Detailed Literature Review

September 8, 2015

A Study of Queues and Appointment Systems in Hospital Out-Patient Departments, with Special Reference to Waiting-Times

Author:	Norman T. J. Bailey
Published:	January 1952

- Investigation of patients' waiting time and time consultant waste waiting for next patient
- Recommend giving patients appointments at regular intervals equal to the average consultation time
- Keeping patients waiting longer is undesirable on humanitarian grounds
- Congestion in waiting room wastes hospital accommodation that is in short supply
- Suggests that consultant's waiting time is 10, 37.5 or 100 times more valuable than the patient's waiting time
- Consider the size of queue that may build up and the consequences of having longer or shorter clinics
- Consultation time vary considerably from patient to patient
- Generally assume patients arrive punctually for their appointment
- Serve patients one at a time
- Assume consultation times are independent although that might not hold in practice
- Fitted the frequency distribution of consultation time to a Gamma distribution
- Used random numbers to randomly generate consultation times (chose a rather simplistic method)
- Assumed average consultation time was 5 minutes (could be changed by multiplying all times by a constant factor)
- Only consider systems where the appointment intervals are all the same
- Two main ways of controlling the progress of a clinic:
  - Using shorter / longer appointment intervals
  - Varying the number of patients already present when the consultant arrives
- Patients' and consultant's waiting times are inversely related
- Consultant's wasted time will likely be less due to patients arriving early
- Is it beneficial to have patients arrive before consultants? (should arrival times be negative for first few patients?)
- Optimum results when consultant arrives at the same time as the second patient

- As the appointment interval increase, the consultant's wasted time increases rapidly in comparison to the decrease in patient's waiting time
- Considered the distributions of waiting times
- The distribution of patients' waiting times is skew with a long tail
- Didn't consider situations with more than 25 patients
- Consultant's idle period does not change much for less than 25 patients if the total time remains the same
- Patients with late appointments will wait longer than those with early appointments
- Possibility of making average waiting time more uniform (i.e., minimise variance of average waiting time)
- The waiting times of the 25 patients in one clinic are fairly correlated
- A great deal of time wasted by patients could be reduced without a significant increase in consultant idle time
- Optimum when patients are giving appointments at regular intervals
- Waiting times are sensitive to small changes in appointment interval (i.e., the average consultation time)
- If patients arrive early, there will be more congestion in waiting rooms
- Greater / fewer number of patients doesn't change general conclusions
- Only consider clinics with one server

A Survey of Queuing Theory Applications in Healthcare

Author:	Samuel F. Fomundam, Jeffrey W. Herrmann
Published:	2007

- Patients arrive, wait for service, obtain service, and then depart
- Queuing theory can be useful in real-world healthcare situations
- Preater (2002) provide a brief history, extensive bibliography (without descriptions), and discussion of the relationship between delays, utilisation, and the number of servers
- Systems at different scales: individual departments, healthcare facilities, and regional health systems
- Does not review simulation studies
- Queuing models are simpler, require less data, and provide more generic results than simulation (Green 2006a)
- Tucker (1999), and Kao and Tung (1981) use simulation to compliment the results obtained from queuing theory
- Resources necessary to attain the goals described by healthcare providers
- Appointment systesm look to reduce patient waiting without greatly increasing server idleness
- Minimising customer waiting times and maximisising server utilisation are conflicating goals
- Reneging: a patient forgoes a service because he doesn't want to wait any longer

