Optimal Queue Design[1]

Yeon-Koo Che[2] Olivier Tercieux[3]

April 4, 2022

Abstract

We study the optimal design of a queueing system when agents' arrival and servicing are governed by a general Markov process. The designer chooses *entry* and *exit* rules for agents, their *service priority*—or *queueing discipline*—as well as their *information*, while ensuring they have incentives to follow recommendations to *join* the queue and, importantly, to *stay* in the queue. Under a mild condition, at the optimal mechanism, agents are induced to enter up to a certain queue length and no agents are to exit the queue; agents are served according to a first-come-first-served (FCFS) rule; and they are given no information throughout the process beyond the recommendations they receive from the designer. FCFS is also necessary for optimality in a rich domain. We identify a novel role for queueing disciplines in regulating agents' beliefs, and their dynamic incentives, thus uncovering a hitherto unrecognized virtue of FCFS in this regard.

JEL Classification Numbers: C78, C61, D47, D83, D61
Keywords: Queueing disciplines, information design, mechanism design, dynamic matching.

[1]We are grateful to Ethan Che, Laurens Debo, Laura Doval, Drew Fudenberg, Refael Hassin, Moshe Haviv, Yash Kanoria, Krishnamurthy Iyer, Ioannis Karatzas, Jinwoo Kim, Jacob Leshno, Shengwu Li, Vahideh Manshadi, Chiara Margaria, Afshin Nikzad, Chris Ryan, Robert Shumsky, Eduardo Teixeira, and seminar participants at Chicago-Booth, Columbia DRO-IEOR, Harvard/MIT, NYU-Stern, Dartmouth-Tuck, NUS, ACM-EC, INFORMS, Hong-Kong virtual, VSET, AMET, and Toulouse, for their helpful comments. We acknowledge research assistance from Will Grimme, Dong Woo Hahm, and Sara Shahanaghi. Yeon-Koo Che is supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5A2A03043516)

[2]Department of Economics, Columbia University, USA. Email: yeonkooche@gmail.com.
[3]Department of Economics, Paris School of Economics, France. Email: tercieux@pse.ens.fr.

1 Introduction

Consider the problem faced by someone, called a designer, who designs a queueing system for agents seeking to receive a service or product. Agents arrive stochastically according to some Markov process, and are served according to another Markov process, both depending on the number of agents in the queue. Aside from these two processes, which are exogenous and thus beyond her control, the designer can choose many aspects of the queue design. She can let an arriving agent enter the queue or turn him away. She can remove an agent from the queue. She can also decide which agent will be served at each point, or more generally, how available service capacity is allocated among agents in the queue at each instant. Finally, the designer can control how much information an agent has about the queue or his expected waiting time throughout the process, both when he arrives at the queue and at any point after he has joined.

Casual observation of real-world queues suggests a wide range of choices available along these different dimensions of queue design. Service call centers sometimes encourage customers to wait in line (i.e., to be put on hold); other times, presumably in the face of high call volume, they tell customers to try another time. Some call centers ask customers to leave the queue and return later. Queue disciplines—the service priorities for agents in the queue—also take a variety of forms: *first-come-first-served* (FCFS) is the oldest and by far the most common queue discipline, but *service-in-random-order* (SIRO) which assigns priority at random, has been also used. Some authors have proposed other rules such as *last-come-first-served* (LCFS) (e.g., Hassin (1985), Su and Zenios (2004), and Platz and Østerdal (2017)). Finally, a range of different information policies are commonly observed. Some queueing systems keep customers completely in the dark about the queue length, their relative positions, or their estimated waiting times. For instance, many offices for social housing do not disclose any information on positions on waiting lists.[1] Other systems provide customers with their estimated waiting time or the number of customers ahead of them. For instance, popular ride-hailing apps provide a customer with not only the estimated arrival time of a vehicle but also its current location on a map.

A queue design, together with the primitive process, induces a Markov chain on the length of the queue, and we focus on the steady state where the chain is at its invariant distribution. This in turn determines an agent's expected waiting time and the service rate in steady state. For these quantities to be feasible, they must be incentive compatible for

[1]This is the case, for instance, for several housing choice voucher programs in California, e.g., PCCDS Housing Service or HACA among others.

agents. While our designer can keep an agent from joining the queue or remove one from the queue, she cannot coerce an agent to enter the queue or to stay in the queue against his will. In other words, when recommended to either join or stay in a queue, an agent must have an incentive to obey this recommendation given the information that he has. Subject to this feasibility requirement, the designer maximizes a weighted sum of agents' welfare and service provider's profit. Since the weight is arbitrary, the designer could be a service provider who maximizes the profit, a consumer advocate who maximizes agents' welfare, or a regulator who values both.

The question is: *how should the designer choose all different aspects of the queue design?* Under a very mild *regularity condition* on the process, our answer is strikingly simple and consistent with many observed practices of queue design. (i) The optimal queue design has a *cutoff* policy: namely, there exists a maximal queue length $K \geq 0$ such that agents are recommended to enter the queue if and only if its length is less than K.[2] (ii) Those who join the queue are then prioritized to receive a service according to FCFS. (iii) No information is provided to agents beyond the recommendations they receive to join or to stay in the queue.[3]

Result (i) (shown in Section 4) means that one can achieve an optimal queue design, without removing agents or incentivizing them to leave the queue once they join the queue. Reneging—or abandonment of the queue—is then never part of our optimal queue behavior.[4] Results (ii) and (iii) (both shown in Section 5) mean that, at least in the canonical model we consider, the most tried-and-true queueing norm is (at least weakly) better than any others, provided that agents receive no information beyond the recommendations from the designer. Our "no" information policy is also common in practice. Call centers often simply suggest "please hold," and the common practice of offering a rough estimate on the waiting time is one way to implement that policy.

The intuition behind the information policy—no information beyond recommendation—is explained as follows. It is well known and intuitive that incentive constraints are relaxed most when agents are given as little information as possible. If an agent has the incentive to join or to stay in a queue for a set of signals, he must also have the same incentives when all these signals are pooled into one, regardless of the queueing discipline. Since this "pooled" signal is precisely what the agent will have given "no information" beyond the

[2] When the queue length is $K-1$, an agent is recommended to enter with a positive probability possibly equal to one. If this probability is less than one, the entry is "rationed" at $K-1$.

[3] Since recommendations contain information about the state, this policy should not be confused with "no information" authors often use, which refers to "no communication" what so ever. Agents can make Bayesian inferences on their expected waiting times, based on the recommendation they receive, the queue design that the designer commits to, and the elapsed time after joining the queue.

[4] Removal of agents can only be consistent with optimality if it occurs when the queue is full or near full.

recommendation, the no information policy is optimal.

To explain why FCFS is optimal, fix an optimal entry and exit policy—i.e., a cutoff policy with some maximal length K. Assuming agents obey the recommendation, this induces a distribution of queue length in the steady state. Since our agents are homogeneous, the expected waiting time *when averaged across possible initial queue lengths* is the same for each agent, and does not depend on the queueing discipline in use. Then, given no information, the incentive for joining the queue will be the same across all queueing disciplines, and on this account, FCFS is not particularly necessary or desirable.

However, the dynamic incentives that agents face—their incentive to "continue" queueing once they join the queue—differ across queueing disciplines, assuming the no information policy. The reason is that the distribution of waiting times differs across queueing disciplines, so one updates beliefs about the remaining waiting times differently as time passes under different queueing disciplines. Our main insight is that, under the *regularity* condition on the primitive process, the evolution of these beliefs become progressively more favorable under FCFS. Consequently, under the condition, agents are willing to stay in the queue under FCFS with no information, thus implementing the optimal queueing outcome.

The progressively improving beliefs under FCFS stem from its fundamental property: namely, that one's service priority can only improve over time under FCFS. Hence, starting with any initial queue length, the elapse of time is indeed *good news* about the remaining waiting time. But there is also a countervailing force. Since an agent is not told about the queue length k when he joins the queue (recall that agents get no information beyond the designer's recommendations), his belief about this will be also updated as time progresses. On this account, the elapse of time is actually *bad news*, since it indicates that the agent likely underestimated the initial length of the queue when he joined it. We show that the good news dominates the bad news under the regularity condition. As noted above, this means that incentive compatibility is maintained throughout once an agent is willing to join the queue under FCFS.

The belief evolution is not as favorable for other queueing disciplines, however. Consider SIRO. Since priority is assigned randomly, one's queue position does not matter; instead, his belief about the current queue length is what matters for his incentives: the more agents there are in the queue, the less likely it is for an agent to receive service. Hence, the elapse of time (without being served) is a signal that there are more agents in the queue than he initially thought. Further, unlike FCFS, there is no "good news" since his priority does not improve over time. So, the agent becomes more pessimistic as time passes. Indeed, we can find simple cases such as the standard $M/M/1$ queue in which the belief worsens over time

to such a degree that an agent leaves the queue after entering it, thus failing the incentive requirement necessary for implementing the optimal cutoff policy.

In fact, there is a sense in which the FCFS is uniquely best in dealing with the dynamic incentives problem. In Section 6, we show that for *any* queueing discipline differing from FCFS, there exists an environment under which it is strictly suboptimal no matter the information policy adopted. In this sense, FCFS is not only optimal under the no information policy, it is also *necessary* for optimality in a rich domain.

Related Literature. The current paper follows the long line of queueing theory research, in particular, the *rational queueing* literature. This literature, which has developed into a significant body of work since the seminal work by Naor (1969), studies the strategic behavior of rational Bayesian agents in a variety of queueing scenarios.[5] While sharing their focus and approach, the current paper is distinguished from standard works in this literature in several respects.

