]
0.24938944623397144 0.255229135234328
```

Ex 4.1. First a test, we set the vector of server capacities `c=[1]` so that we reduce our multi-server queue to a single-server queue. Modify the code of the single server code so that it mimics the code for the multi server queue. Include your code.

BTW: such 'dumb' corner cases are necessary to test code. In fact, it has happened many times that I tested code of which I was convinced it was correct, but I still managed to make bugs. A bit of paranoia is a good state of mind when it comes to coding.

Now that we have tested the implementation (in part), here is the code for a queue served by three servers.

Python Code

```
import numpy as np

np.random.seed(3)

labda = 3
mu = 1.2
N = 1000

```

```python
X = np.random.exponential(scale=1 / labda, size=N + 1)
S = np.random.exponential(scale=1 / mu, size=N)

c = np.array([1.0, 1.0, 1.0])
W = np.zeros_like(S)
w = np.zeros_like(c)
for k in range(len(S)):
    s = w.argmin()  # server with smallest waiting time
    W[k] = w[s]
    w[s] += S[k]  # assign arrival to this server
    w = np.maximum(w - X[k + 1] * c, 0)

print(W.mean(), W.std())
```

```
0.9065985566616623 1.095184908323177
```

Ex 4.2. Run the code, and write down the mean and std of `W`. Then change the code for the multi-server such that the individual servers have different speeds, e.g., `c=np.array([2, 0.5, 0.5]). Like this, the total service capacity remains the same. What is the impact on the mean and std of ~W`? Include your results.

Ex 4.3. Once you researched the previous exercise, provide some consultancy advice. Is it better to have one fast server and several slow ones, or is it better to have 3 equal servers? What gives the least queueing times and variance? If the variance is affected by changing the server rates, explain the effects based on the intuition you can obtain from Sakasegawa's formula.

5 Server adjustments

Finally we consider a server that needs some adjustment during each job service. A very simple model is to assume that such adjustments are $\sim \mathrm{U}((a,b))$, for some a and b.

Ex 5.1. Can you write a computer program to let a server fail for a certain time? If not, explain my code that you can find in the hint. Then do a few experiments to provide insight into how adjustments affect the waiting times.

6 Hints

h.2.3. When we have n *arriving* jobs, how many *interarrival* times do we have?

h.2.4. If I would not do this, and I would want to change the simulation length (the number of jobs), at how many places should I change this number?

h.2.7. Observe that for D_1 we need D_0. If A_0 would be the arrival time of the first job, then what would we take for D_{-1}?

h.3.3. Here is the code.

Python Code

```python
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(3)

num = 4
labda = 3
```

```python
X = np.random.exponential(scale=1 / labda, size=num)
A = np.zeros(len(X) + 1)
A[1:] = X.cumsum()
mu = 0.3 * labda
S = np.random.exponential(scale=1 / mu, size=len(A))
S[0] = 0
D = np.zeros_like(A)

for k in range(1, len(A)):
    D[k] = max(D[k - 1], A[k]) + S[k]

L = np.zeros((len(A) + len(D), 2))
L[: len(A), 0] = A
L[1 : len(A), 1] = 1
L[len(D) :, 0] = D
L[len(D) + 1 :, 1] = -1
N = np.argsort(L[:, 0], axis=0)
L = L[N]
L[:, 1] = L[:, 1].cumsum()
print(L)

plt.clf()
plt.step(L[:, 0], L[:, 1], where='post', color='k')
plt.plot(A[1:], np.full_like(A[1:], -0.3), '^b', markeredgewidth=1)
plt.plot(D[1:], np.full_like(D[1:], -0.3), 'vr', markeredgewidth=1)
plt.savefig("wait4.pdf")
```

h.5. Here is the code

Python Code

```python
import numpy as np

np.random.seed(3)

labda = 3
mu = 4
N = 1000

X = np.random.exponential(scale=1 / labda, size=N)
A = np.zeros(len(X) + 1)
A[1:] = X.cumsum()
S = np.random.exponential(scale=1 / mu, size=len(A))
# R = np.zeros_like(A)
R = np.random.uniform(0, 0.1, size=len(A))

D = np.zeros_like(A)
for k in range(1, len(A)):
    D[k] = max(D[k - 1], A[k]) + S[k] + R[k]

W = D - A - S
print(W.mean(), W.std())
```