The prioritising exclusion process

Jan de Gier

University of Melbourne

GGI workshop, Firenze 2014

With Caley Finn

Motivation

The accumulating priority queue (APQ, Kleinrock 1964)

Queueing Syst (2014) 77:297–330
DOI 10.1007/s11134-013-9382-6

Waiting time distributions in the accumulating priority queue

David A. Stanford · Peter Taylor · Ilze Ziedins

Received: 31 December 2011 / Revised: 30 July 2013 / Published online: 7 December 2013

APQ

The accumulating priority queue (APQ):

- M/M/1 queue (arrival and service rate) with two classes of customers
- High and low priority customers (triage in emergency)
- Customers deterministically acquire priority
- High priority do so at a faster rate
- Customers are ordered according to their priority
- Hence they overtake deterministically
- Change accumulation rates to achieve service waiting times performance levels

Waiting time performance levels

Table 1 CTAS key performance indicators

Category	Classification	Access	Performance level (%)
1	Resuscitation	Immediate	98
2	Emergency	15 min	95
3	Urgent	30 min	90
4	Less urgent	60 min	85
5	Not urgent	120 min	80

of patients whose waiting times before accessing treatment should not exceed the stipulated standard. For example, as is depicted in Table 1, the Canadian Triage and Acuity Scale (CTAS) [4] formulates five priority classifications for assessment in an emergency department, each with its own time standard and compliance target for the proportion of that class's patients that need to meet that standard. The Australasian

Goals

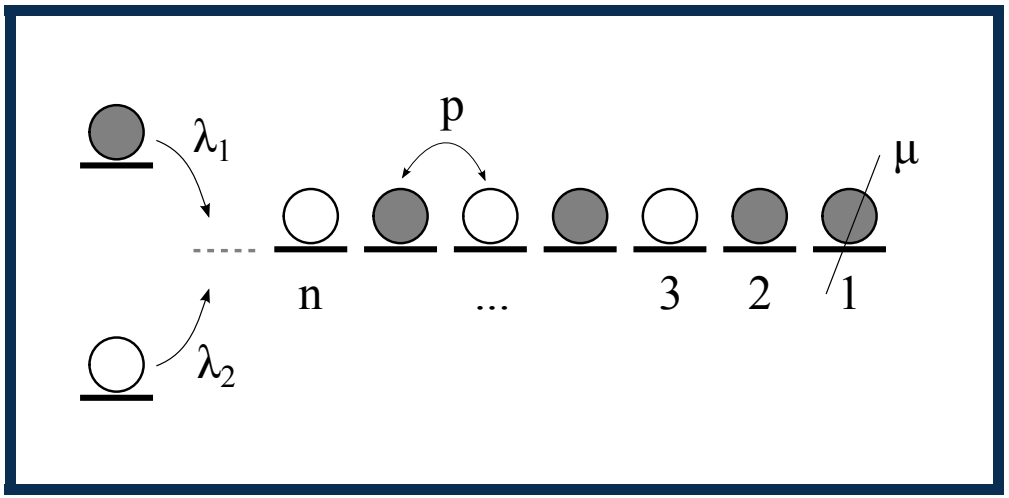
- APQ has fluctuating length
- Stochastic scheduling mechanism?
- Mapping to exclusion process

Goals:

- Stat-phys: Find a “solvable” particle hopping model with fluctuating length
- Queueing: What can we say using physics methods

Prioritising exclusion

The prioritising exclusion process (PEP)



Low and high arrivals with rate λ_1 and λ_2
High overtakes low with rate p
Service rate μ

M/M/1 queue

- The PEP is related to the simple $M/M/1$ queue.
- The total arrival rate of customers to the PEP is $\lambda = \lambda_1 + \lambda_2$, and the service rate is μ.
- Both these rates are independent of the internal arrangement of the queue, and the prioritising parameter p.
- If we are interested only in the total length of the queue, we can treat the system as a $M/M/1$ queue with arrival rate λ and service rate μ.