First, our Markovian model is general and flexible enough to encompass many settings of interest. A typical queueing model tends to focus on a specific process such as $M/M/1$ or $M/M/c$. Similarly, a standard dynamic matching model in economics considers a specific match technology. By contrast, our model allows the arrival and servicing of agents to follow general Markov processes that may depend on the current queue length, which nests $M/M/c$ (which in turn subsumes $M/M/1$) queueing models as well as recent economic models of dynamic matching as special cases.[6]

Second, we consider agents' incentives not only to *join* but more importantly to *stay* in the queue when recommended by the designer to do so. Addressing these latter dynamic incentives distinguishes the current paper from most of the existing ones. There are a few papers that consider incentives by agents to abandon a queue, or to "renege"; see Hassin and Haviv (1995), Haviv and Ritov (2001), Mandelbaum and Shimkin (2000), Sherzer and Kerner (2018), and Cripps and Thomas (2019). However, these papers approach the issue as a positive theory, trying to explain reneging as a rational strategic response to various features such as nonlinear waiting costs or aggregate uncertainty. Our approach is instead to treat the issue from a normative perspective, and more systematically as part of incentive design, following the tradition of mechanism design, as we note next.

Third, the current paper is distinguished in its comprehensive treatment of many as-

[5] See Hassin and Haviv (2003) and Hassin (2016), for an excellent survey of the literature.

[6] As we mention in Section 3, the models in this literature consider agents who can only be matched with "compatible" agents (or objects) in the queue. Assuming that each pair is compatible with some fixed probability, the effective arrival rate (i.e., the rate at which an agent joins the queue) and the effective service rate (i.e., the rate at which an agent leaves the queue) depends on the number of agents in the queue.

pects of queueing system design. Most of the existing papers (inlducing Naor (1969)) do not consider the optimal entry/exit policies explicitly, but rather focus on some, typically unregulated, exogenously given queueing environment. Likewise, queueing disciplines are often assumed in the literature to be FCFS or, less frequently, SIRO. Agents' information is also typically fixed; authors often assume that agents have either full information about the queue or no information whatsoever.[7]

A few papers study the optimal design of queueing disciplines while taking other aspects of queueing system as given. Following Naor (1969)'s seminal observation that FCFS causes agents to queue *excessively*, ignoring the "congestion" externality they inflict on later agents, Hassin (1985) and Su and Zenios (2004) argue that LCFS can "cure" this externality and is optimal for agents.[8] Excessive incentives for queuing are not a problem for our designer since she can control entry by withholding service. On the other hand, FCFS may give too few incentives if the designer maximizes (or close to maximizing) service provider's profit or his service utilization, or there is excessive supply of agents as in the case of Leshno (2019). In these cases, other mechanisms such as SIRO could perform better by providing greater incentives for queueing. But this conclusion rests crucially on agents having full information about the queue length. We show that FCFS is *always* optimal regardless of the agents' incentives if the designer can also choose an optimal information policy—namely, *no information (beyond recommendations)*. Further, its optimality is strict if one also considers agents' dynamic incentive to stay in the queue after joining it, which the above papers do not consider.[9]

Despite its practical relevance, information design has received attention only recently in the queueing literature; see Simhon, Hayel, Starobinski, and Zhu (2016), Hassin and Koshman (2017), Lingenbrink and Iyer (2019), and Anunrojwong, Iyer, and Manshadi

[7]Here no information means that agents truly do not have *any* information, including a recommendation from the designer. In fact, this assumption is typically made in the context of an *unregulated* environment, where there is no designer or supervisory entity overseeing or managing the queue. See Hassin and Haviv (2003) for the canonical description of the unregulated environment.

[8]Platz and Østerdal (2017) find a similar result when there are a continuum of agents who enter at their endogenously chosen times. See also Haviv and Oz (2016) for alternative schemes in the observable environment and Haviv and Oz (2018) for extensions to the unobservable queue environment.

[9]Several papers study alternative queueing disciplines in environments that are less related or comparable to ours. FCFS is shown to be optimal in Bloch and Cantala (2017) and a part of the optimal design in Margaria (2020) in models where, unlike the standard queueing model, the lengths of queues are non-stochastic, either because arrival occurs only when an agent exits (the former) or because there are a continuum of agents (the latter). Further, they do not consider information design, so the reason for the optimality of FCFS is completely different in these models than in our model. Kittsteiner and Moldovanu (2005) consider the allocation of priority in queues via bidding mechanisms where processing time is private information. The crucial difference is the use of transfers implicit in bidding mechanisms, which is not allowed in our model.

(2020).[10] While the last three identify the same optimal information design as the current paper, they do not compare alternative queue designs and they do not consider dynamic incentives. By contrast, we allow all these dimensions of design—entry, exit, queueing disciplines, and information design—to be chosen optimally by the designer in the face of the dynamic incentive problem.

In terms of style, the current paper is closest in spirit to the mechanism design literature, pioneered by Myerson (1981), which takes the underlying physical character of the environment as given but otherwise allows the designer to optimally choose all other aspects of the system.[11] Interestingly, our main findings are also similar in flavor to those of Myerson (1981): *the optimal mechanism is both simple and resembles commonly observed practices.* As mentioned, the cutoff policy conforms to the standard practice of capping the queue length at some level. The optimality of FCFS accords well with its prevalent use in practice, and is reassuring in light of its perceived fairness (see Larson (1987)). The *no information beyond recommendation* policy also conforms to standard practice in call centers which put customers on hold, often with no information on their waiting times, unless they are *explicitly* discouraged from waiting in line.[12]

2 Model and Preliminaries

We consider a generalization of a canonical queueing model in which agents arrive sequentially at a queue to receive a service. Time indexed by $t \in \mathbb{R}_+$ is continuous.

Agents' payoffs. There are three parties: a *designer*, who organizes resource allocation including the queueing policy, a *service provider* who services agents, and *agents* who receive service. As will be seen, the designer may be the service provider, a representative of the agents, or a planner who reflects the welfare of both parties.

[10] In a less related model, Ashlagi, Faidra, and Nikzad (2020) study optimal dynamic matching with information design, showing that FCFS, together with an information disclosure scheme, can be used to implement the optimal outcome. Although similar at first glance, their model is quite different from, and not easily comparable to, ours. There are a continuum of agents in their model, and their information policy pertains to the quality of good rather than to agents' queue position. In particular, the virtue of FCFS in regulating agents' beliefs on where they stand in the queue is orthogonal to Ashlagi, Faidra, and Nikzad (2020)'s insights.

[11] Within the queueing literature, the optimal design or control literature focuses on the ex ante choice of the service and arrival process, which we take as given (see Shaler Stidham (2009) for a survey). Hence, one can view the current work as complementing this literature.

[12] As we already pointed out, offices for social housing often provide applicants with very limited information on their position in the list. In addition, these offices often cap waiting lists when they are too long.

The agents are homogeneous in their preferences. Each agent enjoys a payoff of

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

if she receives service after waiting $t \geq 0$ time period, where $V > 0$ is the net surplus from service (possibly after paying a service fee to the designer) and $C > 0$ is a per-period cost of waiting. The service provider earns profit $R > 0$ from each agent she services. In a customer service context, the profit may not take the form of monetary fees collected from customers but rather the shadow value of fulfilling a warranty service or more generally addressing any customer needs (See Section 7 for a discussion of an endogenously set monetary fee collected from customers). The designer's objective, which will be specified more fully below, is a weighted sum of the service provider's and agents' payoffs. An agent's outside option, which she collects when not joining the queue or from exiting one, yields zero payoff.

Primitive process. At each instant, given the number of agents in the queue, or **queue length**, $k \in \mathbb{Z}_+$, an agent arrives at a Poisson rate of $\lambda_k > 0$ and an agent in a queue is served at a Poisson rate of $\mu_k > 0$. Hence, a pair (λ, μ), where $\lambda \triangleq \{\lambda_k\}$ and $\mu \triangleq \{\mu_k\}$, specifies a **primitive process**. We view (λ, μ) as arrival and service rates that arise in many queueing environments of interest, including $M/M/c$ queue models and dynamic matching models, as illustrated in Section 3; for instance, the possibility of arrival and service rates depending on the current queue length k emerges naturally from a dynamic matching context.

We interpret μ_j as the maximal service rate that *any* set of j or fewer agents may receive in any queue of length $k \geq j$. It is then without loss to assume that μ_k is nondecreasing in k.[13] We also assume that μ_k is bounded uniformly in k. In addition, our results invoke one of the following conditions:

Definition 1. (i) The **service process** $\mu = \{\mu_k\}$ is **regular** if $\mu_k - \mu_{k-1}$ is nonincreasing in k. (ii) The **primitive process** (λ, μ) is **regular** if the service process μ is regular and $\lambda_k - \lambda_{k-1} \leq \mu_k - \mu_{k-1}$ for each $k \geq 2$.

These two regularity conditions are extremely mild. In fact, we are not aware of any queueing model where the regularity is violated; Section 3 shows all the canonical queueing models as well as dynamic matching models satisfy these two conditions.[14]

Designer's policy. The designer has a number of instruments at her disposal. We focus on an anonymous stationary Markovian policy that treats all agents identically based on two

[13] See Section S.1 in the online appendix for details.

[14] In particular, as shown in the online appendix Section S.1, the regularity of the service process, namely, (i), has a desirable axiomatic foundation.

state variables: the queue length k and the queue position ℓ, namely the arrival order of an agent among those in a queue. The stationarity restriction means that the policy does not depend on the calendar time. The designer chooses the following set of policies.

• **Entry and exit rule:** The entry and exit rules specify how the designer regulates entry of agents who arrive to a queue and exit of those who are already in the queue. Formally, an **entry rule** is given by $x = (x_k)$, where $x_k \in [0, 1]$ denotes the probability that an arriving agent is asked to join a queue of length k. An **exit rule** is given by $(y, z) = (y_{k,\ell}, z_{k,\ell})_{k,\ell}$. The designer removes the agent with queue position ℓ from the queue of length $k \geq \ell$ at a Poisson rate $y_{k,\ell} \geq 0$. In addition, upon a new arrival in the queue, the designer can maintain the queue length constant by removing an agent currently in the queue: $z_{k,\ell} \in [0, 1]$ denotes the probability that an agent with queue position ℓ is removed from a queue of length k when another agent is joining the queue.[15] The entry rule could accommodate the possibility of non-entry that is either involuntary or voluntary. Similarly, the exit rules y and z capture both the explicit policy of diverting some agent away from a service pool (e.g., Mandelbaum and Shimkin (2000)) as well as the abandonment induced by a queueing policy (to be described below). The main difference between y and z pertains to whether the removal is conditional on the entry of another agent. In particular, z captures the possibility of an agent being "preempted" by a new arrival, e.g., under a LCFS rule (see Hassin (1985)). We let $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ denote the set of all feasible (x, y, z)'s.

