How do we choose the rates in an Accumulating Priority Queue?

Azaz Bin Sharif*, David Stanford*, Peter Taylor† and Ilze Ziedins‡
*University of Western Ontario,
†University of Melbourne,
‡University of Auckland.

CTAS

The Canadian Emergency Department Triage and Acuity Scale
(http://www.calgaryhealthregion.ca/policy/docs/1451/Admission_over-capacity_AppendixA.pdf)

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

ATS

The Australasian Triage Scale

(http://www.acem.org.au/media/policies and guidelines/P06 Aust Triage Scale - Nov 2000.pdf)

Category	Access	Performance Level
1	Immediate	100%
2	10 minute	80%
3	30 minute	75%
4	60 minute	70%
5	120 minute	70%

In both cases, there are priority classifications, with associated access targets and proportions of time that the target should be met.

The KPIs are in terms of tails of waiting time distributions.

What would we like to know?

1. We would like to know whether it is possible to meet the targets for all customer classes simultaneously.
2. If so, we want to propose a queueing discipline that ensures that the targets are met.

The accumulating priority queue

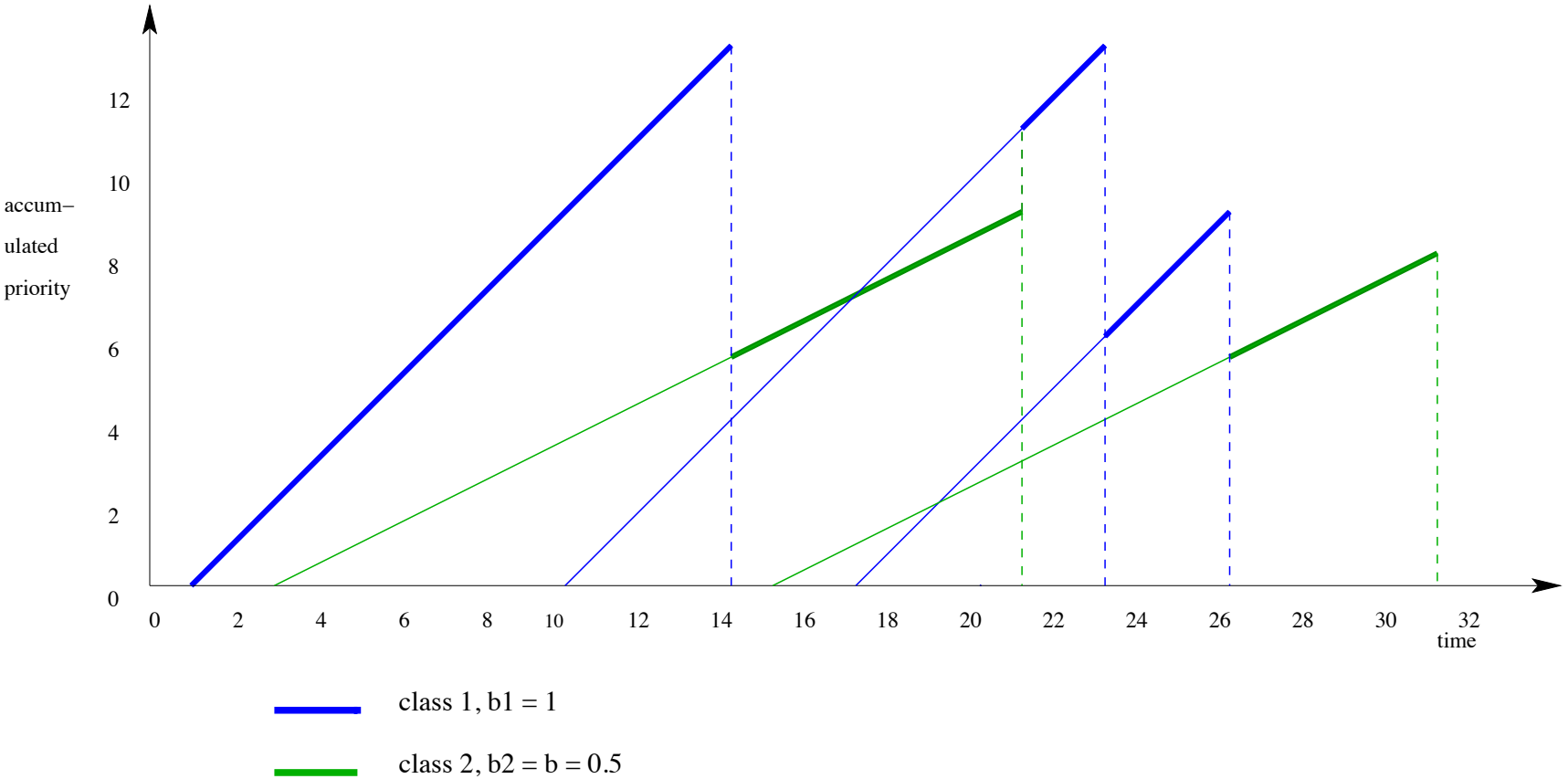
In 1964, Kleinrock proposed a queueing discipline where