M/M/1 queue

The state of a $M/M/1$ queue is characterised by the probability distribution P_n of a queue of lenght n:

$$\frac{\mathrm{d}\, P_0}{\mathrm{d}\, t} = \mu P_1 - \lambda P_0$$
$$\frac{\mathrm{d}\, P_n}{\mathrm{d}\, t} = \lambda P_{n-1} + \mu P_{n+1} - (\mu + \lambda) P_n, \qquad n > 0.$$

The stationary length distribution of the $M/M/1$ queue (and hence for the PEP) is

$$P_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n,$$

when $\lambda < \mu$ (arrival rate less than service rate)

The expected queue length is finite, given by

$$\langle n \rangle = \frac{\lambda}{\mu - \lambda}.$$

We will call this the *bounded* phase of the PEP.

M/M/1 queue

- At $p = 0$, the PEP reduces exactly to a $M/M/1$ queue. Customers are high priority with probability λ_1/λ or low priority with probability λ_2/λ.

- But as there is no overtaking, the probability of high or low at any site is the same as at arrival.

  For $p = 0$, in the bounded phase,

$$P(\tau_n \dots \tau_1) = P_n \left(\frac{\lambda_1}{\lambda}\right)^h \left(\frac{\lambda_2}{\lambda}\right)^l.$$

  where

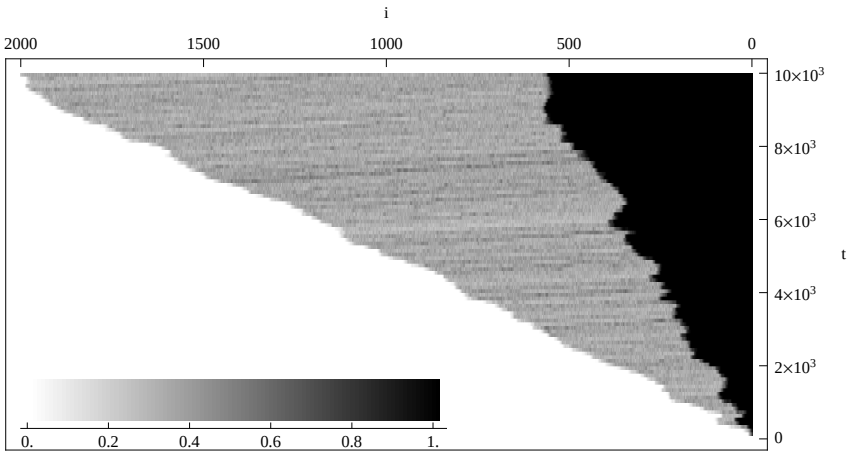
$$h = \sum_{i=1}^{n} \tau_i, \qquad l = n - h.$$

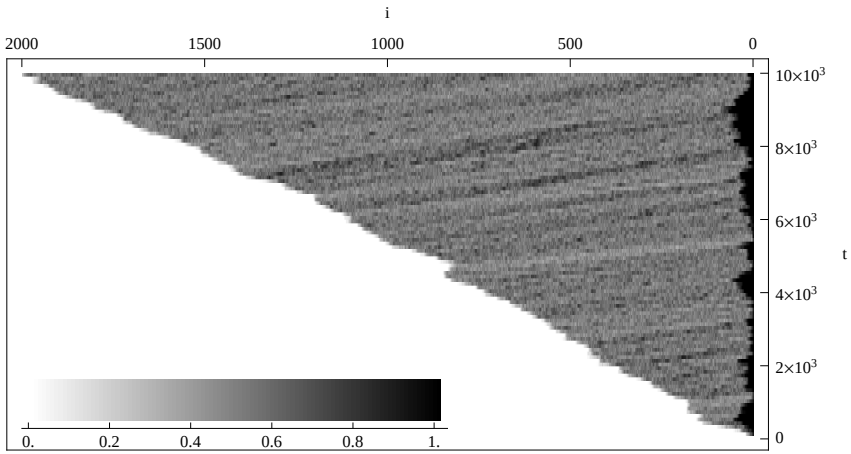
Unbounded phase

When $\lambda > \mu$, the system is *unstable* and the expected queue length grows as

$$\langle n \rangle \sim (\lambda - \mu)t.$$

In the late time limit, we can treat the queue as infinite in length. We call this the *unbounded* phase of the PEP.





