Optimal Mechanisms in Strategic Queues

Marco Scarsini[1] Eran Shmaya[2]

[1]Luiss University

[2]Stony Brook University

LNMB, January 2025


Posteitaliane

Classical queueing models

- The traditional modeling of queues is purely stochastic:
  - Customers enter a queue at random times.
  - According to a specified regime, they spend a random waiting time in the queue.
  - Then, they start getting served.
  - After a random service time, they get served and leave the queue.
- Even when balking or reneging are contemplated, they happen at random.
- In these models no decision is contemplated or analyzed.

M/M/1

- M/M/1 is the simplest queueing model.
- In this model customers arrive according to a homogeneous Poisson process with parameter λ.
- Service times are i.i.d. random variables having an exponential distribution with parameter μ.
- The random time each customer spends in the queue depends on the queueing regime.
- Some examples of regimes are first-come-first-served, last-come-first-served, random order, etc.
- An M/M/1 queue is stable iff $\lambda < \mu$.

Strategic queueing models

- In his seminal paper, Naor (1969) considered an M/M/1 queueing model where customers strategically decide whether or not to enter a queue and when to renege, possibly never.
- They make their decision based on the parameters of the model (service rate, waiting cost, and reward for being served).
- Naor showed that, in a first-come-first-served regime, customers' selfish behavior produces an outcome that is socially suboptimal.
- This is due to the externalities that a customer's behavior produces for other customers who subsequently join the queue.
- Selfish customers tend to join the queue more often than the socially optimum behavior would recommend.

Social planner

- A social planner could achieve optimality either by enforcing a cap on the queue length or by imposing a toll to join it.
- Both these choices require an exact knowledge of the model's parameters.
- Hassin (1985) showed that optimality could be achieved also by adopting a different queuing regime.
- If a new arriving customer is immediately served, preempting the currently served customer, optimality is achieved.
- This regime is usually termed last-come-first-served with preemption.
- This regime achieves optimality in a universal sense, that is, for any possible parameters of the model.

Universally optimal regimes

- There exist other universally optimal regimes.
- For instance, optimality is achieved by any regime where a new arriving customer is put in a position that is not the last.
- We want to characterize universally optimal regimes.

The model

- M/M/1 queuing system:
- Customers arrive according to a Poisson process with rate λ and are served with rate μ.
- Each customer incurs a flow cost rate c while in the system, and receives a reward r upon service completion.
- The queue is governed by a fixed regime.
- The queue is observable.
- When customers arrive at the queue, they can either join it or balk.
- At any time a customer in the queue can renege.
- A customer who either balked or reneged cannot rejoin the queue at a later time.

Equilibrium

- If $r < c/\mu$, then no customer will ever join the queue.
- Consider a customer who arrives at a queue with n customers.
- This customer's expected payoff is

  $$r - \frac{c}{\mu}(n+1),$$

  if they join the queue, and 0, if they balk.
- There exists a value n^e such that

  $$r - \frac{c}{\mu}(n^e) \geq 0 \quad \text{and} \quad r - \frac{c}{\mu}(n^e + 1) < 0.$$

- The optimal strategy for this customer is to join the queue if and only if $n \leq n^e$.
- It is never optimal for a customer to renege.

Social optimum

- The designer incurs a flow cost rate c per customer in the system and receives a reward r upon each customer's service completion.
- The social designer can decide which arriving customers to accept, and when to kick existing customers out of the system.
- The social designer cares about the total welfare of the customers in the long run, but has no other considerations, i.e., the designer does not care about the identity of the customer who is being served.
- The socially optimal strategy would require each customer to join if and only if its size is not larger than some value n^*.
- Naor showed that $n^* \leq n^e$ and, for some values of the parameters, the inequality is strict.

Queuing regimes

- A queuing regime is given by a tuple $(\mathcal{X}, \alpha, \xi, (\rho_i)_i, \pi)$, where
  - $\mathcal{X}$ is a set of states,
  - α, ξ, ρ_i are transition functions,
  - π is a position function.
- The set of states can be partitioned as $\mathcal{X} = \mathcal{X}_0 \uplus \mathcal{X}_1 \uplus \dots$, where, for every $n \in \mathbb{N}$, $\mathcal{X}_n$ is the set of possible states when there are n customers in the system, and $\uplus$ is the disjoint union.
- $\mathcal{X}_0$ is a singleton, representing the idle system.
- For $x \in \mathcal{X}_n$ we define $n(x) = n$.

Queuing regimes

- At every point in time the customers who are currently in the system are ranked according to some order, called queue, the order in which they will be served if no new customer joins and nobody reneges.
- The regime is assumed to be work-conserving, that is, one customer is always being served if the system is not idle.
- The customer who is currently being served has position 1, and the last customer has position n in the queue.
- The system transitions from one state to another when either a new customer arrives, or a customer is served, or a group of customers (possibly only one) reneges.
- Arrivals and service are random and controlled by Nature, whereas reneging is a decision made by the customer.
- We assume that none of these events changes the relative order among the existing customers in the system.