- customers accumulate priority at class-dependent rates as linear functions of their time in the queue.
- when the server becomes free, it selects the waiting customer with the highest amount of accumulated priority at that instant, provided that the queue is non-empty.

The accumulating priority queue

- There is a single server.
- Customers of priority i arrive according to independent Poisson streams with rate λ_i.
- They accumulate priority at rate b_i where $1 = b_1 > b_2 > \ldots > b_l > 0$.
- When the server becomes free it chooses to serve the customer with the highest current priority, if the system is non-empty.
- Service times are chosen independently from the class-dependent distribution function $B_i(t)$ with Laplace-Stieltjes Transform $B_i^*(s)$.

The APQ


accum–
ulated
priority
0
2
4
6
8
10
12
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
time
class 1, b1 = 1
class 2, b2 = b = 0.5

Mean waiting Times

For such a queue, Kleinrock developed a recursion for calculating the expected waiting times of customers from each class.

He showed that

$$W_i = \frac{[\hat{M}/(1-\rho)] - \sum_{j=i+1}^{I} \rho_j W_j [1 - b_j/b_i]}{1 - \sum_{j=1}^{i-1} \rho_j [1 - b_i/b_j]}$$

where

- M_i is the mean service time of type i customers,
- $\rho_i = \lambda_i M_i$ with $\rho = \sum_{i=1}^{I} \rho_i$, and
- $\hat{M}$ is the stationary mean residual service time.

Mean waiting Times

The above recursion can be inverted to deliver the b_i that produce desired ratios W_i/W_1 of the mean stationary waiting times (within a range restricted by these ratios for a pure priority system).

So if our performance standards were specified in terms of

- mean waiting times, rather than
- tails of waiting time distributions,

we would have the answer to both of our questions.

Waiting time distributions

Theorem (Stanford, Taylor and Ziedins (2014))

For a single-server APQ, the LST $\tilde{W}_+^{(i)}(s)$ of the waiting time distribution for a class-i customer, conditional on it being positive, is given by

$$\tilde{W}_+^{(i)}(s) = (1 - b_{i+1}/b_i)\tilde{W}_{acc}^{(i)}(s) + (b_{i+1}/b_i)\tilde{W}_+^{(i+1)}(b_{i+1}s/b_i),$$

where

$$\begin{aligned}\tilde{W}_{acc}^{(i)}(s) &= \frac{1-\rho}{1-\delta_i}\tilde{W}_{acc}^{(i,0)}(s) + \tilde{W}_+^{(i+1)}(b_{i+1}s/b_i)\sum_{j=1}^{i}\frac{\rho_j b_{i+1}}{b_j(1-\delta_i)}\tilde{W}_{acc}^{(i,j)}(s) \\ &+ \sum_{j=i+1}^{I}\frac{\rho_j}{1-\delta_i}\tilde{W}_+^{(j)}(b_i s/b_j)\tilde{W}_{acc}^{(i,j)}(s),\end{aligned}$$