Notation

Specify both the site, i, and the lattice length, n:

$$\langle \tau_i \rangle_n = P(\text{queue length is } n, \text{ and site } i \text{ is occupied}).$$

Correlation functions, number from right to left:

$$\langle \tau_{i_n} \tau_{i_2} \dots \tau_{i_1} \rangle_n, \qquad n \geq i_n > \dots i_2 > i_1 \geq 1.$$

Examples of stationary rate equations are

$$0 = \lambda_1 P_0 + \mu \langle \tau_2 \rangle_2 - (\lambda + \mu) \langle \tau_1 \rangle_1,$$

$$0 = \lambda \langle \tau_1 \rangle_{n-1} + \mu \langle \tau_2 \rangle_{n+1} + p \langle \tau_2 (1 - \tau_1) \rangle_n - (\lambda + \mu) \langle \tau_1 \rangle_n, \qquad n > 1.$$

These couple between different lenghts.

Unbounded queue; Reference frames

For any fixed configuration $\tau_n \dots \tau_1$,

$$\lim_{t\to\infty} P(\tau_n \dots \tau_1) = 0,$$

The *service frame* is fixed at the right hand end of the lattice:

$$P_{\text{serv}}(\tau_m \dots \tau_1; n) = \sum_{\tau_n,\dots,\tau_{m+1}=0,1} P(\tau_n \dots \tau_{m+1}\tau_m \dots \tau_1).$$

The *arrival frame* probability is defined by

$$P_{\text{arr}}(\tau_1 \dots \tau_m; n) = \sum_{\tau_{m+1},\dots,\tau_n=0,1} P(\tau_1 \dots \tau_m\tau_{m+1} \dots \tau_n).$$

In the $t \to \infty$ limit, the expected lattice length is infinite, and thus we are interested in the $n \to \infty$ limits

$$P_{\text{serv}}(\tau_m \dots \tau_1) = \lim_{n\to\infty} P_{\text{serv}}(\tau_m \dots \tau_1; n),$$
$$P_{\text{arr}}(\tau_1 \dots \tau_m) = \lim_{n\to\infty} P_{\text{arr}}(\tau_1 \dots \tau_m; n).$$

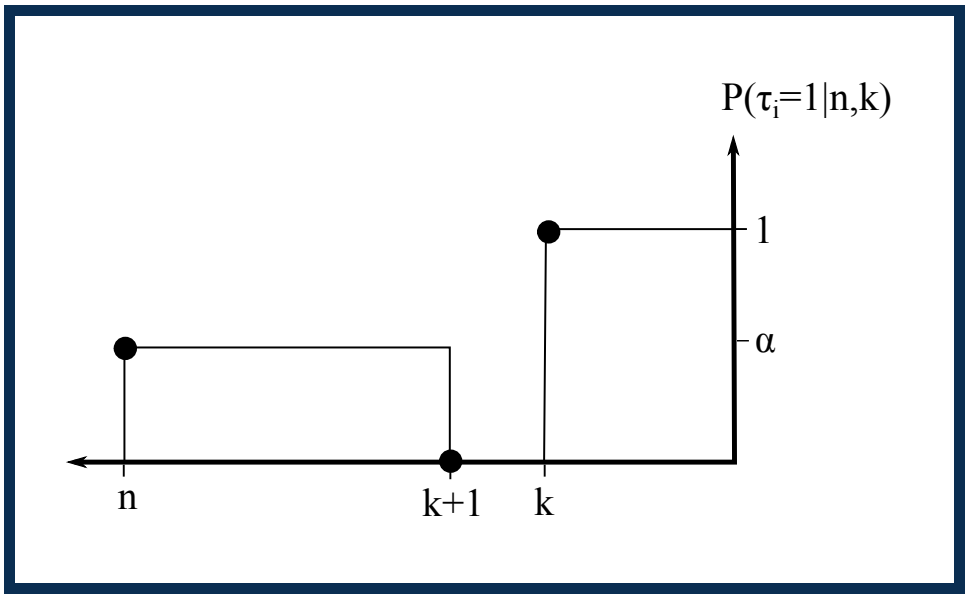
Domain wall ansatz

Consider a general but finite section of length m, with a length k jam:

$$\tau_m \dots \tau_{k+2} 01^k = \tau 01^k.$$

Assume that the conditional probability for a high at site i given a jam of length k is

$$P(\tau_i = 1|k) = \begin{cases} 1 & 1 \le i \le k \\ 0 & i = k+1 \\ \alpha & i \ge k+2, \end{cases}$$



Domain wall ansatz

The probability of the finite segment $\boldsymbol{\tau} = \tau_m \dots \tau_1$ is

$$P_{\text{serv}}(\boldsymbol{\tau} 01^k) = \sum_{\tau_\infty,\dots,\tau_{m+1}=0,1} P(\dots \tau_{m+1}\tau_m \dots \tau_{k+2} 01^k)$$
$$= \alpha^h (1-\alpha)^l P_{\text{jam}}(k).$$

where

$$h = \sum_{i=k+2}^{m} \tau_i, \qquad l = m - h - k - 1,$$

and $P_{\text{jam}}(k)$ is the probability of a length k jam. The jam probabilities are normalised such that

$$\sum_{k=0}^{\infty} P_{\text{jam}}(k) = 1.$$

Service frame

The stationary rate equation for the k-jam configuration with $k \geq 1$ is

$$
\begin{aligned}
0 = {} & \mu P_{\text{serv}}\left(\tau 01^{k+1}\right) + \mu P_{\text{serv}}\left(\tau 01^{k} 0\right) + p\tau_m P_{\text{serv}}\left(0\tau 01^{k}|_{(m+1,m)}\right) \\
& + \sum_{i=k+2}^{m} p(1-\tau_i)\tau_{i-1} P_{\text{serv}}\left(\tau 01^{k}|_{(i,i-1)}\right) + pP_{\text{serv}}\left(\tau 101^{k-1}\right) \\
& - \mu P_{\text{serv}}(\tau 01^{k}) - \sum_{i=k+2}^{m} p\tau_i(1-\tau_{i-1}) P_{\text{serv}}(\tau 01^{k}) - p(1-\tau_m) P_{\text{serv}}(1\tau 01^{k}).
\end{aligned}
$$

Service frame

Substituting the domain wall ansatz, the terms representing hopping combine and telescope to

$$p\left(\sum_{i=k+2}^{m}(1-\tau_i)\tau_{i-1}-\sum_{i=k+2}^{m}\tau_i(1-\tau_{i-1})\right)\alpha^h(1-\alpha)^l P_{\text{jam}}(k)$$
$$=-p\tau_m\alpha^h(1-\alpha)^l P_{\text{jam}}(k).$$

The factor $\alpha^h(1-\alpha)^l$ is common to all terms in the rate equation. Cancelling, and simplifying leaves

$$0=p\alpha P_{\text{jam}}(k-1)+\mu P_{\text{jam}}(k+1)+\mu(1-\alpha)\alpha^k P_{\text{jam}}(0)-(\mu+p\alpha)P_{\text{jam}}(k).$$

No approximation has been made.

Service frame

With boundary condition this implies the recursion

$$P_{\rm jam}(k) = \frac{p\alpha}{\mu} P_{\rm jam}(k-1) + \alpha^k P_{\rm jam}(0), \qquad k \geq 1.$$

with solution

$$P_{\rm jam}(k) = \frac{p\left(\frac{p\alpha}{\mu}\right)^k - \mu\alpha^k}{p-\mu} P_{\rm jam}(0).$$

The normalisation condition fixes

$$P_{\rm jam}(0) = (1-\alpha)(1-\frac{p\alpha}{\mu}),$$

subject to the constraint

$$p\alpha < \mu.$$

No growing jam.

What is α?

