Optimal Queue Design

Yeon-Koo Che[1] and Olivier Tercieux[2]

[1]Columbia University

[2]Paris School of Economics

May 24, 2021
Virtual Market Design Seminar

Introduction

- Waiting in line is very common in every-day-life.
  - 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)
- **Queueing theory** is the mathematical study of waiting lines or queues.
  - what determines queue lengths and
  - waiting time of agents in the queue
- Subfield *rational queueing (e.g., Hassin (2016))* studying agents' incentives to join a queue
  - Agents tradeoff waiting times to get served/matched with outside option
  - Equilibrium queue length vs. socially optimal compare
- Our goal: Systematic design treatment like Myerson for auction design.

A queueing model with general Markov process

- Continuous time $t \in [0, \infty)$
- Agents arrive randomly to a queue to receive service.
- At each instant, if there are k agents in the queue:
  - an agent arrives at a Poisson rate $\lambda_k > 0$
  - service occurs at a Poisson rate $\mu_k > 0$
  - a pair $(\lambda, \mu) = (\{\lambda_k\}, \{\mu_k\})$ is a primitive process
  - We assume μ_k is nondecreasing in k: *Without loss* since we interpret μ_k as the maximum service rate for an agent belong to a set of any k agents.

Examples:

- M/M/1 queueing model: λ_k, μ_k do not depend on k
- M/M/c queueing model: λ_k does not depend on k and $\mu_k = \min\{k, c\}\mu$,
- Dynamic matching model:
  - μ_k = rate of an arriving agent compatible with someone waiting (depends on the nb. of people in the queue)
  - λ_k = rate of an arriving agent *in*compatible with any agent waiting (depends on the nb. of people in the queue)

Assumption on the Primitive Process

We sometimes will assume (μ, λ) to be regular.

1. $\mu_k - \mu_{k-1}$ are nonincreasing in k;
2. $\lambda_k - \lambda_{k-1} \leq \mu_k - \mu_{k-1}$ for all $k \geq 2$

Mild assumption, all the above examples satisfy regularity.

Preferences

Standard queueing model: homogeneous preferences with linear waiting costs.

Individuals' payoffs: When receiving service after waiting $t \in R_+$, agents receive payoff:

$$U(t) = V - C \cdot t,$$

- $V > 0$ is the net surplus from service
- $C > 0$ is the per-period cost of waiting
- Outside option yields a normalized payoff of zero.

Service provider's payoffs. Earns $R > 0$ from each individual who gets served

Designer's objective. Weighted sum of provider's and individuals' payoffs.

Queueing Mechanism

- Entry rule: $x = (x_k)$, where x_k is prob of entry in a queue of length k ["Please hold; somebody will be with you shortly" or "We are experiencing unusual volume of calls, please come back some other time"]
- Exit rule: $y = (y_{k,\ell})$, where $y_{k,\ell}$ is the rate of removal when queue length is k and position is ℓ
  ["We are experiencing unusual call volume, please come back later"]
- Queueing rule: $q = (q_{k,\ell})$
- Information rule: $I = (I_t)$,

Queueing Rule

- $q_{k,\ell}$ the service rate when queue length is k and position is ℓ;
- Feasible queueing rules: For any set $S \subset \{1, ..., k\}$ of size J:

$$\sum_{j \in S} q_{k,j} \leq \mu_J$$

- Work-conserving queueing rules:

$$\sum_{\ell=1}^{k} q_{k,\ell} = \mu_k$$

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

Information Rule

- An information rule $I = \{I_t\}_{t \in \mathbb{R}_+}$, where I_t represents the information an agent has about the state (k, ℓ) after staying in the queue for $t \geq 0$

Special cases are

- Full information
- No information (beyond recommendations)

Overview

- The entry/exit rules (x, y), together with (λ, μ), induces a Markov chain on the queue length k.
- Let $p = (p_k) \in \Delta(\mathbb{Z}_+)$ be the invariant distribution
- We say that p **is generated by policy** (x, y)
- Designer maximizes objective at the inv dist
  - Subject to incentive constraints
  - I.e., incentives to join or stay in queue upon recommendations

Note 1. Prior beliefs of agents = inv. dist.

Note 2. Dynamic IC (i.e., to stay) often disregarded in queueing lit

Preview of the Results

- Optimal cutoff policy: Entry up to some $K \in \mathbb{Z}_+ \cup \{+\infty\}$ but no removal is optimal.

