Optimal Queue Design

Yeon-Koo Che\textsuperscript{1} and Olivier Tercieux\textsuperscript{2}

\textsuperscript{1}Columbia University

\textsuperscript{2}Paris School of Economics

October 21, 2021
Introduction

- Much of goods and services are allocated through non-market means, or “rationed,” by far most commonly through “waiting in line.”

- But waiting is costly and painful:
  - 6 months of life waiting in line for things (e.g., schools, hospitals, bookstores, libraries, banks, post office, petrol pumps, theatres...)
  - 43 days on hold with call centers (Brown et al. 2005)

- **Challenge:** How to efficiently provide incentives for waiting in line?

- Queue designer has several instruments at disposal (imagine call centers): entry and exit, queueing rule, the information policy.

- Existing queueing theory, in particular, *rational queueing* (e.g., Hassin (2016)), treats this issue, but in limited scopes both in terms of the design variables and agents’ incentives.
What We Do

- We take a Myerson-style mechanism design approach to the design of the optimal queueing system,

- allowing all three aspects of queue design—entry/exit, queueing rule, and the information policy—chosen optimally

- while taking into consideration incentives for agents to join a queue and to stay in the queue, whenever necessary.

- We show: Under a mild condition, the optimal policy is implemented by first-come-first-served with the policy of no information given to the agents about the queue.
A queueing model with general Markov process

- Continuous time.
- **Primitive process:** At each instant,
  - an agent arrives (randomly) at a Poisson rate $\lambda_k > 0$
  - and is served (randomly) at a Poisson rate $\mu_k > 0$ (\(=\) average service time is \(1/\mu_k\)),

  when there are \(k\) agents in the queue.
- Dependence on \(k\) allows for extra generality: See next slides.
- Assume **Regularity:**
  1. \(\mu_k\) is nondecreasing and concave in \(k\)
  2. \(\lambda_{k+1} - \lambda_k \leq \mu_{k+1} - \mu_k\), \(\forall k\)

  A mild assumption satisfied in virtually all realistic environments.
Examples:

- M/M/1: $\lambda_k, \mu_k$ do not depend on $k$
- M/M/c: $\lambda_k$ does not depend on $k$ and $\mu_k = \min\{k, c\}\mu$
- Dynamic matching with stochastic compatibility
  - effective arrival rate $= \text{arrival rate} \times \text{prob of not compatible with anybody in the queue (depends on } k)$
  - effective exit rate $= \text{arrival rate} \times \text{prob of somebody in the queue being compatible (depends on } k)$
Preferences

Standard queueing model: homogeneous preferences with linear waiting costs.

- **Agents’ payoffs:**
  \[ U(t) = V - C \cdot t, \]
  where \( t \) the time spent in the system.
  - \( V > 0 \): net surplus from service
  - \( C > 0 \): per-period cost of waiting
  - zero outside option.

- **The firm** receives \( R > 0 \) from each agent served

- **Designer’s objective.** Weighted sum of firm’s and agents’ payoffs.
Queueing Mechanism

- **Entry rule:** $x = (x_k)$, where $x_k$ is prob of recommending entry in a queue of length $k$

  "Please hold; somebody will be with you" or "... please come back some other time; good bye."

- **Exit rule:** $y = (y_{k,\ell})$, where $y_{k,\ell}$ is the rate of removal when queue length is $k$ and position is $\ell$; we also allow for a "lumpy" exit upon a new entry (omitted here)

  "We are experiencing unusual call volume, please come back later"
Queueing Mechanism—Continued

- **Queueing rule:** $q = (q_{k,\ell})$ where $q_{k,\ell}$ the service rate when queue length is $k$ and position is $\ell$, such that
  - **Feasibility:** For any set $S \subset \{1, \ldots, k\}$ of $|S| = m$ agents:
    \[
    \sum_{j \in S} q_{k,j} \leq \mu m
    \]
  - **Work-conservation:**
    \[
    \sum_{\ell=1}^{k} q_{k,\ell} = \mu k
    \]