Arrival frame

The value of α can be determined by considering the arrival frame. One obtains:

$$0 = -\lambda\alpha(1-\alpha) + p\alpha^2(1-\alpha) - p\tau_1\alpha(1-\alpha) + \tau_1\lambda_1(1-\alpha) + (1-\tau_1)\lambda_2\alpha,$$

which for both $\tau_1 = 0$ and $\tau_1 = 1$ reduces to

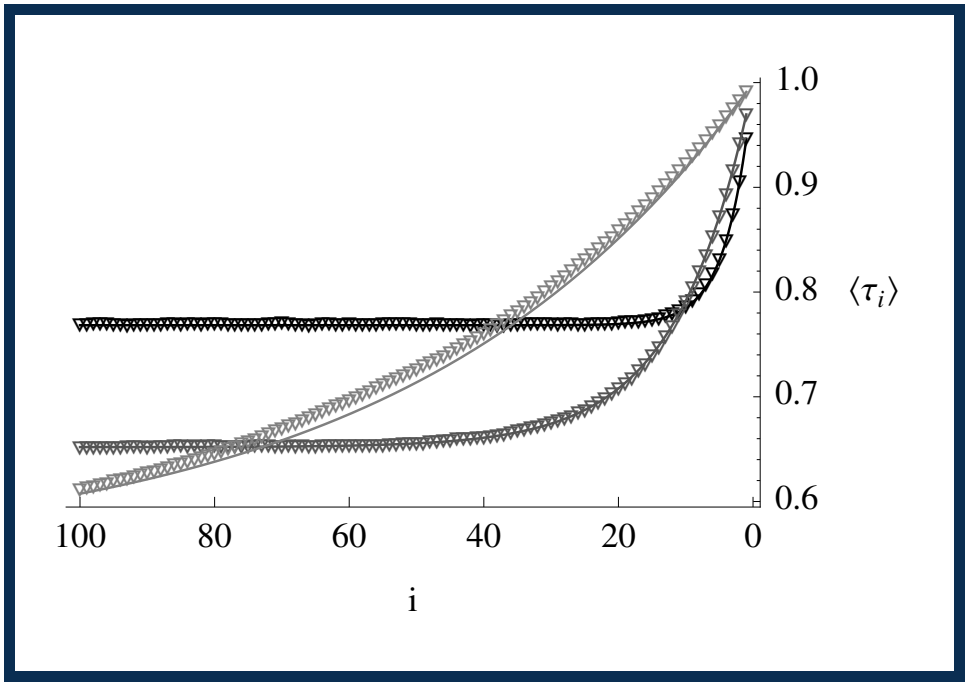
$$p\alpha^2 - (p+\lambda)\alpha + \lambda_1 = 0.$$

The physical solution is the smallest root of this quadratic.

Density profile

The density at site i in the service frame, $\langle \tau_i \rangle_\infty$, is computed from the domain wall solution as

$$\begin{aligned} \langle \tau_i \rangle_\infty &= \alpha \sum_{k=0}^{i-2} P_{\text{jam}}(k) + \sum_{k=i}^{\infty} P_{\text{jam}}(k) \\ &= \alpha + (1-\alpha) \left(\frac{p\alpha}{\mu} \right)^i . \end{aligned}$$



Conserved current

The bulk equations can be written

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\langle\tau_i\rangle_\infty = J_\infty^{(i+1)} - J_\infty^{(i)},$$

where

$$J_\infty^{(1)} = \mu\langle\tau_1\rangle_\infty, \qquad J_\infty^{(i)} = \mu\langle\tau_i\rangle_\infty + p\langle\tau_i(1-\tau_{i-1})\rangle_\infty.$$

Stationary current:

$$J_\infty = p\alpha(1-\alpha) + \mu\alpha$$

Low priority customers leave the queue at rate

$$\mu - J_\infty = (\mu - p\alpha)(1-\alpha).$$

For $p\alpha > \mu$, the jam becomes unbounded, and low priority customers will no longer be served.

Bounded phase $\lambda < \mu$

In the bounded phase the domain wall ansatz breaks down at the arrival end.