• **Queueing rule:** A queueing rule specifies the allocation of an available service rate among agents in the queue based on its length and the agents' queue positions.[16] Formally, a **queueing rule** is given by $q = (q_{k,\ell})$, where $q_{k,\ell} \geq 0$ is the Poisson rate at which an agent receives service when the queue length is k and her position in the queue is ℓ. Feasibility requires that $\sum_{\ell \in S} q_{k,\ell} \leq \mu_{|S|}$, for all k and all $S \subset \{1, ..., k\}$: that is, the total service rate received by a subset of agents in the queue cannot exceed the service rate available for the number of those agents.[17] As is standard, we also require a feasible queueing rule to be *work conserving*: $\sum_{\ell=1}^{k} q_{k,\ell} = \mu_k$, for all queue length k. This means that the allocation of

[15] By definition, if an agent ℓ is removed, no other agent $\ell' \neq \ell$ is removed.

[16] In fact, we can allow queueing rules to be fully general, i.e., without limiting ourselves to those that depend only on (k, ℓ); examples include rules that allow service probabilities to vary with time and to depend on the history leading up to the current queue length and positions. However, our class entails no loss since the optimal rule in this fully general class belongs to the current class that we focus on.

[17] Recall that we interpret μ_j as the maximal rate at which a set of j (or fewer) agents in the queue can be served collectively. Hence, the feasibility condition simply requires that any subset of agents of size j must be collectively served at rate no greater than this maximal service rate μ_j. For instance, in the $M/M/c$ queue model, there are c servers each able to serve an agent at rate, say μ. Then, any j agents can be served at most at rate $\mu_j = \min\{j, c\}\mu$ in total.

service is "non-wasteful," or exhausts the available service capacity. We let $\mathcal{Q}$ denote the set of all work-conserving queueing rules. The set $\mathcal{Q}$ encompasses all standard queueing disciplines. For instance, assuming the service process is regular, **first-come-first-served (FCFS)** satisfies $q_{k,\ell} \triangleq \mu_\ell - \mu_{\ell-1}$. Namely, the agent in position 1 enjoys the highest possible service rate μ_1 for any single agent; given this, the agent in position 2 receives the highest possible service rate, $\mu_2 - \mu_1 \geq 0$, and so on. The regularity condition guarantees the service rate can only fall as one's position gets worse. (We will see in Section 3 how this corresponds to more familiar expressions in the canonical queuing models such as $M/M/1$, $M/M/c$ or dynamic matching models.) Similarly, **last-come-first-served (LCFS)** satisfies $q_{k,\ell} \triangleq \mu_{k-\ell+1} - \mu_{k-\ell}$, and **service-in-random-order (SIRO)** satisfies $q_{k,\ell} \triangleq \mu_k/k$, for all $k \in \mathbb{N}, \ell \leq k$.[18]

• **Information rule:** An information rule specifies the payoff relevant-information available to the agents who are recommended to stay in the queue after each time $t \geq 0$ he has spent in the queue, including $t = 0$ when he has just arrived at the queue. Since an agent has a linear waiting cost, the only payoff-relevant information at each time spent $t \geq 0$ is the probability $\sigma_t \in [0, 1]$ that he will be eventually served and the expected remaining waiting time $\tau_t \in [0, \infty]$ on the queue (before exiting the queue either because he received service or was removed from the queue).[19] Given the memoryless nature of the process $(\lambda, \mu, x, y, z, q)$, these two variables depend only on the current queue length k and one's queue position $\ell \leq k$ and are independent of the time t one has spent in the queue, so we write $(\sigma_{k,\ell}, \tau_{k,\ell}) \in [0, 1] \times [0, \infty]$ for each (k, ℓ). An agent's (payoff-relevant) information then boils down to his information regarding (k, ℓ) at each time $t \geq 0$. As is well-known, say from Kamenica and Gentzkow (2011), this information can be represented as a distribution of "posterior beliefs" about (k, ℓ), which does in general depend on time in the queue $t \geq 0$. Formally, an **information rule** is given by $I = (I_t)_{t \in \mathbb{R}_+}$, where $I_t \in \Delta(\Delta(\mathbb{Z}_+ \times \mathbb{N}))$ specifies a distribution of posterior beliefs on (k_t, ℓ_t) at time t. Feasibility requires that posterior beliefs at each t must be adapted to the filtration generated by the process $(\lambda, \mu, x, y, z, q)$ and must satisfy Bayes rule given his prior beliefs and knowledge of the process $(\lambda, \mu, x, y, z, q)$. Let $\mathcal{I}$ denote the set of all feasible information rules. (We suppress the dependence both of $(\sigma_{k,\ell}, \tau_{k,\ell})$ and $\mathcal{I}$ on $(\lambda, \mu, x, y, z, q)$ for notational ease.) The set $\mathcal{I}$ is large enough to include

[18] The regularity of the service process ensures that these standard queueing disciplines are work conserving. Conversely, the regularity property is necessary if one requires FCFS and LCFS to be work conserving. See Section S.1 in the online appendix.

[19] The waiting time refers to the duration of time an agent spends in the queue, including the service time. In the queueing literature, this is sometimes referred to as *sojourn time*.

all realistic information rules that are feasible.[20] Special cases include **full information**, in which case I_t coincides with the true distribution of (k_t, ℓ_t), and **no information**, in which case the posterior I_t is degenerate on the belief obtained by Bayes updating via $(\lambda, \mu, x, y, z, q)$ from the prior beliefs I_0. There are many "intermediary" information rules between full and no information. For instance, the designer may simply reveal whether, upon joining the queue, the agent's expected waiting time is below or above some predetermined threshold.[21]

Given the primitive process (λ, μ), a Markov policy (x, y, z, q) generates a Markov chain—more specifically, a birth-and-death process—on the queue length k. Given (λ, μ), we only consider a Markov policy that induces an invariant distribution $p \triangleq (p_0, p_1, ...p_\infty)$ on the queue length. Specifically, this means that the distribution p must satisfy the following balance equation:

$$\lambda_k x_k (1 - \sum_\ell z_{k,\ell}) p_k = (\mu_{k+1} + \sum_\ell y_{k+1,\ell}) p_{k+1}, \ \forall k \in \text{supp}(p) \tag{B}$$

The LHS of the equation is the rate at which the queue length transits from k to $k+1$: with probability p_k the queue length is k, in which case an agent arrives at rate λ_k, is recommended to join the queue with probability x_k, and no agent is removed from the queue with probability $1 - \sum_\ell z_{k,\ell}$. The balance equation (B) requires this rate to equal the rate at which the queue length transits from $k+1$ to k, namely its RHS: with probability p_{k+1} the queue length is $k+1$, in which case an agent is served at rate μ_{k+1} or is removed at rate $\sum_\ell y_{k+1,\ell}$ from the queue. We say that an entry/exit policy $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ **generates** an invariant distribution p if (x, y, z, p) satisfies (B), and call the associated tuple (x, y, z, p) an **outcome**.

Incentives. We assume that the designer may prohibit an agent from joining the queue and may also remove an agent from a queue, but that the designer cannot coerce an agent to join or stay in the queue against her preference. Consequently, when recommended to enter the queue or to stay in the queue, an agent must have an incentive to obey that recommendation, given the information available to her.

Formally, this obedience constraint is specified in terms of an agent's beliefs about the queue length and position (k_t, ℓ_t) at each time, which in turn determines the conditional ser-

[20] Just like the queueing rule, we can allow for a more general information design, one that may allow the information to vary depending on the rich history of the process beyond (k, ℓ). This additional generality is irrelevant since our optimal information rule (which is Markovian) attains the upper bound in the designer's objective regardless of such unrestricted information.

[21] For many queueing rules (e.g., FCFS), this will mean specifying whether the agent's position is above or below a certain predetermined integer L. Formally, I_0 will put weight only on two possible posterior beliefs, one with support in $\{1, \ldots, L\}$, the other one with support in $\{L+1, \ldots, K\}$.

vice probability and expected residual waiting times $(\sigma_{k,\ell}, \tau_{k,\ell})$. We evaluate these variables when the system is at its invariant distribution p. Obedience then requires:

$$\sum_{k,\ell} \gamma^t_{k,\ell} \left[V \cdot \sigma_{k,\ell} - C \cdot \tau_{k,\ell} \right] \geq 0, \forall \gamma^t \in \text{supp}(I_t), \forall t \geq 0, \tag{IC}$$

where $(\sigma_{k,\ell}, \tau_{k,\ell})$ is induced by the policy (x, y, z, q) and where $(\gamma^t_{k,\ell})$—which stand for the posterior beliefs about state (k_t, ℓ_t)—are induced by a feasible information rule I_t for each t, given the initial beliefs which satisfy:

$$\int_{\gamma^0 \in \text{supp}(I_0)} \gamma^0_{k,k} I_0(d\gamma^0) = \frac{\lambda_{k-1} x_{k-1} p_{k-1}}{\sum_{j \in \mathbb{N}} \lambda_{j-1} x_{j-1} p_{j-1}}, \forall k \in \mathbb{N}.$$

In words, (IC) states that each agent, when recommended to join or stay in the queue, must find the prospect of being served to be high enough to justify the remaining waiting cost, given each possible belief $(\gamma^t_{k,\ell})$ at each $t \geq 0$. The last condition simply means that the belief one has on each (k, k) at the time of entry—the probability of being the k-th entrant to the queue at $t = 0$—must on average (over all possible realizations of posterior beliefs) equal the probability of the queue length being $k - 1$ conditional on joining p_{j-1} evaluated at the invariant distribution p given the entry policy x.

In the sequel, we refer to the incentive constraint for t by (IC_t). We say that a queueing/information policy $(q, I) \in \mathcal{Q} \times \mathcal{I}$ **implements** an outcome (x, y, z, p) if (IC) holds. Even though we interpret an implemented outcome as resulting from the designer's policy choice, this is without loss, due to the revelation principle. Our model can capture any equilibrium outcome, both regulated and unregulated. For instance, consider the textbook unregulated and unobservable $M/M/1$ queue governed by FCFS, in which agents make their entry decisions without any recommendation or any information about the queue length (see Hassin and Haviv (2003) for instance). There is an equilibrium of such a model in which each agent enters the queue with some probability $e \in (0, 1]$ and stay in the queue until he is served.[22] In our model, this corresponds to our entry policy of $x_{k,\ell} = e$ and $y_{k,\ell} = z_{k,\ell} = 0$, for all k, ℓ (along with FCFS and no information).