- Probability that a patient reneges increases with queue length and with the patient's estimate of their future waiting time
- Broyles and Cochran (2007) look at the percentage of patients that renege in an emergency department
- Many healthcare systems have a variable arrival rate due to time of day, day of week, or season of year
- Arrival rates can depend on system state
- A system with congestion discourages arrivals
- Worthington (1991) models a system with arrival rates that discrease linearly with queue length and expected waiting time
- McQuarrie (1983) discuss minimising waiting times by giving priority to clients who require shorter service times
- Similar to shortest processing time rule (although this is perceived as unfair in practice)
- Priority discipline reduces the average wait time for all patients, but lower priority patients endure a longer average waiting time
- Probability that a patient would have to wait more than a certain amoutn of time
- Fiems et al. (2007) investigate the effect of emergency request on the waiting times of scheduled patients with deterministic processing times and discrete time - emergency patients interrupt the scheduled patients
- Blocking occurs when a queuing system places a limit on queue length and turns away walk in patients when its waiting room is full (e.g., limited number of beds)
- Limited waiting times is an important objective in a healthcare systems
- Bailey (1954) establishes threshold capacity where service sypply equals demand
- Agnihothri and Taylor (1991) seek optimal staffing at a hospital scheduling department
- Green (2006b) adjusts staffing to reduce the percentage of patients that renege
- Bruin et al. (2005) determine the number of beds required to achieve a turn away rate of 5%
- Bruin et al. (2007) suggest too few beds downstream is the primary cause of reduced admissions upstream
- Assign costs to patient waiting time and to each server
- Find the resource allocation that costs the least
- Gupta et al. (1971) discuss a system where non-routine requests are superimposed on top of routine, scheduled requests
- Rosenquist (1987) suggests scheduling patients when possible and segregating patients based on expected examination duration would reduce variability and decrease expected waiting times
- Systems with appointments reduce the arrival variability and waiting times at the facility
- Systems with appointments require patients to wait outside the facility (lower cost to the patient and facility)
- Bailey (1952, 1954) proposes appointment interval and consultant arrival time as being the two key variables that determine the efficacy of an appointment system
- Relative values of patient time and consultant time

- Appointment times at intervals equal to the average patient processing time
- Consultant should arrive at the same time as the second patient
- Brahimi and Worthington (1991) design an appointment system to reduce the number of patients in the queue at any time, and reduce patient waiting time without significantly increasing doctor idle time
- Also explore the effect of patients who do not show up and if there is a maximum number of patients allowed at any time
- Vasanawala and Desser (2005) describe a system where emergency request may require rescheduling of scheduled requests
- DeLaurentis et al. (2006) point out that patient no shows could lead to waste of resources
- Bottlenecks lead to high utilization and increase patient waiting times even though other nodes might have low utilitization
- Distinguish between three different scales of healthcare organisations
- Larger organisations with more patients are able to attain the same quality of service at higher utilizations than smaller organisations
- The ways in which distinct queuing systems within an organisation interact - links between subsystems

Appointment Policies in Service Operations: A Critical Analysis of the Economic Framework

Author:	Susana V. Mondschein, Gabriel Y. Weintraub
Published:	May 2002

- Literature assumes that demand is exogenous and independent of customers' waiting times
- Objective functions used in literature are appropriate only in the case of a central planner with demand unresponsive to waiting time
- Long waiting times cause customers to become dissatisfied with the service received
- Two important characteristics are customer's perception of waiting time and actual waiting time
- Can reduce perception of waiting time through more comfortable environment and a socially fair (first-in-first-output) policy
- If the time between appointments is short, then customer's waiting times will be long and expected server idle time will be short
- Study appointment policies under different scenarios through simulation
- Central planner maximises social welfare including the utility of the server and the customers
- Common to find institutions with poor services from customer's point of view
- Chilean IRS has reduced customers' waiting time significantly
- Question the assumption that demand is exogenous
- In practice, it is common to find services that never reach steady state so can't use results from queuing theory
- Some papers explore scheduling N people in K predetermined instances of time
- Other papers look at optimal vector $X = (x_1, \dots, x_N)$, which explores all possible alternatives

- No universal appointment rule
- Appointment policies for customers with different service time distributions
- Best policy is increasing order of service time variance
- Punctuality and variance of service time are key variables
- Obtive function is linear combination of expected total customer waiting time and server completion time
- Assume N is fixed and independent of waiting times
- Can include other ideas in objective function
- Optimal policy depends on $\frac{\beta}{\alpha}$ (ratio of relative costs), but these are difficult to estimate
- Crucial to know if central planner or private server
- Customers are risk averse in regards to waiting times (need to minimise variance)
- Appointment policy that leads to long waiting times results in a reduction in quantity demanded
- Overbooking can be implemented when absenteeism is high
- Demand being independent of waiting times is a good assumption when the service custome sells a 'prime necessity' service (e.g., public hospital) as customers place a high value on the service
- Attempt to include an objective function that does not have to schedule all N available slots (maybe it can be run for a range of N values and choose the best)
- If $\frac{\beta}{\alpha}$ is correctly estimated, then the objective function should be reasonable
- Numerically determine the impact of an appointment policy
- Ho and Lau (1992) study the performance of eight different appointment rules under different scenarios
- Customer's behaviour is not considered in the models
- α and β are functions of several fundamental parameters (e.g., customers' willingness to pay for the service, customers' value of waiting time and the server's value of his/her time)
- Percentage of absenteeism can be estimated using historical data
- Yang, Lau and Quek (1998) construct a general appointment rule, but this could be extended to include a larger number of servers or walk-ins
- Demand must depend on waiting time
- Vast majority of servers are now private (including medical services), so face competitive environments
- More realistic models are more difficult to solve than those traditionally used in the literature