Define the probability of a length k jam in a length n queue

$$P(n,k) = \sum_{\tau_n,\ldots,\tau_{k+2}=0,1} P(\tau_n \ldots \tau_{k+2} 01^k). \tag{1}$$

Approximation *Length assumption*:

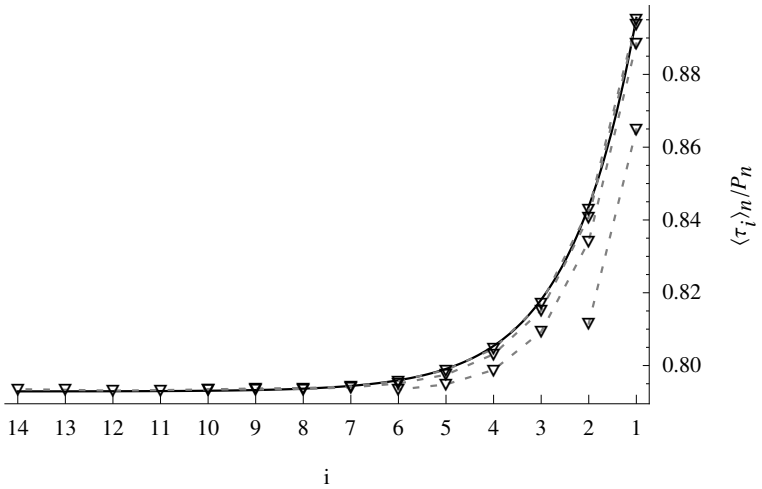
$$P(n,k) = P_n P^*_{\rm jam}(k).$$

The equations for $P^*_{\rm jam}(k)$ have the same form as for $P_{\rm jam}(k)$ in the unbounded case but with λ in place of μ.

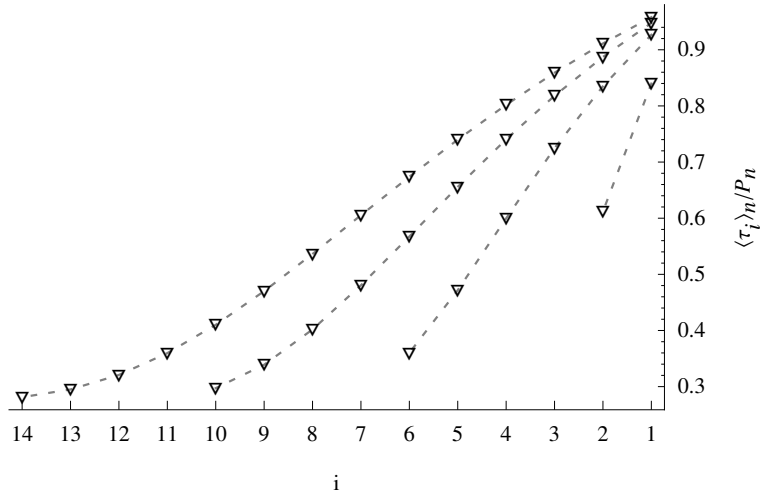
$$P^*_{\rm jam}(k) = \frac{p\left(\frac{p\alpha}{\lambda}\right)^k - \lambda\alpha^k}{p-\lambda} P^*_{\rm jam}(0).$$

Density profile

$$\langle \tau_i \rangle_n = P_n \left(\alpha + (1-\alpha) \left(\frac{p\alpha}{\lambda} \right)^i \right).$$



(c) $p\alpha < \lambda$



(d) $p\alpha > \lambda$

Triangle markers show simulation results, with points for each length connected by dashed lines.

Domain wall approximation invalid for $p\alpha > \lambda$.