Problem statement. The designer's objective is evaluated at the invariant distribution

[22] More specifically, one can show that, if λ is sufficiently low (more precisely, if $(\mu - \lambda)V > C$), then agents enter with probability $e = 1$. If not, then there exists a random entry probability $e \in (0, 1)$ such that if all agents adopt this mixing strategy, each agent becomes indifferent to entry, making it an equilibrium behavior.

$p = (p_k)$ of the Markov chain. It can be written as follows:

$$W(p) \triangleq (1-\alpha)R\sum_{k=1}^{\infty} p_k\mu_k + \alpha\sum_{k=1}^{\infty} p_k(\mu_k V - kC),$$

where $\alpha \in [0,1]$. The first term is the flow expected profit for the service provider: with probability p_k, the queue has k agents, and an agent is served at rate μ_k, generating a profit (or shadow value) of R for each agent served. The second term is the flow expected utility for agents: again with probability p_k, the queue has k agents, each of whom pays holding/waiting cost of C per unit time (the second term), and an agent is served, and realizes a surplus of V, at rate μ_k. The objective is a weighted sum of these two terms, with weight $\alpha \in [0,1]$.

The designer's problem is to choose $(p, x, y, z, q, I) \in \Delta(\mathbb{Z}_+) \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{Q} \times \mathcal{I}$ to

$$[P] \qquad \text{Maximize } W(p) \text{ subject to } (B) \text{ and } (IC),$$

where the conditional service probabilities and residual waiting times $(\sigma_{k,\ell}, \tau_{k,\ell})$ in (IC) are induced by (p, x, y, z, q).[23] In words, the designer picks the outcome that maximizes her objective among those that are implementable by some queueing/information policy. Let $\mathcal{W}$ denote the supremum of the value of program $[P]$.

3 Scope of Applications

Our model encompasses a variety of queueing and dynamic matching models considered by the existing literature.

• $M/M/1$ ***queue model:*** This is the most canonical queueing model in which the arrival rate λ_k and service rate μ_k do not depend on the queue length k. Hence, regularity is trivially satisfied. Our queueing formula simplifies to $q_{k,\ell} = \mathbf{1}_{\{\ell=1\}} \cdot \mu$ under FCFS, $q_{k,\ell} = \mathbf{1}_{\{\ell=k\}} \cdot \mu$ under LCFS, and $q_{k,\ell} = \mu/k$ under SIRO, for some $\mu > 0$. Naor (1969) and Hassin (1985) investigate agents' incentives to join the queue under FCFS and LCFS, respectively. More recently, Simhon, Hayel, Starobinski, and Zhu (2016), Hassin and Koshman (2017), and Lingenbrink and Iyer (2019) study information design to manage agents' incentive to join the queue, both under FCFS queueing rule. Compared with these papers, our model is more general with respect to the queueing environment (we allow for state-dependent arrival and

[23] While the entry/exit policy (x, y, z) uniquely pins down the invariant distribution, we include p as part of the designer's choice.

service rates), agents' incentives (we consider their incentives to stay in, not just to join, a queue), and queueing rules (we allow for fully general queueing rules, not just FCFS).

• $M/M/c$ ***queue model:*** This generalizes the $M/M/1$ queue model to allow for multiple $c \geq 2$ servers, each with exponential service time. As with M/M/1, the arrival rate is independent of the queue length k, but the service rate is linear up to the number of available servers, so $\mu_k = \min\{k, c\}\mu$, where $\mu > 0$ is a service rate by a single server. One can see that regularity is satisfied. Our queueing formula simplifies to $q_{k,\ell} = \mathbf{1}_{\{\ell \leq c\}} \cdot \mu$ under FCFS, $q_{k,\ell} = \mathbf{1}_{\{k-\ell+1 \leq c\}} \cdot \mu$ under LCFS, and $q_{k,\ell} = \min\{k, c\}\mu/k$ under SIRO. In fact, our model can be seen as a generalization of $M/M/c$ model in which servers are heterogeneous in their service rates: server j serves at rate $\tilde{\mu}_j := \mu_j - \mu_{j-1}$, with $\mu_0 := 0$.[24]

• ***Team servicing model:*** Suppose there are m customers (or machines) each having a service need arising at an independent Poisson rate while operating (see Gnedenko and Kovalenko (1989), p. 42). There are c servers who can serve a customer at rate μ. When there are k agents in the queue, the new arrival rate is $\lambda_k = (m-k)\lambda$ and the service occurs at rate $\mu_k = \min\{k, c\}\mu$. Again, our regularity condition holds. (The standard queueing rules are the same as above.)

• ***Dynamic one-sided matching with stochastic compatibility:*** Suppose each agent is compatible with another agent with probability $\theta \in (0, 1]$. In this model, an agent joins a queue only when he arrives at some rate η *and* is incompatible with the agents already in the queue, which occurs with probability $(1-\theta)^k$, or else, she matches with a compatible partner and does not join the queue, which occurs with probability $\eta(1-(1-\theta)^k)$. This is a special case of our model in which $\lambda_k = \eta(1-\theta)^k$ and $\mu_k = \eta(1-(1-\theta)^k)$. Observe that μ_k is increasing at a decreasing rate, and λ_k is decreasing, in k, so the process is regular. Our queueing formula for FCFS, for instance, yields the service rate for ℓ-th positioned agent to be $q_\ell = \mu_\ell - \mu_{\ell-1} = \eta(1-\theta)^{\ell-1}\theta$, the probability that all agents ahead of him are incompatible, and he is compatible, with an incoming agent. Likewise, LCFS and SIRO formula have intuitive interpretations. Doval and Szentes (2018) consider such a model with $\theta = 1$ and study agents' incentive to join a queue under FCFS. Akbarpour, Li, and Gharan (2020) study the limit as $\theta \in (0, 1)$ tends to 0 but the arrival rate increases. Their focus differs from ours; for instance, they do not consider the incentive to join or stay in a queue, the queueing rule, or information design.[25] One can see that the regularity condition holds for all θ. Note also that, if $\theta \in (0, 1)$, even under FCFS a low-priority agent may be "served"

[24] Strictly speaking, under FCFS, this means that when a more efficient server completes his service, the agent in the next position moves up to be served by that server.

[25] Instead, they study the benefit from thickening the market, which we do not consider.

(or matched) if all higher-priority agents are incompatible with the newly arriving agent.

• ***Dynamic two-sided matching with stochastic compatibility:*** Heterogeneous agents on one side match with heterogeneous agents or objects (e.g., housing) on the other side. If the types of the matched pair are compatible, then high surplus is realized; if not, a low surplus is realized. The designer operates buffer queues for different types of agents or objects to keep the agents waiting until a compatible match is found. Leshno (2019) and Baccara, Lee, and Yariv (2020) consider such models. In these models, if one buffer queue is active, the other is empty. Hence, the system can be analyzed as a one-dimensional Markov chain. Some of our results below rely on the system induced by a given policy to exhibit birth and death processes. Indeed, this feature is satisfied under the optimal policy under Baccara, Lee, and Yariv (2020) but not under Leshno (2019). Nevertheless, our central results apply to the latter setup, as we show in Section S.7 of the online appendix. Baccara, Lee, and Yariv (2020) consider optimal matching policy under both FCFS and LCFS, whereas Leshno (2019) considers a general class of queueing rules, and finds FCFS to be suboptimal. Both papers assume complete information, i.e., neither considers information design. Again, the current paper is differentiated by its consideration of broad incentive issues (i.e., the incentive to stay in, not just to join, a queue) and a general class of queueing rules as well as information design. The fact that we draw a different conclusion on the optimal queueing rule—namely, FCFS—relative to Leshno (2019) is attributed to our ability to combine information design with the choice of a queueing rule (see Section 7 for further discussion).

4 Optimality of the Cutoff Policy

The designer's problem $[P]$ is in general difficult to solve. Instead, we consider the following relaxed problem:

$[P']$
$$\max_{p\in\Delta(\mathbb{Z}_+)} W(p)$$

subject to

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

$$\lambda_k p_k - \mu_{k+1} p_{k+1} \geq 0, \forall k. \tag{B'}$$

Here, the planner maximizes the designer's objective subject only to individual rationality (IR) and a weakening (B') of the balance equation (B). The problem constitutes a linear program (LP) involving an infinite-dimensional measure p.

It is clear that $[P']$ is a relaxation of $[P]$. First, (IR) must be implied by (IC). If the former condition fails, the agents do not ex ante break even. Then, there must exist *some* agent and *some* belief induced by that mechanism such that the agent with that belief would not wish to join a queue when called upon to do so. Hence, (IC) would fail.[26] Next, since the $y_{k,\ell}$ are nonnegative and $z_{k,\ell}, x_{k,\ell}$ are all in $[0,1]$, (B) implies (B'). Let $\mathcal{W}^*$ denote the supremum of the value of program $[P']$. Then, whenever $\mathcal{W}^* < \infty$, we must have $\mathcal{W}^* \geq \mathcal{W}$.

The program $[P']$ is interesting in its own right: it can be interpreted as the problem facing a planner who chooses the invariant distribution p directly to maximize her objective, simply facing the primitive process (λ, μ), but disregarding agents' incentives altogether, except for guaranteeing some minimal payoff for them. Ultimately, however, we are interested in $[P']$ as an analytical tool for characterizing an optimal queue design that solves $[P]$, since a solution to this relaxed program $[P']$ may be attained by a mix of policy tools (x, y, z, q, I).