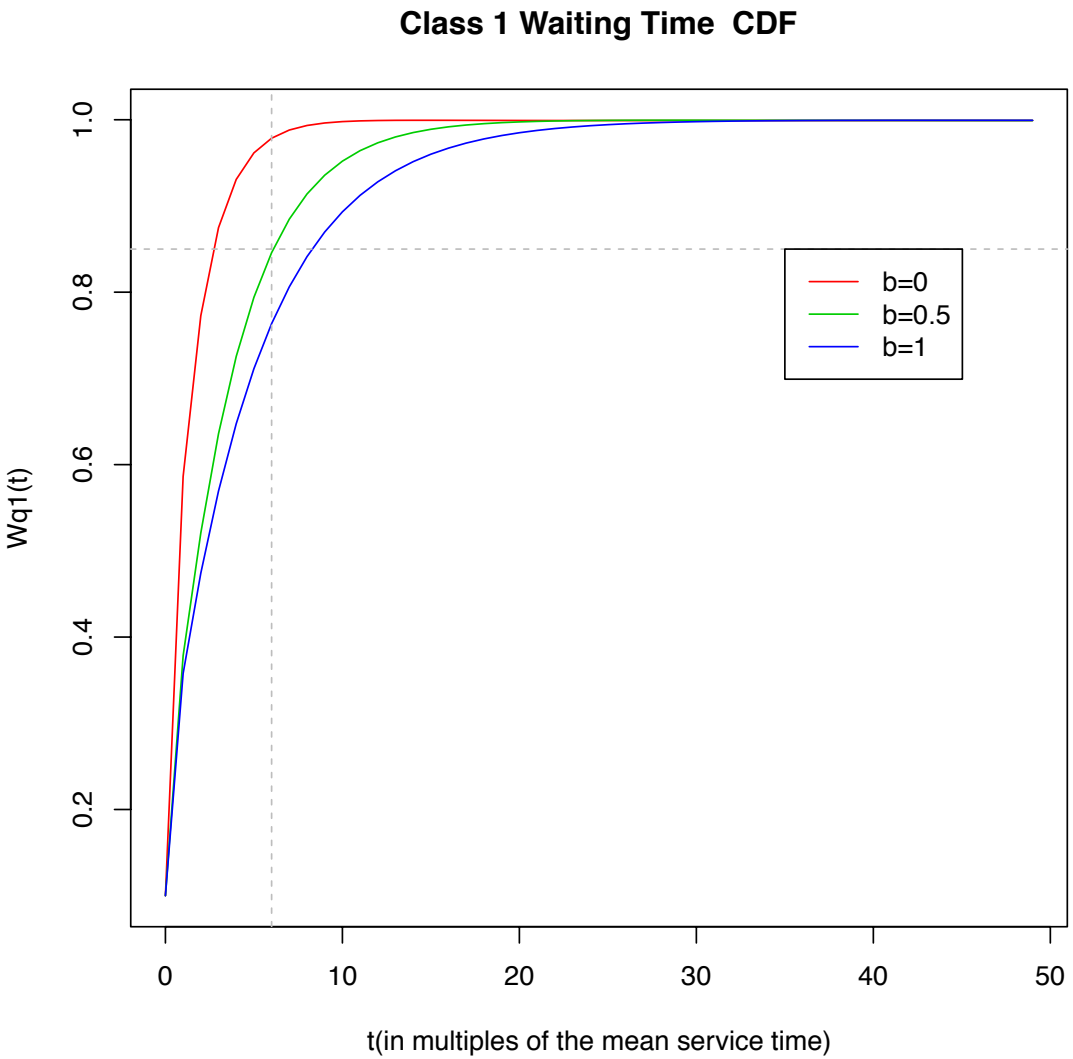
and $\delta_i = \sum_{j=1}^{i}\rho_j(1 - b_{i+1}/b_j)$.