Optimal Outpatient Appointment Scheduling

Author:	Guido C. Kaandorp, Ger Koole
Published:	23 May 2007

- Local search procedure to find optimal schedule
- Weighted average of expected waiting times, idle tile of doctor and tardiness as objective

- Trade off the interest of physicians and patients
- Bailey and Welch introduced first advanced scheduling rule
- Cayirli and Veral give a good overview
- Service time durations are exponentially distributed with rate μ
- Patients arrive on time
- No shows are allowed
- Setting is discrete time
- Tries to find neighbouring appointment schedules that are better
- As proved multimodularity, locally optimal schedule is globally optimal
- Tardiness is the probability that the session exceeds the planned finishing time multiplied by the average excess
- For many intervals, the computation times can be quite long
- Some papers evaluate schedules (often using simulation) and other design algorithms to find good schedules
- Can also consider patients not arriving on time and non-exponential service times
- Other papers focus on continuous time, which deal with finding the optimal interarrival intervals
- Pegden and Rosenshine find the optimal arrival moments assuming convexity of the objective in the interarrival times
- Lau and Lau give a procedure for finding optimal arrival instants assuming convexity
- Hassin and Mendel extend the model to no shows
- Branch and bound only works for small instances (for discrete time)
- Inclusion of different types of patients is relatively straightforward
- N patients to be scheduled
- Assume service time of patients are independent and exponentially distributed
- Assume patients always come on time
- Due to exponential service times, the potential number of departures in an interval has a Poisson distribution
- No shows occur frequently in practice
- Assume that all patients have the same no show probability and no shows are independent
- Need a search algorithm to reduce computation time by starting with a feasible solution and trying iteratively to find a better solution in the neighbourhood of that solution
- For a well chosen neighbourhood, it is possible to show that the algorithm finds a global minimum
- Neighbourhood can be interpreted as moving patients from time slot t to slot $t-1$ (online for computational reasons only moves one patient)
- Take $\dfrac{1}{\mu} = 20$ min and the probability of no-shows as 10%

- If the waiting time has a big weight then the patients are more spread out
- The times between consecutive arrivals first increases and then decreases again (dome shaped form commonly discussed in the literature)
- Individual block schedule: same number of intervals as patients and one patient in each interval
- Bailey-Welch rule: same but last patient is moved to beginning of day
- Optimal schedule is always better
- Bailey-Welch rule is indeed optimal for certain parameter values
- If the probability of no-shows increases, then the waiting time, idle time and tardiness becomes larger
- Changing the number of intervals would lead to more simultaneous arrivals

Outpatient Scheduling in Health Care: A Review of Literature

Author:	Tugba Cayirli, Emre Veral
Published:	January 2009

- Comprehensive survey of research on appointment scheduling
- Resources are better utilized and patient waiting times are minimized
- Health care providers are under a great deal of pressure to reduce costs and improve quality of service provided (Goldsmith 1989)
- Excessive waiting time is often the major reason for patients' dissatisfaction in outpatient services
- Focus on outpatient scheduling
- Find an appointment system for which a particular measure of performance is optimized
- In the static case, all decisions must be made prior to the beginning of a clinic session (most common appointment system in health care)
- Presence of unpunctual patients, no-shows, walk-ins and/or emergencies may upset the schedule
- Doctors may be late to star a clinic session
- Almost all student model a single-stage system where patients queue for a single service
- Doctors usually have thir own list of patients (desirable)
- Positive relationship between waiting times and the number of appointments in a session
- Unpunctuality is defined as difference between a patient's appointment time and actual arrival time truncated at zero
- Empirical studies suggest patients arrive early more often than late
- Patient earliness can be undesirable as it creates congestion
- Patient unpunctuality is assumed to be independent of scheduled appointment time
- The larger the no-show probability, the greater risk that the doctor will stay idle and decrease the waiting time of patients
- No-show probability is a major factor that affects performance of an appointment system
- Presence of walk-ins (regular and emergency) is neglected in most studies