Aggregate density profile and current

Another approximation can be derived for aggregate densities:

$$\overline{\langle \tau_i \rangle} = \sum_{n=i}^{\infty} \langle \tau_i \rangle_n.$$

$$\frac{\mathrm{d}}{\mathrm{d}\, t} \overline{\langle \tau_i \rangle} = J_{\mathrm{ext}}^{(i)} + \overline{J}^{(i+1,i)} - \overline{J}^{(i,i-1)},$$

where

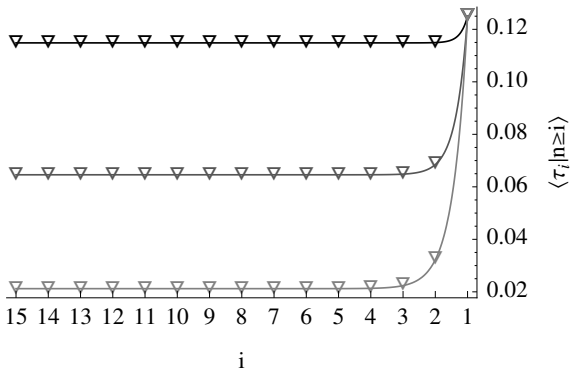
$$\overline{J}^{(i,i-1)} = \mu \overline{\langle \tau_i \rangle} + p \overline{\langle \tau_i (1 - \tau_{i-1}) \rangle}.$$

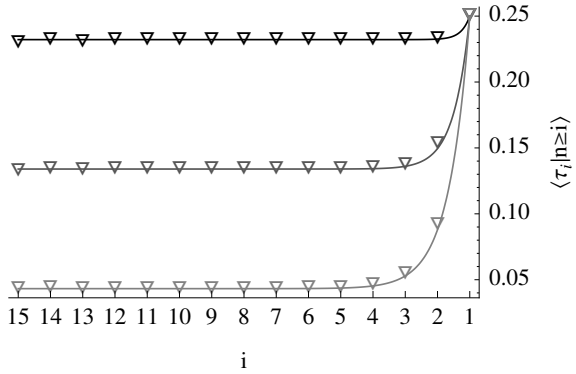
with stationary solution

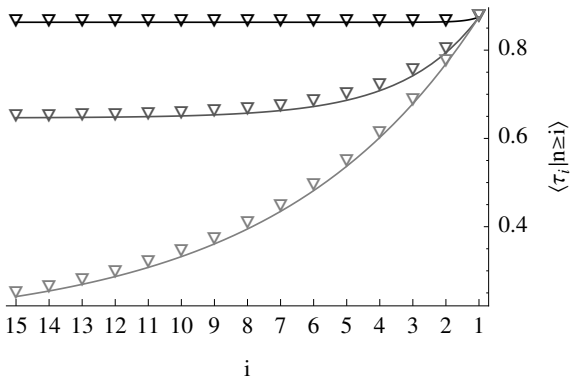
$$\overline{J}^{(i,i-1)} = \lambda_1 \left(\frac{\lambda}{\mu} \right)^{i-1}.$$

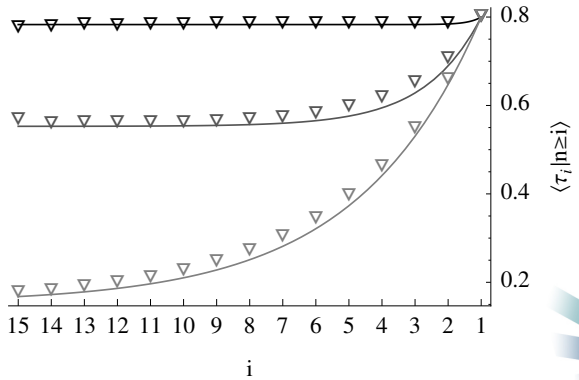
Aggregate density profile

$$\overline{\langle \tau_i \rangle} = \left(\frac{\lambda}{\mu}\right)^i \left[\alpha + (1-\alpha)\left(\frac{p\alpha}{\mu}\right)^i\right],$$