Queueing regimes

- Let $[n] := \{1, \dots, n\}$.
- The transition rules of the system and the position of new customers in the queue are governed by the the transition functions ρ_i, ξ, α, and the position function π as follows:
  - If the system is at state $x \in \mathcal{X}_n$ and a new customer arrives, the system transitions to state $\alpha(x) \in \mathcal{X}_{n+1}$ and the arriving customer is placed at position $\pi(x) \in [n+1]$ in the queue.
  - If the system is at state $x \in \mathcal{X}_n$ with $n \geq 1$ and the customer who is being served completes service, the system transitions to state $\xi(x) \in \mathcal{X}_{n-1}$.
  - If the system is at state $x \in \mathcal{X}_n$ and the customer whose current position is $i \in [n]$ reneges, the system transitions to state $\rho_i(x) \in \mathcal{X}_{n-1}$.
- For $I = (i_1 < i_2 < \cdots < i_k)$, we let $\rho_I := \rho_{i_1} \circ \rho_{i_2} \circ \cdots \circ \rho_{i_k}$.

Examples

Example (First-come-first-served)

In the *first-come-first-served* (FCFS) regime the state only encodes the number of customers in the system, so $\mathcal{X}_n$ is a singleton for every n; hence $\mathcal{X} = \mathbb{N}$.
The transition functions are:

$$\alpha(n) = n + 1, \quad \xi(n) = n - 1, \quad \rho_i(n) = n - 1.$$

The position function is $\pi(n) = n + 1$.

Examples

Example (Last-come-first-served)

The *last-come-first-served* (LCFS) has the same state space and transition functions of FCFS.

$$\alpha(n) = n + 1, \quad \xi(n) = n - 1, \quad \rho_i(n) = n - 1.$$

In the *last-come-first-served with preemption* (LCFS-PR) regime $\pi(x) = 1$ for every state x.
In the LCFS without preemption $\pi(x) = \min(2, n(x) + 1)$ for every state x.

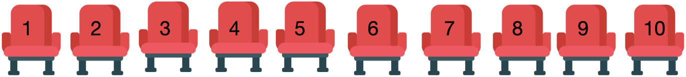
Examples

Example (Priority-slots, Wang (2016))

In the *priority-slots* (PS) regime there is a countable set $\mathbb{N}$ of slots and the state space is given by the set of occupied slots, so an element of $\mathcal{X}_n$ is a subset of $\mathbb{N}$ of cardinality n.
If $x = \{x_1, \ldots, x_n\} \in \mathcal{X}_n$ with $x_1 < \cdots < x_n$, then

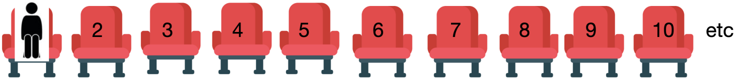
$$\begin{aligned}
\alpha(n) &= x \cup \{\min(\mathbb{N} \setminus x)\}, \\
\pi(n) &= \min(\mathbb{N} \setminus x), \\
\xi(x) &= x \setminus \{x_1\}, \\
\rho_i(x) &= x \setminus \{x_i\}.
\end{aligned}$$