- No information is optimal.

- FCFS is optimal: can implement the optimal cutoff policy, provided that no information is given to agents.

- FCFS necessary for optimality in a rich domain: For any queueing discipline differing from FCFS, there exists a queueing problem (λ, μ, V, C) such that it is not optimal under any information design.

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:
  - Hassin and Koshman (2017), Lingenbrink and Iyer (2019), Anunrojwong, Iyer, and Manshadi (2020)

Our paper:

- General mechanism design approach with information and queue design;
- We consider dynamic incentives (i.e., incentives to stay in the queue);
- We consider a general primitive process not only M/M/1.

Designer's problem

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

$[P]$ Max weighted sum of agents' flow payoffs at the inv. dist. p,

subject to balance equation,

(B) p is generated by (x, y)

and subject to incentive constraints, i.e.,

(IC) Recommended to join or stay $\Rightarrow$ incentives to do so

Designer's problem

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

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

subject to balance equation,

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

and subject to incentive constraints, i.e.,

(IC) Incentive constraints for every signal at each time t

Remark: Difficult to solve.

A relaxed LP problem

The designer chooses (only!) p

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

subject to relaxed balance equation,

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

subject to relaxed incentive compatibility,

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

(IR): Aggregating (IC) at $t=0$ across beliefs $\gamma^0 \in \text{supp}(I_0)$ "$=$" (IR)
$\Leftrightarrow$ (IC) at $t=0$ with no information

Optimality of Cutoff Policy

Definition

A **cutoff policy** is a pair (x, y) where $y \equiv 0$ and $x_k = 1$ for $k \leq K^* - 2$ and $x_k = 0$ for all $k \geq K^*$, for some $K^* \in \mathbb{Z}_+ \cup \{+\infty\}$.

Theorem

Assume the primitive process is regular. An optimal solution p^* of $[P']$ can be generated by a cutoff policy.

Note: No need for removal. Random rationing possible for $k = K^* - 1$.

Optimality of FCFS with no information

- Fix a cutoff policy (x^*, y^*) generating p^* a solution to $[P']$
- Let $q^* =$ FCFS and $I^* =$ "no information"

Theorem

Assume the primitive process is regular. (x^*, y^*, q^*, I^*) *is an optimal solution to* $[P]$*—the designer's exact problem.*

Argue in two steps.

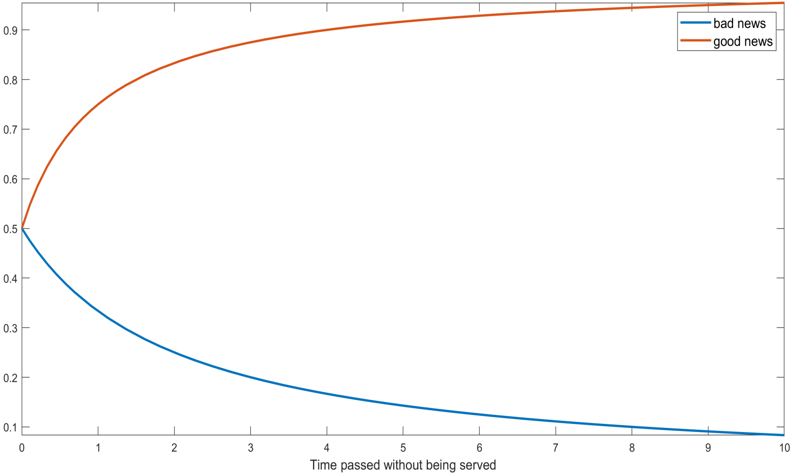
1. Show that (IC) holds at $t = 0$; Holds since (IR) is satisfied at p^*
2. Show that (IC) holds at $t > 0$. Need to study dynamic evolution of beliefs.