Choosing the accumulation rates

For a given set of workload parameters $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_I)$ and tuning parameters $\mathbf{b} = (b_1, \ldots, b_I)$, we can

- use the STZ recursion to evaluate the Laplace transform of the waiting time at any point s that we are interested in,
- numerically invert the transform to derive the waiting time distributions $W^{(i)}(t)$, and
- test whether the performance standards in terms of tails are met.

Choosing the accumulation rates



Choosing the accumulation rates

So,

- we have a way of testing whether any $\mathbf{b} = (b_1, \ldots, b_l)$ is feasible, but
- we do not have a way of defining the feasible region for $\mathbf{b} = (b_1, \ldots, b_l)$ analytically.

- We also haven't specified any objective function.

What should our objective be?

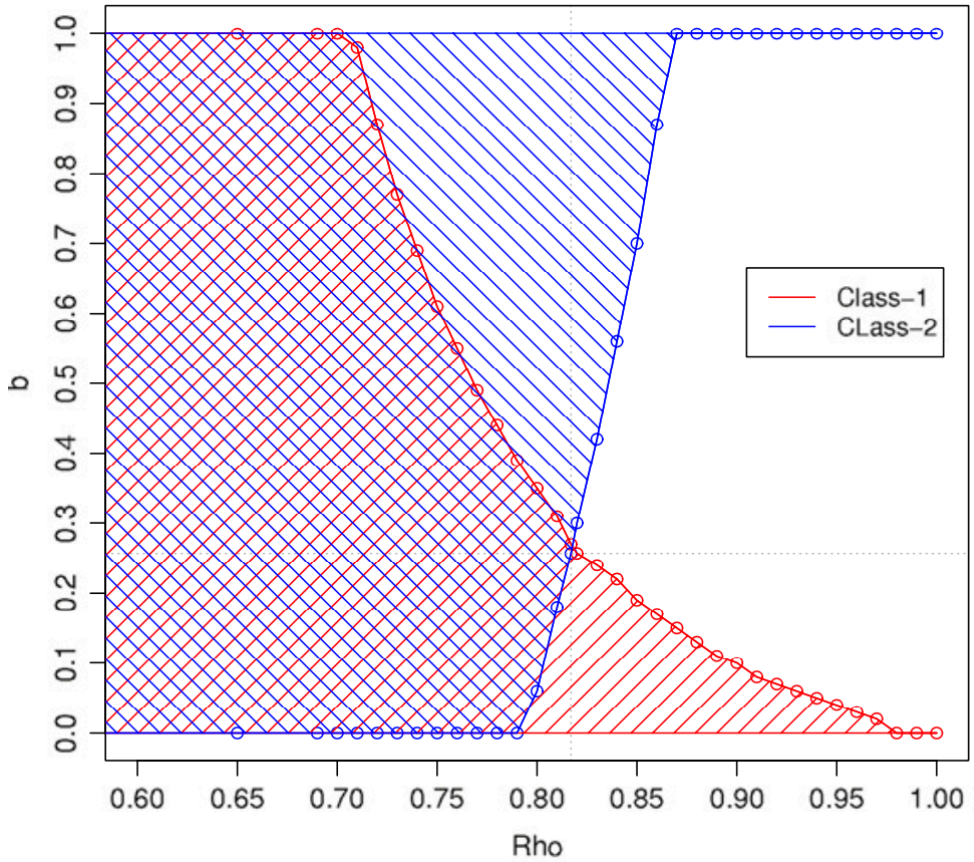
David Stanford gave a talk at the ‘Queues, Modelling and Markov Chains’ Workshop in which he suggested that we should minimise the expected waiting time of those patients that do not meet the performance KPIs.