- Few (if any) model the reneging behaviour of walk-ins
- Even less focus on emergencies who might preempt the current consultation
- Companions can increase congestion even though they aren't served
- Consultation time is the all the time that the patient is claiming the doctor's attention
- Majority of studies assume patients are homogeneous
- General assumption that service times and waiting times may be questionable (doctors speed up service when busy)
- Service times are observed to be unimodal and right skewed
- Most studies use exponential distributions to make model tractable
- Coefficient of variation: $CV = \frac{\sigma}{\mu}$ is a common measure for variability of service times (range from 0.35 to 0.85)
- Most studies show the relative performance of an appointment system is not affected by skewness, but only by mean and variance
- High variability of service times deteriorates both the patients' waiting times and doctor's idle time
- Larger the CV, the smaller the optimal appointment interval
- Performance of the system is very sensitive to small changes in the appointment intervals (Bailey 1952)
- Doctor's unpunctuality is lateness to first appointment, which has a big effect on patient waiting times as delay factor builds up
- When clinics are run under more credible appointment systems, patients and doctors become more punctual
- In almost all studies, patients are served on first come first served basis, which may destroy the whole purpose of the appointment system
- First priority is given to emergencies, then to scheduled patients and finally walk-ins
- Objective is minimising the expected total cost of the system
- Assumes a linear relationship between waiting cost and waiting time, which is not true in practice
- Modelling unpunctual patients and walks-in means the assumption of homogeneous waiting times might need to be relaxed
- Might be a threshold over which patients' tolerance declines steeply
- Cost of doctor's idle time is annual doctor's salary divided by hours worked (need to also include idle facility)
- Cost of patient's idle time is minimum wage (might be better to use median wage)
- True patient waiting time is difference between consultation start time and max of appointment time and arrival time (truncated to zero)
- Overtime is positive difference between expected and actual completion time of doctor
- Congestion leads to longer waiting times and (possibly) reduced service times
- Some studies include 'fairness' (i.e., uniformity of perfomance of an appointment system

- 7 major appointment rules (of varying sophistication) have been explored in the literature
- Dome shape: optimal appointment intervals exhibit a common pattern where they initially increase towards the middle of a session and then decrease
- As scheduling requests are handled dynamically, the use of patient classification is somewhat limited
- For surgical scheduling, the scheduler has a complete list of all requests for the day and patient availability is guaranteed
- Having predefined slots of patient classes leads to an inflexible appointment system
- Not entirely possible to eliminate no shows (Barron 1980)
- Add extra patients to make up for anticipated no shows
- Walk-ins and no shows won't necessarily cancel each other out
- Hospitals are rarely able to reschedule patients in case of an emergency
- Earlier queuing models assume steady state behaviour, which is never reached in a real clinic environment
- Optimal order of two procedures is increasing variance when service times are uniform or exponential
- Simulation modelling can model more complex patient flows
- Bailey's rule: an initial block of two patients and fixed intervals
- Ho and Lau (1992, 1999) and Ho, Lau and Li (1995) evaluate 50 appointment rules under various operating environments
- No rule will perform well under all circumstances
- Rohleder and Klassen (2000) consider the possibility that the scheduler can make an error when classifying patients
- Possibily could look at accidentally double booking two patients
- More fair if explictly try to increase appointment intervals towards the end of the day (Yang et al. 1998)
- Simulation research fails to report statistical significance of results
- Case studies have major drawback of lack of generalisation
- Rockart and Hofmann (1969) show individual block systems lead to more punctual doctors and patients and less no-shows
- The more time patients have to wait for an appointment, the greater the percentage of no-shows
- Simple grouping of inpatients and outpatients results in a substantial improvement of doctor's idle time (Walter 1973)
- O'Keefe (1985) had their proposed appointment system of classifying return and new patients rejected by staff who prefer a more simple and uniform appointment system
- The appointment scheduling problem is to a large extent a 'political' one as doctors are often unwilling to change their old habits
- Babes and Sarma (1991) initially tried to apply steady-state queuing theory, but found their results tended to be very different than those observed in real operation

- Huarng and Lee (1996) cdoul not implement an individual based appointment system because of staff resistance
- Bennett and Worthington (1998) found their recommendations weren't implemented successfully due to lack of control of doctor's behaviour
- It is often not practical to come up with estimates of individual consultation times
- Despite much published work, the impact on outpatient clinics has been limited
- Future research should attempt to close the gap between theory and practice
- Need to develop easy-to-use heuristics to choose the best system for individual clinics
- Empirical evidence can show exponential service distribution assumption is too restrictive
- Model walk-in seasonality
- Important to try to include a 'fairness' measure
- Lack of emphasis on real life performance of an appointment system (e..g, ease of use, physician's behaviour)
- Arguable today if doctor's time is more valuable than patients