- **Examples:**
  - **First-Come First-Served (FCFS):** $q_{k,1} = \mu_1$, $q_{k,2} = \mu_2 - \mu_1$, $q_{k,\ell} = \mu_{\ell} - \mu_{\ell-1}$. ($M/M/1$, $q_{k,\ell} = \mu$ if $\ell = 1$ and 0 o/wise)
  - **Last-Come First-Served (LCFS):** $q_{k,k} = \mu_1$, $q_{k,k-1} = \mu_2 - \mu_1$, $q_{k,\ell} = \mu_{k-\ell+1} - \mu_{k-\ell}$ ($M/M/1$, $q_{k,\ell} = \mu$ if $\ell = k$ and 0 o/wise)
  - **Service-In-Random-Order (SIRO):** $q_{k,\ell} = \mu_k / k$
Queueing Mechanism—Continued

- **Information rule:** \( I = (I_t) \), where \( I_t \) specifies the information an agent gets about the state—i.e., the queue length \( k \) and her position \( \ell \)—for each time \( t \geq 0 \) spent on the queue.

- **Examples:**
  - Full information
  - No information (beyond recommendations)
Overview

- The entry/exit rules \((x, y)\), together with \((\lambda, \mu)\), induces a Markov chain on the queue length \(k\) with an invariant distribution \(p = (p_k) \in \Delta(\mathbb{Z}_+)\).

- We focus on the problem at steady state, or invariant distribution:

  “Maximize designer objective (at the invariant dist)
  
  subject to: Agents have incentives to join and stay whenever needed.”

- Why IC? Agents can be denied entry or removed without consent, but they cannot be coerced to join the queue or staying in it against their will.
Related Literature

Queueing Design with fixed information rule:
- Naor (1969), Hassin (1985), Su and Zenios (2004): Excessive incentives for queueing under FCFS, corrected by LCFS
- Leshno (2019): Insufficient incentives for queueing under FCFS, corrected by SIRO or LIEW
- Bloch and Cantala (2017), Margaria (2020),...
- Ashlagi, Faidra, and Nikzad (2020)

Information Design with fixed queueing rules:

Current work distinguished by:
- the generality of the primitive Markov process and designer objective
- the comprehensiveness of mechanism design approach
- the consideration of dynamic incentive issues
Designer’s problem

Designer chooses \((x, y, q, I)\) to solve:

\[
\begin{align*}
[P] & \quad \text{Maximize designer objective at } p, \\
(B) & \quad p \text{ is an invariant distr given by } (x, y) \\
(IC) & \quad \text{incentives to join or stay when recommended}
\end{align*}
\]
Designer’s problem

Designer chooses \((x, y, q, l)\) to solve:

\[
[P] \quad \text{Maximize } (1 - \alpha) \sum_{k=1}^{\infty} p_k \mu_k R + \alpha \sum_{k=1}^{\infty} p_k (\mu_k V - kC),
\]

subject to

\[(B) \quad \lambda_k x_k p_k = (\mu_{k+1} + \sum_{\ell} y_{k+1,\ell}) p_{k+1}, \quad \forall k\]

and

\[(IC) \quad \text{Incentive constraints for every signal at each time } t\]

**Remark:** Difficult to solve.
A relaxed LP problem

The designer chooses (only!) $p$

$$[P'] \text{ Maximize } (1 - \alpha) \sum_{k=1}^{\infty} p_k \mu_k R + \alpha \sum_{k=1}^{\infty} p_k (\mu_k V - kC),$$

subject to relaxation of balance equation:

$$(B') \quad \lambda_k p_k - \mu_{k+1} p_{k+1} \geq 0$$

subject to relaxed incentive compatibility:

$$(IR) \quad \sum_{k=1}^{\infty} p_k (\mu_k V - kC) \geq 0.$$ 

Remark: $(IR)$ equivalent to “agents having incentives to join under no information.”
Theorem

If $\mu$ is regular, then an optimal solution of relaxed program $[P']$ is a cutoff policy, meaning there exists $K^* \geq 0$ such that agents are allowed to queue up to $K^*$.

Note: Random rationing possible at $K^* - 1$.

Implication: No need for removing agents.
Optimality of FCFS with no information

**Theorem**

Assume the primitive process is regular. **FCFS + no information** (i.e., beyond that inferred by recommendation) is optimal.

- Can implement the cutoff policy that solves the relaxed program \([P']\) with FCFS + No information.

**Proof:**

1. Incentives to join the queue: Holds since (IR) is satisfied at the solution.
2. Incentives to stay in the queue until served: non-trivial.