Optimality of FCFS with no information

- Question: Is "the elapse of time without getting served" good news or bad news?
  - Good news: *conditional on the initial queue length*, under FCFS, position in queue can only improve (i.e., likely that agents ahead of me got served)
  - Bad news: reveals that the initial queue length may have been longer, yielding a pessimistic updating about one's position

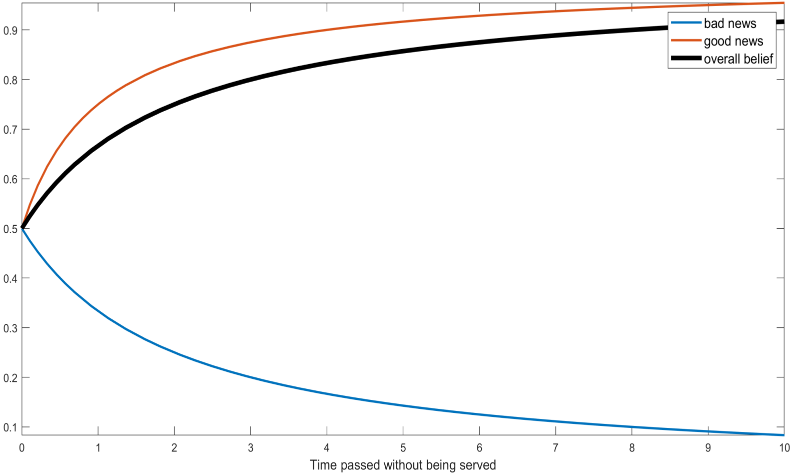
We show that the regularity of the primitive process ensures that good news dominates bad news.

Belief about position $\ell = 1$



M/M/1 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

- Let γ_ℓ^t belief that position is ℓ after spending time $t \geq 0$ in the queue.
- We focus on the likelihood ratio of beliefs over positions:

$$r_\ell^t \triangleq \frac{\gamma_\ell^t}{\gamma_{\ell-1}^t}$$

  where $\ell = 2, ..., K^*$.

- $r^t := (r_\ell^t)_\ell$ forms a system of ODEs
  (existence and uniqueness is shown)
- We show: Under regularity, starting from initial beliefs, r^t decreases in time t

  $\Rightarrow$ Beliefs about queue position improve over time

  $\Rightarrow$ Residual waiting time falls.

Evolution of beliefs under FCFS with no information

- We show: Under regularity, given r^0, r^t decreases in time t

  $\Rightarrow$ Beliefs about queue position improve over time

  $\Rightarrow$ Residual waiting time falls.

Intuition for the role of regularity.

- If $\lambda_k - \lambda_{k-1} > \mu_k - \mu_{k-1}$ for all $k \geq 2 \Rightarrow \lambda_k$ increases quickly
  - The initial beliefs put higher weights on long queues
  - Belief that the queue is long given elapse of time is higher
- Bad news is stronger

Evolution of beliefs under FCFS with no information

System of ODEs on the likelihood ratios:

$$\dot{r}_\ell^t = r_\ell^t \left(-(\mu_\ell - \mu_{\ell-1}) + (\mu_\ell r_{\ell+1}^t - \mu_{\ell-1} r_\ell^t) \right)$$

Evolution of beliefs under FCFS with no information

System of ODEs on the likelihood ratios:

$$\dot{r}_\ell^t = r_\ell^t \left(-(\mu_\ell - \mu_{\ell-1}) + (\mu_\ell r_{\ell+1}^t - \mu_{\ell-1} r_\ell^t) \right)$$

Intuition: One could be at position j at $t + dt$ because

- he was at position j at time t: Since μ_j increasing in j, more likely to stay at his position starting at $\ell - 1$ rather than at $\ell \Rightarrow$ likelihood ratio decreases

Evolution of beliefs under FCFS with no information

System of ODEs on the likelihood ratios:

$$\dot{r}_\ell^t = r_\ell^t \left(-(\mu_\ell - \mu_{\ell-1}) + (\mu_\ell r_{\ell+1}^t - \mu_{\ell-1} r_\ell^t)\right)$$

Intuition: One could be at position j at $t + dt$ because

- he was at position j at time t: Since μ_j increasing in j, more likely to stay at his position starting at $\ell - 1$ rather than at $\ell \Rightarrow$ likelihood ratio decreases

- he was at position $j + 1$ at time t: Since μ_j increasing in j, more likely to move from $\ell + 1$ to ℓ rather than from ℓ to $\ell - 1 \Rightarrow$ likelihood ratio may increase