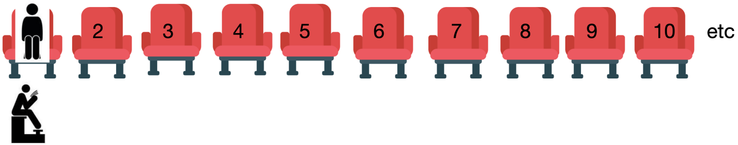

#STANDFORSOMETHING

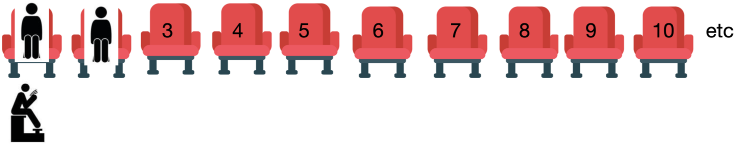


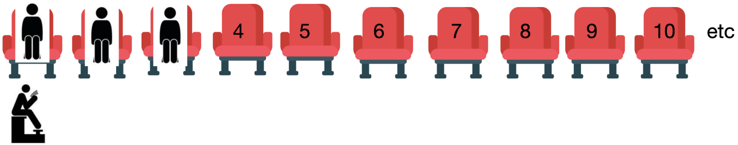
etc

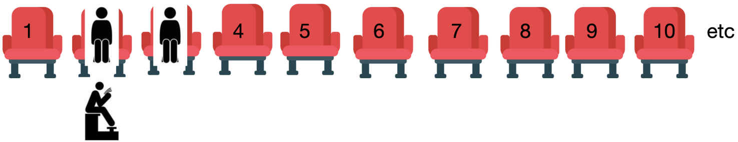
Idle state

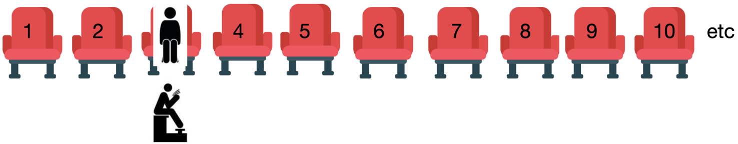

2
3
4
5
6
7
8
9
10
etc


2
3
4
5
6
7
8
9
10
etc

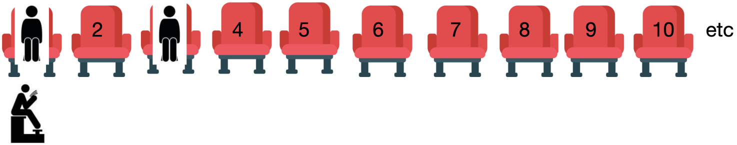

3
4
5
6
7
8
9
10
etc


4
5
6
7
8
9
10
etc


1
4
5
6
7
8
9
10
etc


1
2
4
5
6
7
8
9
10
etc

Preemption



Strategies and equilibrium

- A Markov strategy profile is a function σ defined over non-idle states such that $\sigma(x) \subset [n]$ for every $x \in \mathcal{X}_n$, with the interpretation that $\sigma(x)$ are the positions of players that abandon at state x.
- We assume that abandoning happens simultaneously whenever the system reaches this state.
- Naor proved that the social optimum is achieved by a strategy profile σ such that $|\sigma(x)| = (n - n^*(\lambda, \mu, c, r))_+$.
- A Markov strategy profile is a Markov perfect equilibrium if, for every state x, it is a Nash equilibrium in the game that starts at state x, in which players can decide whether to stay in the queue or abandon it.

Universally optimal regimes

- A regime is universally optimal if, for every environment (λ, μ, c, r), the game admits a Markov perfect equilibrium that induces the socially optimal behavior.
- Our goal is to characterize the class of universally optimal regimes.

Maximal states

- A state x is maximal if it satisfies the following property:
- If $x_0, x_1, \ldots, x_k = x$ is a sequence of non-idle states such that for every $1 \leq j \leq k$ either $x_j = \alpha(x_{j-1})$ or $x_j = \xi(x_{j-1})$, then $n(x_0) \leq n(x)$.
- A maximal state is a state that cannot be reached by arrival and service from a state with a larger number of customers without going through an idle state.

Examples

Example (First-come-first-served)

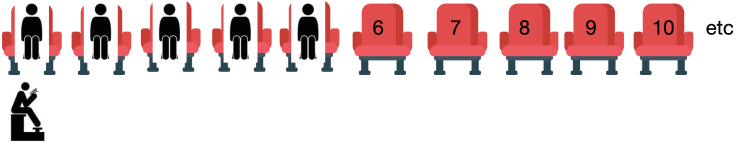
- The state is the number of customers in the system.
- There are no maximal states.
- Indeed, it is possible to have n customers in the system now and to have had $n + 1$ customers in the past.

Examples

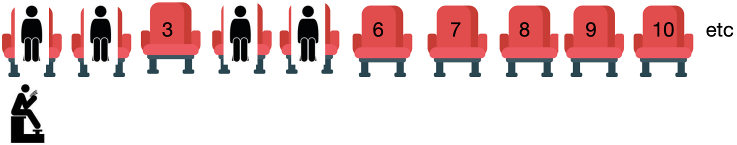
Example (Priority slots)

- A state is given by the set of occupied slots.
- A state $x \in \mathcal{X}_n$ is maximal if and only if $x = [n]$, that is, the slots that are occupied are exactly $1, \ldots, n$.

Maximal state



Non-maximal state



Characterization

Theorem

The following two conditions are equivalent for a queuing regime:

(a) *The regime is universally optimal.*

(b) *For every state x that is not maximal, we have*

$$\pi(x) < n(x) + 1. \tag{1}$$

- Hassin (1985) proved that if, for every state x, condition (1) holds, then the regime is universally optimal.
- On the other hand, there exist universally optimal regimes, such as the priority slots, that do not satisfy this property.

Preemption

We say that preemption occurs at a non-idle state x if $\pi(x) = 1$.

Corollary

If a regime is universally optimal, then preemption occurs at some non-idle state.

- The stochastic properties of the M/M/1 queue guarantee that a social planner does not need to use preemption to achieve the social optimum.
- Replacing the customer being served with another one does change the expected performance of the regime.
- The role of preemption is purely strategic, in the sense that it affects the customers' equilibrium behavior.


QUEUE HERE
SIGH HEAVILY FROM HERE
TUT LOUDLY FROM HERE
Royston
roystonrobertson.co.uk