Indeed, our ultimate goal is to prove such a policy mix exists, which will then imply that it optimally solves $[P]$, the real object of interest. The analysis will involve demonstrating three claims: (i) an optimal solution p^* to $[P']$ exists, (ii) there exists an entry/exit policy (x^*, y^*, z^*) that generates p^*; namely, the associated outcome (x^*, y^*, z^*, p^*) satisfies (B); and (iii) there exists a queueing/information policy (q^*, I^*) that implements the optimal outcome (x^*, y^*, z^*, p^*), meaning $(x^*, y^*, z^*, p^*, q^*, I^*)$ satisfies (IC). Since $\mathcal{W}^* \geq \mathcal{W}$, it would then follow that $(x^*, y^*, z^*, p^*, q^*, I^*)$ solves $[P]$. The remainder of this section will address (i) and (ii), while claim (iii) will be taken up in the next section.

With respect to (ii), we establish not only that the optimal p^* is well-defined and can be implemented by some entry/exit policy (x^*, y^*, z^*), but also that, under the regularity of the service process (λ, μ), (x^*, y^*, z^*) takes a particularly intuitive form:

Definition 2. An entry/exit policy (x, y, z) is a **cutoff policy** if there exists $K \in \mathbb{Z}_+ \cup \{+\infty\}$ such that $x_k = 1$ for all $k = 0, 1, ..., K-2$, $x_{K-1} \in (0, 1]$, and $x_k = 0$ for all $k \geq K$ and that $y_{k,\ell} = z_{k,\ell} = 0$ for all k, ℓ.

[26] This can be shown more precisely. Fix any (x, y, z, p, q, I) that satisfies (IC_0). Aggregating (IC_0) across all beliefs $\gamma^0 \in \text{supp}(I_0)$, we get

$$\int_{\gamma^0 \in \text{supp}(I_0)} \sum_{k,\ell} \gamma^0_{k,\ell} [V\sigma_{k,\ell} - C\tau_{k,\ell}] I_0(d\gamma^0) \geq 0.$$

Clearly, the *ex ante* probability of receiving service, $\int_{\gamma^0 \in \text{supp}(I_0)} \sum_{k,\ell} \gamma^0_{k,\ell} \cdot \sigma_{k,\ell} I_0(d\gamma^0)$, equals $\sum_k p_k \mu_k / [\sum_k p_k \lambda_k x_k]$—the average rate of receiving service divided by the average rate of entering the queue at p. Next, by Little's law, the *ex ante* expected waiting time, $\int_{\gamma^0 \in \text{supp}(I_0)} \sum_{k,\ell} \gamma^0_{k,\ell} \tau_{k,\ell} I_0(d\gamma^0)$, equals $\sum_k p_k k / [\sum_k p_k \lambda_k x_k]$—the average queue length divided by the average entry rate. Substituting these two expressions and simplifying the terms, the above inequality implies (IR).

In words, under a cutoff policy p, the designer sets a maximum queue length K and recommends that an arriving agent join a queue as long as $k \leq K-1$ and that those who join the queue stay in the queue until they are served. Thus, no agent is diverted away or induced to abandon his queue once he has joined it. It is possible that $x_{K-1} \in (0,1)$, in which case the K-th entrant may be randomly rationed.[27] Although a cutoff policy seems natural, it may not arise without policy intervention. For instance, the aforementioned unregulated/unobservable $M/M/1$ queue (governed by FCFS) will not exhibit a cutoff structure if agents randomize on entry with an interior probability $e \in (0,1)$. (Recall from Footnote 22 that such a mixing is an equilibrium if $(\mu-\lambda)V < C$.)

Observe that an invariant distribution p is generated by a cutoff policy (x,y,z) with maximal length K (potentially infinite) if and only if $\text{supp}(p) = \{0,...,K\}$ and (B') binds for all $k = 0,...,K-2$ and holds for $k = K-1$ (with weak inequality).[28] Our cutoff characterization, which we present next, will focus on establishing this latter feature. All proofs of the paper are relegated to the Appendix.

Theorem 1. An optimal solution of $[P']$ exists. If μ is regular, there is an optimal solution to $[P']$ implemented by a cutoff policy with maximal queue length $K^* \geq \arg\max_k \mu_k V - kC$.

The intuition behind the result can be traced to the fundamental reason for queueing agents. Although keeping agents in the queue beyond the service capacity may at first glance appear wasteful, it serves a valuable purpose as an "insurance" against the risk of the service capacity going idle and wasted when too few agents show up in the future. This insurance benefit falls as more agents are added to the queue, however, due the concavity of μ_k in k (as required by a regular service process), while the cost of entailing additional wait time does not fall as more agents are added, so the cutoff policy emerges as optimal. This can be shown more precisely when the coefficient of p_k in the objective as well as in (IR) is decreasing in $k \geq 1$—this occurs, for instance, if μ_k is constant in k as in the $M/M/1$ queue model.[29] In this case, one can increase the value of the objective and relax (IR) by shifting probability mass p_k toward lower values of k in the sense of first-order stochastic dominance

[27] While we assume $y_{k,\ell} = z_{k,\ell} = 0$ for all k,ℓ, this is just a convenient normalization. If $x_{K-1} \in (0,1)$ in a cutoff policy, the same p^* can be implemented by any (x',y',z') such that $x'_{K-1} = \frac{\mu_K + \sum_\ell y'_{K,\ell}}{\mu_K(1-\sum_\ell z'_{K-1,\ell})} x_{K-1}$; see (B). In this sense, the reader should interpret the cutoff policy as an equivalence class involving a set of such pairs. This means that while it is unnecessary to induce an agent to exit from a queue after he joins it, doing so when the queue length is $K-1$ (and $x_{K-1} \in (0,1)$) or K is consistent with a cutoff policy. In other words, encouraging a customer to come back later is not at odds with a cutoff policy.

[28] The characterization follows from the following observation. Assuming $p_{k+1} > 0$, (B') binds at k if and only if (B) is satisfied for $x_k = 1, z_{k,\ell} = 0$ for all $\ell = 1,...,k$ and $y_{k+1,\ell} = 0$ for all $\ell = 1,...,k+1$.

[29] The coefficient is zero at $k = 0$, so it is undesirable to reduce the queue length all the way down to zero.

while keeping constant the mass at state 0.[30] This suggests that, for some $K \in \mathbb{Z}_+$, (B') must bind for all $k = 0, ..., K-2$ and $p_k = 0$ for all $k > K$, as required by the cutoff policy. This logic suggests that the mixing equilibrium in the aforementioned unregulated/unobserved $M/M/1$ queue is suboptimal for any welfare weight α. The agents will be collectively better off if agents are encouraged to enter fully if $k < K^*$ but never if $k \geq K^*$, for some K^*. In short, a cutoff policy is optimal.

To show this for a general "regular" service process, we use the fact that the coefficient of p_k in the Lagrangian function—$f(k) \triangleq \mu_k((1-\alpha)R + (\alpha+\xi)V) - (\alpha+\xi)ck$, where $\xi \geq 0$ is the Lagrangian multiplier for (IR),—is single-peaked when μ_k is regular: namely, f is increasing when $k < k^*$, is constant when $k^* \leq k \leq k^{**}$, and is decreasing when $k > k^{**}$, where k^* is possibly zero and k^{**} is possibly infinite.[31] The single peakedness of f means that on the increasing region, one would like to put the largest possible mass on a higher k within that region—a goal that is accomplished by binding (B') for k in that region. For k in the decreasing region, the intuition provided above applies.

The proof requires some care since the Lagrangian characterization of optima may not be valid in an infinite dimensional LP. Our approach follows several steps. First, we consider a finite K-dimensional LP—one in which $p_k = 0$ for all $k > K$—and prove by using the Lagrangian method that its optimal solution $p^K = (p_0^K, ..., p_K^K)$ exhibits a cutoff policy. Second, we show that an optimal solution $\overline{p} = (\overline{p}_0, ..., \overline{p}_\infty)$ to $[P']$ exists. This follows from the observation that the set of p's satisfying (IR) is closed and its elements have a vanishing tail sum so that the feasible set of solutions forms a "tight" set of measures and is therefore sequentially compact, by Prokhorov's theorem. This, together with the upper semi-continuity of the objective, gives us the existence of an optimal solution. Third, the same observation means that the value of the (normalized) K-truncation of $\overline{p}$, $\overline{p}^K \triangleq (\frac{\overline{p}_0}{\sum_{i=0}^K \overline{p}_i}, ..., \frac{\overline{p}_K}{\sum_{i=0}^K \overline{p}_i})$, which is feasible for $[P']$, converges to $\mathcal{W}^*$ as $K \to \infty$. Fourth, by definition, p^K attains a weakly higher value than $\overline{p}^K$ for each K, so its limit p^* as $K \to \infty$ attains $\mathcal{W}^*$. Finally, the set of feasible p's exhibiting a cutoff policy is closed, so the limit p^*, which is optimal, retains the cutoff structure.

[30] The only remaining issue then is that the LP, which is infinite dimensional, admits an optimal solution. The space of p's satisfying (IR) and with a cutoff structure is "tight" and is thus sequentially compact by the Prokhorov's theorem. Lingenbrink and Iyer (2019) uses this method to arrive at a similar conclusion. However, this simple approach does not work for our theorem due to the greater generality.

[31] In either case, f is monotonic.

5 Optimality of FCFS with No Information

In this section, we establish the general optimality of FCFS with no information. From now on, we assume that the service process is regular (i.e., part (i) of Definition 1). Then, by Theorem 1, the optimal solution p^* to $[P']$ is implemented by a cutoff policy (x^*, y^*, z^*) with a maximal queue length $K^* \in \mathbb{Z}_+ \cup \{+\infty\}$. Recall that the optimal cutoff policy may involve random entry at $k = K^* - 1$; recall that $x^*_{K^*-1} \in (0, 1]$ stands for the optimal randomization at $k = K^* - 1$. To avoid the trivial case, we assume that $K^* > 1$. Further, recall that the optimal cutoff policy has $y^*_{k,\ell}, z^*_{k,\ell}$ all equal to 0. To ease on notations, we sometimes simply write this optimal cutoff policy as x^*, and similarly, write the optimal policy (x^*, y^*, z^*, p^*) as (x^*, p^*).

In what follows, we fix the optimal outcome x^* and the maximal queue length $K^* > 1$. We will show that FCFS, together with an optimal information design, implements (x^*, p^*); namely, (IC) holds under that policy. Since $[P']$ is a relaxation of $[P]$, this will prove that the identified policy mix solves $[P]$.