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

This is a good idea from a ‘human’ point of view, but ....

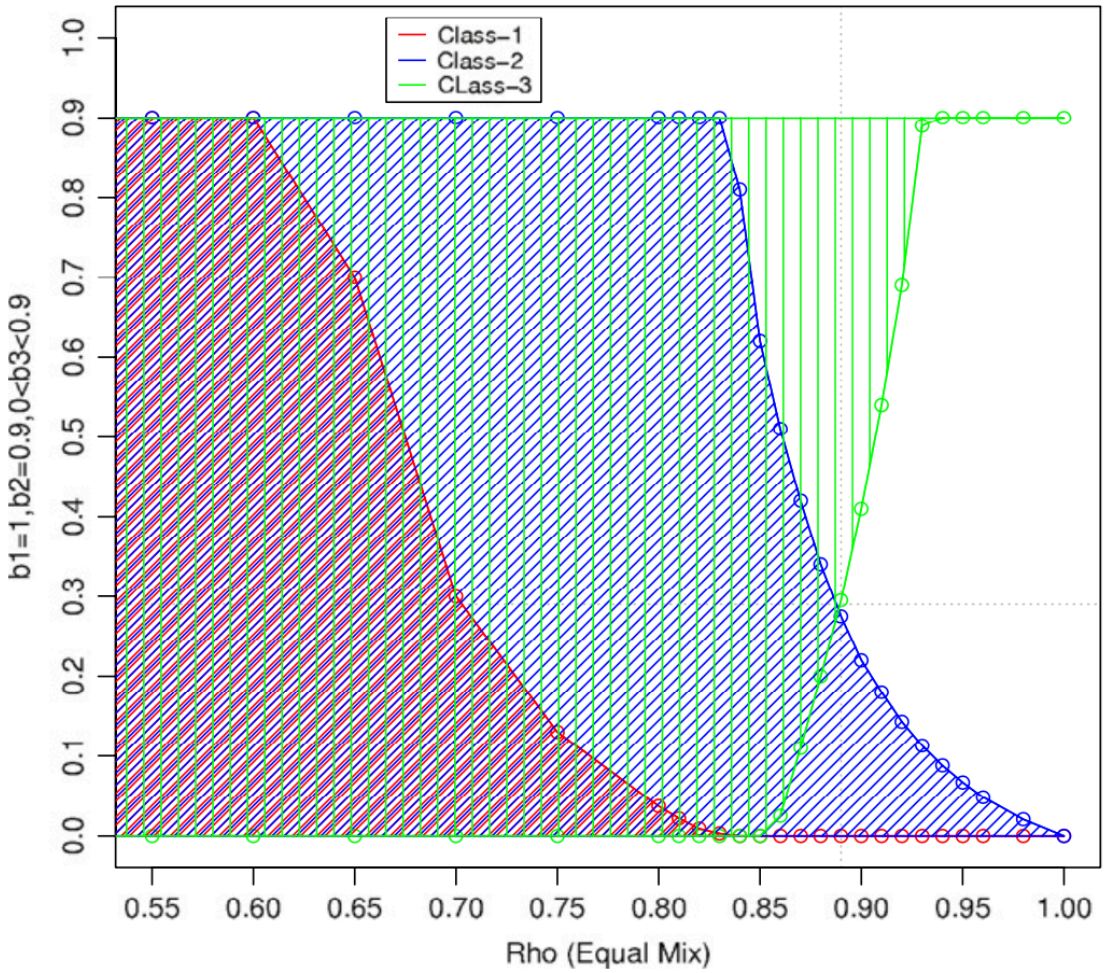
What should our objective be?

Another possibility is to choose the value of the $\mathbf{b} = (b_1, \ldots, b_l)$ that satisfy the constraints for the greatest range of $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_l)$.



Feasibility range for increasing ρ with $\lambda_1 = \lambda_2$.

What should our objective be?



What should our objective be?