Scheduling Arrivals to Queues: A Model with No-Shows

Author:	Sharon Mendel
Published:	July 2006

- Interested in finding the optimal schedule in an appointment system
- n arrivals of indpendent customers to a single server system with exponential service times
- Include probability of customer not showing up
- Minimise the sum of expected customers' and server's costs
- Reduce the waiting time for patients and increase service utilisation
- Control customers' waiting times and server's idle time
- Include tardiness in schedule in objective function
- Base point is analytical model presented by Pegden and Rosenshine (1990)
- Customer can be scheduled to arrive at same time as preceeding customer
- As showing up probability decreases, the optimal appointment intervals decreases and expected customer's waiting times decreases
- Based on Pegden and Rosenshine (1990) and Stein and Cote
- Server provides a Markovian service
- Determine t_1 (time of first arrival) and vector $x^* = (x_1, \ldots, x_{n-1})$ of intervals between scheduled arrivals
- w_i is expected waiting time of i-th customer in queue
- Minimise $\Phi(x^*) = c_w \sum_{i=1}^{n} w_i + c_s \left[t_1 + \sum_{i=1}^{n-1} x_i + w_n + \frac{1}{\mu} \right]$

- Obvious that $t_1 = 0$ (thus $w_1 = 0$) in optimal solution so can simply write $\phi(x^*) = c_w \sum_{i=2}^{n-1} w_i + c_s \sum_{i=1}^{n-1} x_i + (c_s + c_w)w_n$
- Pedgen and Roshenine develop a recursive algorithm to solve it (need to assume convexity to get global optimal)
- Stein and Cote use a D/M/1 queue model (assuming steady state)
- S(n,p)/M/1 is a system with n independent customers who each have probability p of showing up, and are serviced by a single Markovian server
- Can write the expected waiting time of the i-th customer (if they show up) as a function of x^*
- The probability of k departures from the queue in any interval is poisson
- For zero no shows, get same solution as Pegden and Roshenshine
- Need to constrain solution by $x_i \geq 0$
- Can normalise (without loss of generality) to $\mu = 1$ (assuming homogeneous)
- No closed form solution, so need to solve numerically using Newton-Raphson (by sequential quadratic programming using Matlab optimisation toolbox)
- Pegden and Roshenshine state the objective function is believe to be convex despite the presense of non-convex terms
- Interval increases for first few customers, then stays almost constant before decreasing (dome shape)
- As the probability of no shows increases, first few customers arrive together and last few customers arrive together
- Bailey recommended the first few customers arriving together
- No-shows lead to lower expected waiting time for customers who do show up
- Maximum expected waiting time of customer who shows up is up to twice the average expected waiting time and more than five times the average service time
- Cost of no-shows is relative cost of system where all customers show up to one with no-shows but the same expected number of customers who show up
- As the probability of no-shows increases the cost of no-shows increases
- Equally spaced model is a realistic restriction as more easily implementable
- Equations still do not have a closed form solution so use Newton-Rhapson approximation in Matlab
- Forcing equally spaced intervals does not materially change the value of the objective function
- Possibly should also look at fixed intervals but a variable number (≥ 0) scheduled to arrive at the start of each interval
- Theorem 1.1 of Hajek: average customer waiting time is at a minimum for constant inter-arrival times (if one exponential server)
- Does not always hold in a system with no-shows due to other factors in objective function