We denote the first-come-first-served (FCFS) rule by q^*, where, as defined before, the service rate is given by $q^*_{k,\ell} = \mu_\ell - \mu_{\ell-1} \triangleq q^*_\ell$ for each (k, ℓ) with $k \geq \ell$. Not surprisingly, under FCFS the expected waiting time depends only on one's queue position ℓ, so we use τ^*_ℓ to denote the expected waiting time for an agent with queue position ℓ. Given the primitives, this can be pinned down exactly.

Lemma 1. For any $\ell = 1, ..., K^*$, $\tau^*_\ell = \ell/\mu_\ell$. τ^*_ℓ is nondecreasing in ℓ. If $2\mu_1 > \mu_2$, then τ^*_ℓ is strictly increasing in ℓ.

We next introduce the information rule. We call an information rule $I^* \in \mathcal{I}$ **no information** if *no information is provided to each agent both at the time of joining the queue and after joining the queue, beyond what he can infer from the recommendations to join or stay in the queue.* This means that when he joins the queue, he forms a belief about his position ℓ, or the length of queue, based on the invariant distribution and the recommendation to join the queue. From then on, he updates the belief about his queue position at each $t > 0$ according to Bayes rule without any further information (given that he is recommended to stay from then on). In practice, the no information rule can be implemented by sending a message consisting of either "join" or "leave," or by providing a coarse (i.e., binary) estimate of the "expected" waiting time, to an arriving agent.

Given the cutoff policy x^*, the queueing and information rules (q^*, I^*), the incentive

constraint at time t is given by

$$(IC_t) \qquad V - C\sum_{\ell=1}^{K^*} \tilde{\gamma}_\ell^t \cdot \tau_\ell^* \geq 0,$$

where $\tilde{\gamma}^t = (\tilde{\gamma}_1^t, ..., \tilde{\gamma}_{K^*}^t) \in \Delta(\{1, ..., K^*\})$ is the belief on his position in the queue after spending time t on the queue.[32] Since the expected waiting time depends only on one's position, the belief on other variables such as the queue length k does not affect the agent's incentive to join or stay in the queue.

Given the information rule I^*, the belief at the time of joining the queue must be:

$$\tilde{\gamma}_\ell^0 = \begin{cases} \frac{p_{\ell-1}^* \tilde{\lambda}_{\ell-1}}{\sum_{i=0}^{K^*-1} p_i^* \tilde{\lambda}_i} & \text{if } \ell = 1, ..., K^* \\ 0 & \text{if } \ell > K^*, \end{cases} \tag{1}$$

where $\tilde{\lambda}_k$ is an "effective" arrival rate given by: $\tilde{\lambda}_k \triangleq \lambda_k$ for $k = 0, ..., K^* - 2$, and $\tilde{\lambda}_{K^*-1} \triangleq x_{K^*-1}^* \lambda_{K^*-1}$. This formulation rests on the consistency of an agent's belief about the rule in place—namely, (x^*, q^*, I^*)—, as well as the invariant distribution p^*. Specifically, (1) computes the probability of an agent occupying position ℓ conditional on entering the queue. Its numerator is the probability that an agent joins the queue in state $\ell - 1$, which equals the probability of there being $\ell - 1$ agents already in the queue multiplied by the probability of entry per unit time in that state $\tilde{\lambda}_{\ell-1}$.[33] Its denominator is the total probability of entering the queue per unit time.

It is easy to show that the candidate policy (q^*, I^*) provides the agents with incentives to enter the queue, i.e., it satisfies (IC_0).[34] As noted in the Introduction, (IC_0) is not particularly demanding, for all other queueing rules also satisfy the condition under no information I^*. It will prove much more challenging to satisfy (IC_t) for $t > 0$. To examine whether this condition holds under (q^*, I^*), we need to study how an agent's belief evolves once he joins the queue. Since no agent is recommended to abandon the queue, (IC_t) for $t > 0$ boils down to whether agents' beliefs about their queue positions become (at least

[32] Note that $\sigma_{k,\ell} = 1$ for all k, ℓ since, by definition of the cutoff policy, the designer never removes agents from the queue.

[33] The formula in (1) is justified by the logic of PASTA (Wolff (1982)). More precisely, one can think of an agent's *unconditional* (i.e., before conditioning on her arrival) belief about the state as given by the invariant distribution over states. The conditional belief is then obtained by conditioning based on the entry $\{x_k\}$ policy as well as the heterogeneity in the arrival rate $\{\lambda_k\}$.

[34] Footnote 26 shows how this is implied by (IR). See also online appendix S.4 for an alternative argument using directly the characterization of waiting times under FCFS given in Lemma 1.

weakly) more favorable—or put more probability at lower ℓ's—as time passes.

Suppose that an agent has belief $\tilde{\gamma}^t$ after spending time $t \geq 0$ in the queue. By Bayes rule, after time $t + dt$, his belief is updated to:

$$\tilde{\gamma}_\ell^{t+dt} = \frac{\tilde{\gamma}_\ell^t(1 - \sum_{i=1}^{\ell} q_i^* dt) + \tilde{\gamma}_{\ell+1}^t \sum_{i=1}^{\ell} q_i^* dt}{\sum_{i=1}^{K^*} \tilde{\gamma}_i^t (1 - q_i^* dt)} + o(dt).$$

The numerator is the probability that his queue position is ℓ after staying in the queue for length $t + dt$ of time. This event occurs if either (i) the agent already has position ℓ in the queue at time t and none of the agents ahead of him and himself have been served during time increment dt; or (ii) if he has position $\ell + 1$ at t and one agent ahead of him is served by $t + dt$.[35] The denominator in turn gives the total probability that the agent has not been served by time t. Hence, given that an agent has not been served by t, the above expression gives the conditional belief that his position in the queue is ℓ at time $t + dt$. We can use the feasibility requirement $\sum_{i=1}^{\ell} q_i^* = \mu_\ell$ to rewrite the belief updating rule as follows:

$$\tilde{\gamma}_\ell^{t+dt} = \frac{(1 - \mu_\ell dt)\tilde{\gamma}_\ell^t + \mu_\ell dt \tilde{\gamma}_{\ell+1}^t}{\sum_{i=1}^{K^*} \tilde{\gamma}_i^t (1 - q_i^* dt)} + o(dt). \tag{2}$$

We now study how the belief updates dynamically over time under (q^*, I^*). The statistic we focus on is the **likelihood ratio** $r_\ell^t \triangleq \frac{\tilde{\gamma}_\ell^t}{\tilde{\gamma}_{\ell-1}^t}$ in beliefs of being in queue position ℓ to being in queue position $\ell - 1$ after spending time t on the queue. One can use (2) to derive a system of ordinary differential equations (ODEs) on the likelihood ratios:

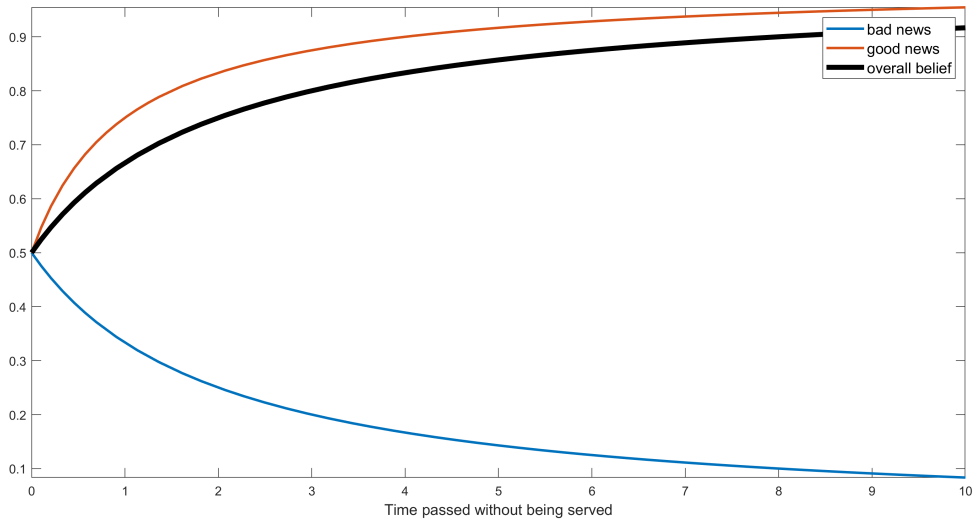
$$\dot{r}_\ell^t = r_\ell^t \left(\mu_{\ell-1} - \mu_\ell - \mu_{\ell-1} r_\ell^t + \mu_\ell r_{\ell+1}^t \right), \tag{3}$$

where $\ell = 2, ..., K^*$. Further, the invariant distribution p^* can be used to obtain the boundary conditions, $r_\ell^0 = \frac{\tilde{\lambda}_{\ell-1}}{\mu_{\ell-1}}$, for $\ell = 2, ..., K^*$, where we recall that $\tilde{\lambda}_k$ is the effective arrival rate. Appendix B.2 derives this system of ODEs and establishes existence of a unique solution.

We will argue that the regularity of the primitive process (in particular part (ii) of Definition 1) is sufficient for these likelihood ratios—the solution to the above ODEs—to decline over time, meaning one's belief about his position becomes progressively favorable under (q^*, I^*). At first glance, this seems obvious under FCFS: conditional on starting at any position ℓ at $t = 0$, an agent's queue position can *only* improve as time passes. Since the agent begins with no information, however, this is not the only event about which the

[35] The probability of multiple agents ahead of him being served during $[t, t + dt)$ has a lower order of magnitude denoted by $o(dt)$.

Figure 1: Belief about position $\ell = 1$



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

agent updates his beliefs. The agent is also updating his belief about his initial position ℓ. On this account, however, the time t spent on the queue is "bad" news, as it suggests that he may have been too optimistic about his position initially, causing him to revise his initial queue position pessimistically as time passes.