Evolution of beliefs under FCFS with no information

System of ODEs on the likelihood ratios at $t = 0$:

$$\begin{aligned}
\dot{r}_\ell^0 &= r_\ell^0 \left(-(\mu_\ell - \mu_{\ell-1}) + (\mu_\ell r_{\ell+1}^0 - \mu_{\ell-1} r_\ell^0)\right) \\
&= r_\ell^0 (-(\mu_\ell - \mu_{\ell-1}) + (\mu_\ell \frac{\lambda_\ell}{\mu_\ell} - \mu_{\ell-1} \frac{\lambda_{\ell-1}}{\mu_{\ell-1}})) \\
&= r_\ell^0 \left(-(\mu_\ell - \mu_{\ell-1}) + (\lambda_\ell - \lambda_{\ell-1})\right) \leq 0
\end{aligned}$$

for $\ell = 2, ..., K^*$

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

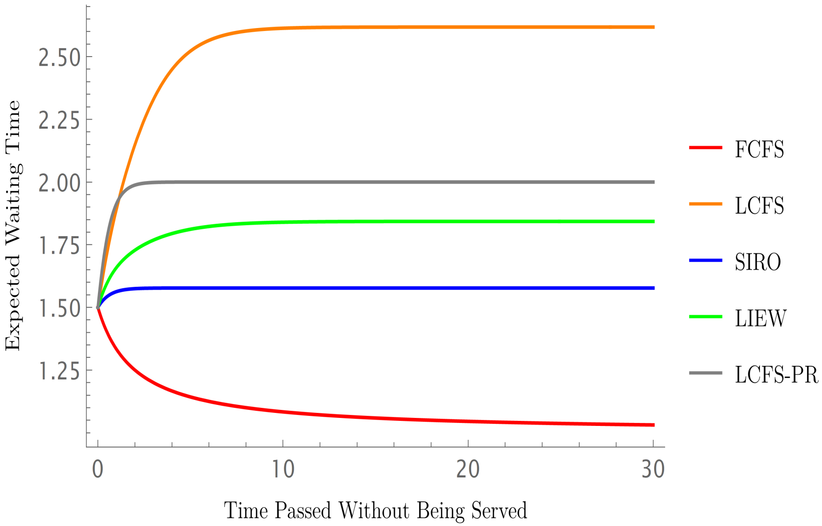
- Giving more information is not optimal: No information pools incentive constraints and helps to incentivize agents to join the queue.

- Queueing disciplines differing from FCFS are suboptimal under any information design: Beliefs about residual waiting time are less favorably updated over time. E.g., under SIRO dynamic IC will be violated for instance when service rate is small compared to entry rate.

Theorem

For any $q \neq$ FCFS, there exists (λ, μ, V, C) such that q fails (IC) under the optimal cutoff policy and under any information rule.

Residual waiting time under alternative queueing rules.



M/M/1 with $K^* = 2$; $\lambda = \mu = 1$.

Concluding Thoughts

- Without information design, the outcome is strictly worse and optimal policy is unknown and is probably complex.
- With information design, FCFS is (uniquely) optimal
- Of course, there may be unmodeled benefits of getting information on queue position or expected waiting times
  - transparency
  - ambiguity aversion...
- Novel role for queueing disciplines in regulating agents' beliefs, and their dynamic incentives
- Reveals a hitherto-unrecognized virtue of FCFS in this regard.

Thank You!

References

ANUNROJWONG, J., K. IYER, AND V. MANSHADI (2020): "Information Design for Congested Social Services: Optimal Need-Based Persuasion," *EC '20: Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 349–350.

ASHLAGI, I., M. FAIDRA, AND A. NIKZAD (2020): "Optimal Dynamic Allocation: Simplicity through Information Design," Discussion paper, Stanford.

BLOCH, F., AND D. CANTALA (2017): "Dynamic assignment of objects to queuing agents," *American Economic Journal: Microeconomics*, 9, 88–122.

HASSIN, R. (1985): "On the optimality of first come last served queues," *Econometrica*, 53, 201–202.

HASSIN, R. (2016): *Rational Queueing*. CRC Press.

HASSIN, R., AND A. KOSHMAN (2017): "Profit maximization in the M/M/1 queue," *Operations Research Letters*, 45, 436–441.

LESHNO, J. (2019): "Dynamic matching in overloaded waiting lists," Discussion paper, SSRN Working Paper 2967011.

LINGENBRINK, D., AND K. IYER (2019): "Optimal signaling mechanisms in unobservable queues," *Operations Research*, 67, 1397–1416.

MARGARIA, C. (2020): "Queueing to learn," Discussion paper, Boston University.

NAOR, P. (1969): "The regulation of queue size by levying tolls," *Econometrica*, 37, 15–24.

SU, X., AND S. ZENIOS (2004): "Patient choice in kidney allocation: The role of the queueing discipline," *Manufacturing and Services Operations Management*, 6, 280–301.