Scheduling Arrivals to Queues

Author:	Claude Dennis Pegden, Matthew Rosenshine
Published:	1990

- Arrival process consists of n arrivals whose arrivals times may be scheduled
- Algorithm is developed to determine the schedule for n arrivals
- Objective function is convex for $n \leq 4$ and conjectured to be convex for $n \geq 5$
- Appointment system in which an arrival cannot be refused an appointment
- If customers are scheduled to arrive early, then the total time the server must be available is reduced, but long waiting times for customers
- If arrivals are spread out, customer waiting time is reduced at the expense of a large required server availability time
- How to incorporate a 'last' possible scheduled time? (truncate the times?)
- Minimise the total system cost
- Kendall's notation with n scheduled customers: $S(n)/M/1$ implies a Markovian service with one server
- $N(t_i)$ is the number of customers in the system just prior to the time of i-th arrival
- x_i is the time interval between the scheduled arrival times of the i-th and the $(i+1)$-th arrival
- Objective is to determine t_1 and $\overline{x} = (x_1, \ldots, x_n)$
- Optimisation problem is $\min\limits_{t_1, \overline{x}} z = c_w \sum\limits_{i=1}^{n} w_i + c_s \left[t_1 + \sum\limits_{i=1}^{n-1} x_i + w_n + \dfrac{1}{\mu} \right]$
- Take $t_1 = 0$ and ignore constants to get simpler expression
- If service time is deterministic, then finding the optimal solution with zero waiting time and idle time is trivial
- Can find closed form solution for $n = 2$
- Due to memoryless property of exponential distribution, the expected waiting time of the second customer is $w_2 = \dfrac{\mathbb{P}(N(t_2) = 1)}{\mu} = \dfrac{\exp(-\mu x_1)}{\mu}$
- Then solve $\dfrac{dz}{dx_1} = 0$ to get $x_1 = \dfrac{-1}{\mu} \ln \left(\dfrac{c_s}{c_s + c_w} \right)$
- As $\dfrac{d^2 z}{dx^2} > 0$, z is convex and have global minimum
- For $n = 3$, can find closed form for x_2 but not x_1
- For $n \geq 3$, can't find closed form for $\overline{x}$
- Can set $\mu = 1$ such that solutions found for x_i are actually solutions for μx_i
- In general, expected waiting time is $w_i = \sum\limits_{j=1}^{i-1} \left(\dfrac{j}{\mu} \right) \mathbb{P}(N(t_i) = j)$

- In general, for $j > 0$, $\mathbb{P}(N(t_i) = j) = \sum_{k=0}^{i-j-1} \mathbb{P}(N(t_{i-1}) = j + k - 1)$ $\times\mathbb{P}(k$ departures between the $(i-1)$-th arrival and the i-th arrival$)$
- Probability of k departures is $\frac{(\mu x_{i-1})^k}{k!} \exp(-\mu x_{i-1})$ as a Poisson process
- Gives an algorithm for computing the value of the objective function
- If the objective function is convex, algorithm will find a global minimum
- Extensions to include batch arrivals, nonexponential services and balking

Scheduling Arrivals to Queues

Author:	William E. Stein, Murray J. Côte
Published:	1994

- Two competing interests must be considered
- Determine the optimal solution for equally spaced arrivals
- Large n approximation using steady-state
- Extend Pegden and Rosenshine by obtaining numerical results for situations with more than three customers
- Optimal time between successive customers becomes nearly constant as n grows
- c_w is the cost of waiting time and c_s is the cost of server time
- Mean cost of waiting is $c_w \sum_{i=1}^{n} w_i = c_w \sum_{i=2}^{n} w_n$ as $w_1 = 0$
- Mean cost of server is $c_s \left[\sum_{i=1}^{n-1} x_i + w_n + \frac{1}{\mu} \right]$
- Mean cost of idle only differs by a constant: $c_s \left[\sum_{i=1}^{n-1} x_i + w_n - \frac{n-1}{\mu} \right]$
- $\gamma = \frac{c_s}{c_s + c_w}$ is the relative importance of the two costs
- Formulas for expected waiting times can be stated more succintly using transition matrices
- Use a reduced-gradient algorithm to solve numerically
- Set $\mu = 1$ without loss of generality
- Optimal interval width does not increase as i increases
- Last few customers are scheduled to arrive sooner to avoid keeping the server idle
- A simplification of the model can be made by requiring equally spaced intervals: $x_1 = \ldots = x_{n-1} = x$
- Realistic restriction to scheduling problem
- Theorem 1.1 of Hajek: average waiting time of a customer will be at a minimum for constant interarrival times (doesn't hold here due to cost of server, but will hold with $\gamma = 0$)

- Constant interarrival times leads to a deterministic D/M/1 arrival queue with an initial size of 0
- Optimal interval is almost an average of the unequally spaced widths
- Assume queue can be described by its steady state distribution (doesn't hold in reality)
- Can solve numerically to find global minimum (e.g., Newton's method)
- Similar results between non-steady and steady state when $\gamma < 0.4$
- x converges as n increases
- Have an upper bound for the value of x, which provides a good estimate when $\gamma \geq 0.3$
- Steady state provides good approximation to x for large n
- Equally spaced intervals reduce the computational effort
- Steady state solution is independent of n