Figure 1 displays these two competing effects in an $M/M/1$ queue with $K^* = 2$. Its top graph depicts the good news effect: an agent's belief about being at the top position ($\ell = 1$) is improving over time *when the belief about his initial queue position is held fixed at the prior*. The bottom graph depicts the bad news effect: the belief about his initial queue position being $\ell = 1$ falls over time. The middle graph displays the overall evolution of the belief—namely about $\ell = 1$ conditional on not being served by t. Its increase means that the former "position-improvement" effect dominates the worsening posterior about the initial position.

The regularity of the primitive process is sufficient for the good news effect to dominate the bad news effect:

Lemma 2. Assume that the primitive process (λ, μ) is regular. Then, for all $\ell \in \{2, ..., K^*\}$, r_ℓ^t is nonincreasing in t for all $t \geq 0$.

Intuitively, regularity ensures that the arrival rate does not rise faster than the service rate as the queue length increases. This keeps the adverse inference about initial position

from worsening one's belief about the residual waiting time.[36] We are now in a position to state our main theorem.

Theorem 2. Assume that the primitive process is regular. Then, FCFS with no information (q^*, I^*) implements the optimal outcome (x^*, p^*) where p^* solves $[P']$. Consequently, (x^*, q^*, I^*) is an optimal solution of $[P]$.

Proof. This theorem is a consequence of Lemma 2. Indeed, it suffices to prove that, under FCFS with no information, (IC_t) holds for all $t \geq 0$. Note first that, as we already stated (see Lemma S4 in the online appendix), (IC_0) holds. Next consider (IC_t) for any $t > 0$. Lemma 2 proves that $r_\ell^t \leq r_\ell^0$ for each ℓ. Since τ_ℓ^* is nondecreasing in ℓ (Lemma 1), this means that

$$\sum_{\ell=1}^{K^*} \tilde{\gamma}_\ell^t \cdot \tau_\ell^* \leq \sum_{\ell=1}^{K^*} \tilde{\gamma}_\ell^0 \cdot \tau_\ell^*,$$

so we have

$$V - C \sum_{\ell=1}^{K^*} \tilde{\gamma}_\ell^t \cdot \tau_\ell^* \geq V - C \sum_{\ell=1}^{K^*} \tilde{\gamma}_\ell^0 \cdot \tau_\ell^* \geq 0,$$

where the last inequality follows from (IC_0) being satisfied. Hence, (IC_t) holds for any $t > 0$. ∎

To the extent that regularity is extremely mild, one may view this theorem as suggesting that the combination of FCFS and No Information is optimal in a broad set of circumstances. Nevertheless, the dynamic incentives provided by FCFS, or the role played by regularity conditions, should not be taken for granted. Intuitively, with the failure of regularity, delay is more of a signal about the initial queue length being long than about predecessors having been served, and thus one's belief, and therefore one's incentive to stay in the queue, may get worse over time. One can indeed build somewhat artificial examples where regularity fails and where the optimal solution to $[P']$ is not implementable under (q^*, I^*).

6 Necessity of FCFS for Optimality in a Rich Domain

We have shown that FCFS with no information is optimal in all regular environments. This result raises a question of whether a different queueing/information policy may be also

[36] Our proof method differs from the standard queuing analysis which focuses on the increasing or decreasing hazard rate of an agent's waiting time in the $M/M/1$ and $M/M/c$ queue models (see Gnedenko and Kovalenko (1989)). Analyzing the evolution of hazard rates appears difficult in our general Markovian model. We believe that the current method that tracks the evolution of posterior beliefs are of independent analytical interest for queuing theory.

optimal in some (or all) environments. Indeed, one can show that, when $\alpha = 1$, FCFS is optimal under full information. In appendix Section S.6, we generalize Naor (1969) to show that FCFS can provide sufficient incentives for queueing under full information if $\alpha = 1$, so the optimum can be achieved with the entry controlled appropriately.[37] Other rules such as LCFS and LIEW (Load Independent Expected Waiting) are known to perform well in some situations. For instance, Hassin (1985) and Su and Zenios (2004)) have shown that versions of LCFS, possibly *with preemption* (i.e., where a newly arriving agent replaces one under service), are optimal under full information when $\alpha = 1$. In LCFS with preemption, once the maximal queue length K^* is reached, new agents still enter the queue, but old agents (currently being served) are removed. In a different model with overloaded queues, again with $\alpha = 1$ and full information, and absent dynamic incentive issue, Leshno (2019) has shown LIEW to be optimal in a class of buffer-queue mechanism. One may wonder how these mechanisms perform more generally with $\alpha < 1$ in the presence of dynamic incentive constraints.

We show below that none of these queueing rules can be optimal in all regular environments. Instead of studying these queueing rules separately, we consider *all* feasible queueing rules and show that FCFS is the only queueing rule that is optimal for all (regular) queueing environments. Or equivalently, for any queueing rule differing from FCFS, we exhibit a (regular) environment in which this rule is suboptimal under any information rule. For this purpose, we focus on the most canonical and simplest environment: the M/M/1 environment in which a uniquely optimal solution to $[P']$ involves (i) $K^* = 2$, (ii) no rationing when $k = K^* - 1 = 1$, and (iii) a binding (IR). Specifically, we fix any service rate $\mu > 0$. We then consider a sufficiently small arrival rate λ by letting it approach zero. When we do this, we simultaneously adjust the values of (V, C, α) to ensure that properties (i), (ii), and (iii) continue to hold.[38]

Since $K^* = 2$, there are only three relevant "states," $(k, \ell) = (1, 1), (2, 1), (2, 2)$, based on the queue length k and one's queue position ℓ. Hence, we can denote a queueing rule by $q = (q_{1,1}, q_{2,1}, q_{2,2})$. Recall that FCFS corresponds to $q^* = (\mu, \mu, 0)$. For any feasible work-conserving queueing rule, we must have $q_{1,1} = \mu$ and $q_{2,1} + q_{2,2} = \mu$. Hence, a queueing rule $q \in \mathcal{Q}$ can differ from FCFS q^* if and only if $q_{2,1} < \mu$, or equivalently, $q_{2,2} > 0$. While

[37] In particular, as proved in Section S.6, a dynamic incentive is not a problem under complete information with FCFS, as long as agents have sufficient incentives for queueing, which is the case when $\alpha = 1$.

[38] These requirements can be met by choosing $V/C = \frac{2\lambda+\mu}{(\lambda+\mu)\mu}$ and $\alpha = 0$. In that case, there is a unique optimal solution p to $[P']$ and any outcome (x, y, z) implementing p satisfies (i), (ii) and (iii). Note that assumption (iii) precludes $\alpha = 1$ under which (IR) is non-binding at the optimal policy as long as the value of the objective may be strictly positive.

most standard queueing rules—such as LCFS or SIRO—do not depend on the arrival rate λ,[39] LIEW as defined in Leshno (2019) does. In an attempt to maximize queueing incentives for incoming agents, LIEW equalizes their expected waiting times across all possible queue lengths they may encounter upon arrival—in our context between an empty queue and a queue with one agent. To equalize waiting time across queue lengths, an agent who enters an empty queue must be later "penalized" in service priority when another agent joins, to counterbalance the fast service she initially receives when there is no other agent. The extent of this penalization must then depend on the arrival rate, causing q to depend on λ.[40] Even under LIEW, however, $q_{2,2}$ is bounded away from 0 for all values of λ. This motivates the following definition. We say that a queueing rule **differs from FCFS** if $q_{2,2}$ is bounded away from 0 for all possible values of λ.[41] All queueing rules studied in the literature such as SIRO, LCFS or LIEW differ from FCFS in this sense. We are now in a position to state the main result of this section:

Theorem 3. Fix any queuing rule q that differs from FCFS. Then, there exists an $M/M/1$ queue with values $(V, C, \alpha, \lambda, \mu)$ such that the queueing rule q fails (IC_t) for some $t > 0$ under any information policy. Hence, q cannot implement the optimal cutoff policy under any information policy.

The intuition for this result is most clear under LCFS. Under this rule, an agent loses his service priority when another agent enters. So, if an agent were initially indifferent to queueing (as implied by a binding (IR)), he will definitely wish to abandon the queue once a new agent enters. Consequently, (IC_t) fails at time t when new entry occurs if he had full information. Even under no information, as time passes without getting served, an agent will suspect that a new entry is increasingly likely and he will lose his priority as a consequence. This feature destroys his dynamic incentive. A similar problem arises with LIEW. Recall that equalization of waiting time across queue lengths means that an agent who enters an empty queue must be "penalized" in service priority later when a new agent enters. This very feature undermines the dynamic incentive of an agent. The root cause of the problem under these rules is: $q_{2,1} < q_{1,1} = \mu$—namely, the loss of priority an agent suffers when a new

[39] Naturally, a queueing rule will change as the value of μ changes, but recall we have fixed the value of μ.

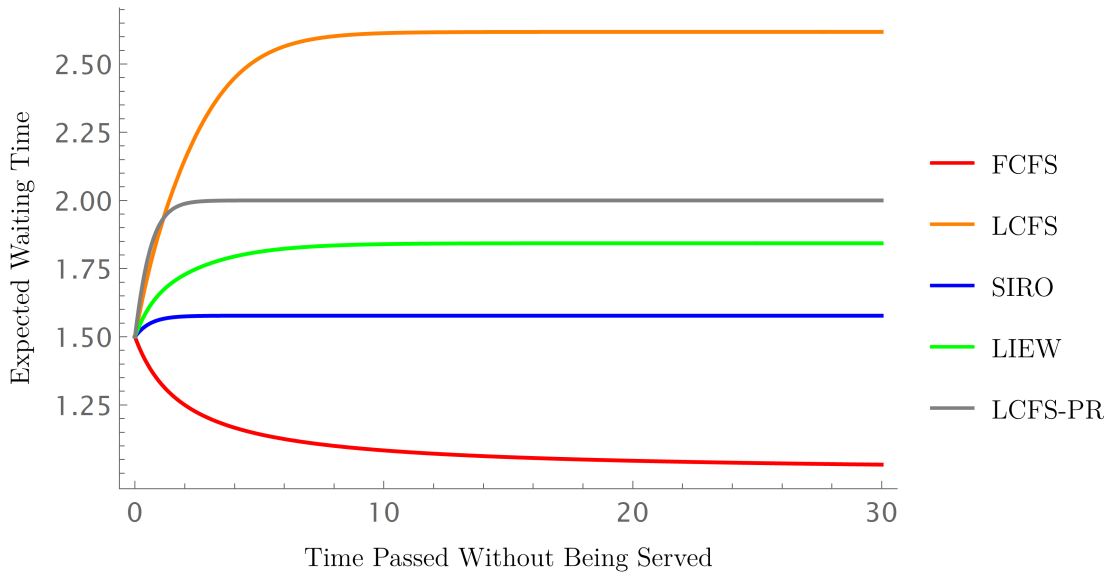
[40] Indeed, as the arrival rate λ increases, the agent will move to state $(2,1)$ more quickly and so the "penalty" can be smaller (i.e., $q_{2,1}$ can be larger). More precisely, assuming that agents can never join the queue when the queue length is equal to 2, LIEW has $q_{2,1} = \frac{\lambda}{2\lambda+\mu}\mu$ and $q_{2,2} = \frac{\lambda+\mu}{2\lambda+\mu}\mu$. Note that $q_{2,2} \geq \frac{1}{2}\mu$.