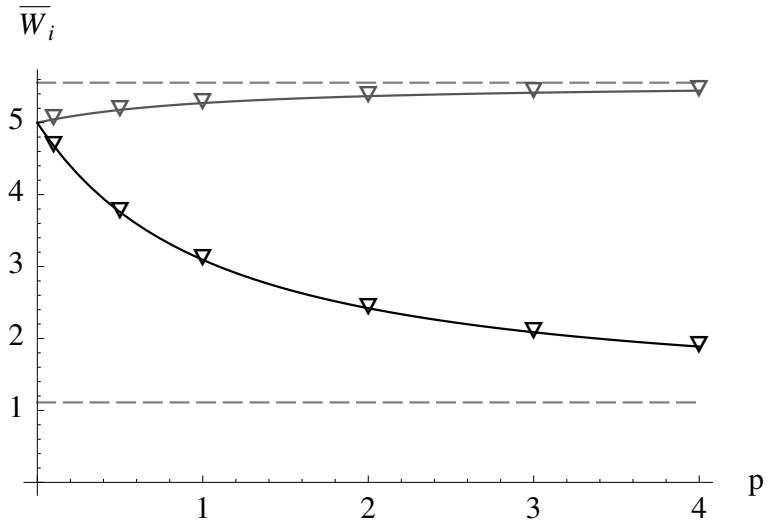
Waiting times

We can use the aggregated densities to compute the average number of customers in the queue
Define

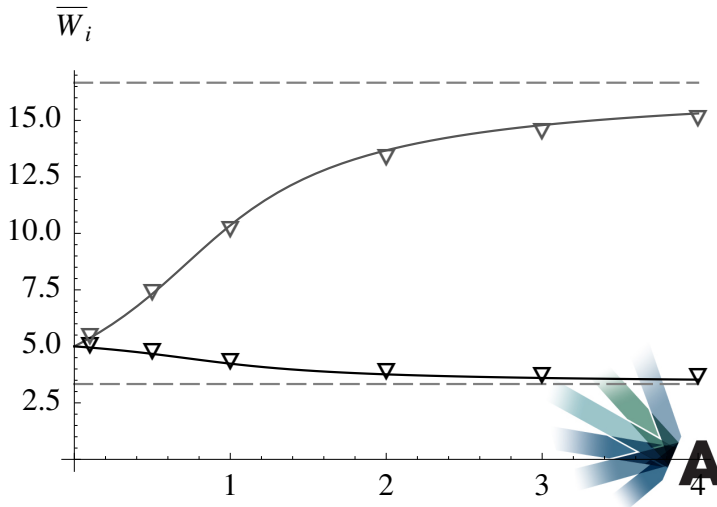
$$\overline{N}_1 = \sum_{i=1}^{\infty}\sum_{n=i}^{\infty}\langle\tau_i\rangle_n = \sum_{i=1}^{\infty}\overline{\langle\tau_i\rangle}.$$

Average waiting time of high priority customers is

$$\overline{W}_1 = \frac{1}{\lambda_1}\overline{N}_1 = \frac{1}{\lambda_1}\left(\alpha\frac{\lambda}{\mu-\lambda} + (1-\alpha)\frac{p\alpha\lambda}{\mu^2 - p\alpha\lambda}\right).$$



(i) Low density



(j) High density

Conclusions

Conclusion

- Queuing model with stochastic scheduling
- Fluctuating length
- Domain wall ansatz exact for the unbounded phase
- Two approximations for bounded phase
- Waiting times

Outlook

- Full exact solution
- More classes of particles
- Waiting time distributions
- Fluctuations and large deviations?

Papers

Exlusion processes:
arXiv:1403.5322
Exclusion in a priority queue
Jan de Gier and Caley Finn

arXiv:1107.2744, J. Phys. A 44 (2011) 405002
Relaxation rate of the reverse biased asymmetric exclusion process
Jan de Gier, Caley Finn and Mark Sorrell

Traffic modelling:
arXiv:1311.0230, Physica A 401 (2014), 82-102
Traffic disruption and recovery in road networks
Lele Zhang, Timothy M. Garoni and Jan de Gier

arXiv:1112.3761, Transport. Res. B: Meth. 49 (2013), 1-23
A comparative study of Macroscopic Fundamental Diagrams of arterial road networks governed by adaptive traffic signal systems
Lele Zhang, Timothy M. Garoni and Jan de Gier