We show: Under regularity, beliefs about queue position improve over time

⇒ Residual waiting time falls.
Intuition

Question: Is “time spent in the queue” good news or bad news?

- Good news: *conditional on the initial queue length*, under FCFS, position in queue can only improve

- Bad news: “the initial queue length may have been longer” ⇒ pessimistic updating

We show that, given the regularity of the primitive process, good news dominates bad news.
Belief about position $\ell = 1$

\[
\frac{1}{27}
\]

\[
M/M/1 \text{ with } K^* = 2; \lambda = \mu = 1.
\]
Belief about position $\ell = 1$

$M/M/1$ with $K^* = 2$; $\lambda = \mu = 1$. 
Evolution of beliefs under FCFS with no information

- $\gamma^t_\ell$ = belief that position is $\ell$ after spending time $t \geq 0$ on the queue.

- Consider likelihood ratios: $r^t_\ell \triangleq \frac{\gamma^t_\ell}{\gamma^t_{\ell-1}}$, for all $\ell = 2, \ldots, K^*$.

- Suffices to show: the likelihood ratios $(r^t_\ell)_\ell$ fall in $t$.
  - ⇒ Beliefs about queue position improve over time
  - ⇒ Residual waiting time falls.
Do the likelihood ratios fall?

- How does \((r^t_\ell)\) evolve?

\[
\begin{align*}
    r^{t+dt}_\ell &= \frac{\gamma^{t+dt}_\ell}{\gamma^{t+dt}_{\ell-1}} \\
    &= \frac{(1 - \mu_\ell dt)\gamma^t_\ell + (\mu_\ell dt)\gamma^{t+1}_{\ell+1}}{(1 - \mu_{\ell-1} dt)\gamma^t_{\ell-1} + (\mu_{\ell-1} dt)\gamma^t_\ell} + o(dt).
\end{align*}
\]

\[\Rightarrow\] System of ODEs of the likelihood ratios:

\[
\dot{r}^t_\ell = r^t_\ell \left( -(\mu_\ell - \mu_{\ell-1}) + (\mu_\ell r^{t+1}_\ell + \mu_{\ell-1} r^t_\ell) \right)
\]

- Generally ambiguous. The “initial beliefs” matter!
Evolution of beliefs under FCFS with no information

- The likelihood ratios at $t = 0$ given by the invariant distr. (cf. PASTA):
  $\forall \ell = 2, \ldots, K^*$,

\[
\dot{r}_\ell^0 = r_\ell^0 \left(- (\mu_\ell - \mu_{\ell-1}) + (\mu_\ell r_{\ell+1}^0 - \mu_{\ell-1} r_\ell^0)\right)
\]

\[
= r_\ell^0 \left(- (\mu_\ell - \mu_{\ell-1}) + (\mu_\ell \lambda_\ell \mu_\ell - \mu_{\ell-1} \lambda_{\ell-1} \mu_{\ell-1})\right)
\]

\[
= r_\ell^0 \left(- (\mu_\ell - \mu_{\ell-1}) + (\lambda_\ell - \lambda_{\ell-1})\right) \leq 0.
\]

- The system of ODEs is "cooperative":

\[
\dot{r}^0 \leq 0 \Rightarrow \dot{r}^t \leq 0 \text{ for all } t
\]
Necessity of FCFS for Optimality

In principle, other queueing rules or information rules may work under some environments. But FCFS with no information is uniquely optimal in a maximal domain sense.

Theorem

For any queueing rule differing from FCFS, there exists a queueing environment \((\lambda, \mu, V, C)\) under which the rule can’t implement the optimal policy regardless of the information policy.
Residual waiting time under alternative queueing rules.

\[ M/M/1 \text{ with } K^* = 2; \lambda = \mu = 1. \]
Concluding Thoughts

- Without info design, FCFS is typically suboptimal, and optimal policy is unknown and difficult to find.

- With information design, FCFS is (uniquely) optimal

- Of course, there may be unmodeled benefits of providing information on queue position or expected waiting times
  - intrinsic value of transparency
  - ambiguity aversion...

- We have identified a novel role for queueing disciplines in regulating agents’ beliefs, and their dynamic incentives

- Revealed a hitherto-unrecognized virtue of FCFS in this regard.
Thank You!
References