[41] Formally, a queueing rule q assigns, to each of our M/M/1 environments, characterized by (μ, λ), a vector $q(\mu,\lambda) = \{q_{1,1}(\mu,\lambda), q_{2,1}(\mu,\lambda), q_{2,2}(\mu,\lambda)\}$ of service rates. Our assumption is as follows: for any given μ, there is some $\eta > 0$ such that $q_{2,2}(\mu,\lambda) > \eta$ for all $\lambda > 0$.

agent arrives. Although LCFS and LIEW are extreme in this regard, any rule that assigns $q_{2,1} < \mu = q_{1,1}$, including SIRO, suffers from the same fundamental issue.

To illustrate, Figure 2 plots the expected waiting times against time elapsed on the queue under five queueing disciplines: FCFS, SIRO, LIEW, LCFS, and LCFS-PR, where LCFS-PR is the LCFS with preemption that Hassin (1985) argued to be optimal when $\alpha = 1$. As is clearly seen, and consistent with Theorem 3, as time passes, an agent in the queue expects to wait increasingly longer under all these disciplines, except for FCFS under which his expected wait decreases.

Figure 2: Expected waiting times under alternative values of q.



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

7 Concluding Remarks

While we have focused on a canonical queueing model, the insights we obtain appear general and apply beyond our model. Here we discuss how one may extend our analysis to other settings of potential interest.

Dynamic two-sided matching. A topic closely related to queueing is dynamic matching; see Akbarpour, Li, and Gharan (2020), Akbarpour, Combe, Hiller, Shimer, and Tercieux (2020), Baccara, Lee, and Yariv (2020), Leshno (2019), Doval and Szentes (2018), and Ashlagi, Nikzad, and Strack (2019), among others. The primary focus of this literature is the optimal timing of matching and assignment, rather than queueing incentives. Exceptions

are Leshno (2019) and Baccara, Lee, and Yariv (2020), who study incentives by two different types of agents for queueing to match with either two different types of objects (e.g., housing) or agents. In such a model, efficiency calls for accumulating agents in a queue until a right type of object or agent arrives, to avoid mismatching. Leshno (2019) assumes overloaded demand so that the planner wishes to incentivize the agents to queue as much as possible, and demonstrates that, given complete information, SIRO outperforms FCFS in this regard, and LIEW outperforms all other mechanisms. Despite the ostensible difference relative to our model, Section S.7 in the online appendix shows that our analysis applies without much modification to this model, and points out that the main results from Leshno (2019) rest crucially on his full information assumption. With optimal information design, the FCFS could do just as well as any other mechanism, including LIEW, in incentivizing agents to enter a queue. If one includes the dynamic incentive problem, which Leshno (2019) does not consider,[42] then FCFS does strictly better than other queueing disciplines. Baccara, Lee, and Yariv (2020)'s model is similar to that of Leshno (2019), except that there are agents on both sides. Hence, our main insight in Theorem 2 applies, except for one difference. Unlike Leshno (2019), agents' incentives to enter a queue may be excessive under FCFS with full information. While this is an issue in their decentralized matching, in our setting the designer can easily solve the problem by preventing an agent from entering a queue, as is often done in practice. Meanwhile, they also show that a queueing discipline admits insufficient entry under LCFS. In that case, information design can be useful even with their assumption. In addition, they too do not consider dynamic incentives, for which our analysis will prove useful. Moving forward, both information design and dynamic incentives, largely missing in this literature, will add interesting new elements to incorporate for the dynamic matching research.

Monopolist problem with endogenously set fee. If $\alpha = 0$, one could interpret the service provider/designer as a monopolist who provides the service. We treated the fee (R) as exogenous, representing the shadow value of addressing customer needs. In many contexts, however, one may think of this profit as a monetary fee collected and set by the

[42]The dynamic incentive issue does not arise in SIRO or FCFS under complete information: any agent who joins the queue will have the incentive to stay in the queue. Recall, however, that neither discipline would implement the optimum under complete information. Under *no information* (which is optimal), dynamic incentives will be an issue. Although an agent may not leave the queue and unilaterally "claim" a mismatched object, which is presumably under the designer's control, he/she may leave the queue without claiming any object. If the value of outright exit is not very low (e.g., in comparison with the value of a mismatched object), then the dynamic incentives will matter just as they do in our model. Specifically, both SIRO (under no information) and LIEW (under *any* information) would be vulnerable to reneging, and cannot implement the optimal outcome, as stated in our Theorem 3.

service provider. With this interpretation, the designer/monopolist will choose this fee R and the net surplus of customers for service now equals $V - R$. Hence, we can rewrite problem $[P]$ assuming that R is part of the decision variables and incorporating the new net surplus of customers into the (IC) condition. Our framework can be easily adapted to this environment. Indeed, one can write $[P']$ assuming that the designer chooses both the invariant distribution and the fee level. Given the optimal choice of fee, the rest of the proof applies without any modification. Namely, a cutoff policy is optimal. Clearly, Lemma 2 must still hold, so Theorem 2 extends to this context. Incidentally, one can also characterize the optimal fee in this context. Intuitively, when choosing the fee, the monopolist should now consider both its impact on his profit and also on the incentives of agents to join (and stay) in the queue. For instance, a higher fee directly increases profit but may also discourage agents from joining the queue, which increases the probability that the servers go idle and thus jeopardizes the opportunity to collect the fee. The fee is chosen optimally to balance the tradeoff.

Time preferences. The current model follows the standard convention of the queueing literature in assuming linear waiting cost. This convention has the usual benefit in admitting analytical tractability and easy comparability with existing queueing models. Another, more important, benefit in the current model is that it isolates the effect of dynamic incentives generated by alternative queueing rules. Given linear waiting costs, differences in waiting time distributions across alternative queueing rules do not matter when one focuses, as existing queueing models do, on static incentives for queueing (i.e., incentives to join a queue), but they do matter when one considers, as we do, dynamic incentives for queueing. In this regard, what helps FCFS is its feature that its waiting time involves least amount of dispersion in comparison with other queueing rules;[43] this helps to minimize the adverse updating from a "missing" an early service. For this reason, introducing nonlinear time preferences will confound this effect of dynamic incentives, since it will make the waiting-time distribution under alternative queueing rules payoff-relevant. A reasonable conjecture is, though, that risk-averse time preferences will reinforce the optimality of FCFS whereas risk-loving time preferences (such as exponential discounting) will counteract it.

Heterogenous preferences. Following the standard queueing models, we have assumed that agents have homogeneous preferences. For added realism, however, it is important to allow agents to differ in their waiting costs, value of service, or in their service requirements. Such heterogeneities will introduce the need by the designer to treat agents differently based

[43] This feature reflects the fairness property of FCFS that an agent who arrives first gets served first (see Shanthikumar and Sumita (1987)).

on their types, for instance prioritizing service toward those agents with high waiting costs, high value of service and small service requirements.[44] This will again confound the analysis by making allocation of service priority directly payoff-relevant, above and beyond making it relevant from the perspective of dynamic incentives—the central focus of the current study. In particular, if the agents' characteristics are unobservable, one must deal with additional incentive issues with screening agents based on this additional informational asymmetry. Such an extension is therefore beyond the scope of the current paper. Nevertheless, one may conjecture that the main logic and thrust of the current paper will extend to a model with heterogeneous preferences. At least within each type of agents, allocating service according to FCFS contributes to their dynamic incentives for queueing, and will be desirable.

We leave these and other worthy extensions of the current model for future research.

References

Akbarpour, M., J. Combe, V. Hiller, R. Shimer, and O. Tercieux (2020): "Unpaired Kidney Exchange: Overcoming Double Coincidence of Wants without Money," *NBER Working paper number 27765*. 26

Akbarpour, M., S. Li, and S. O. Gharan (2020): "Thickness and information in dynamic matching markets," *Journal of Political Economy*, 128(3), 783–815. 14, 26

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. 6, 29

Ashlagi, I., M. Faidra, and A. Nikzad (2020): "Optimal Dynamic Allocation: Simplicity through Information Design," Discussion paper, Stanford. 7

Ashlagi, I., A. Nikzad, and P. Strack (2019): "Matching in dynamic imbalanced markets," *Available at SSRN 3251632*. 26

Baccara, M., S. Lee, and L. Yariv (2020): "Optimal dynamic matching," *Theoretical Economics*, 15, 1221–1278. 15, 26, 27

Bloch, F., and D. Cantala (2017): "Dynamic assignment of objects to queuing agents," *American Economic Journal: Microeconomics*, 9, 88–122. 6

[44] See Anunrojwong, Iyer, and Manshadi (2020) for a simple model of heterogenous waiting costs—i.e., zero cost and positive costs.

Cripps, M. W., and C. D. Thomas (2019): "Strategic experimentation in queues," *Theoretical Economics*, 14, 647–708. 5

Doval, L., and B. Szentes (2018): "On the efficiency of queuing in dynamic matching," Discussion paper, Caltech and London School of Economics. 14, 26

Gnedenko, B., and I. Kovalenko (1989): *Introduction to Queueing Theory*. Birkhauser. 14, 23

Hassin, R. (1985): "On the optimality of first come last served queues," *Econometrica*, 53, 201–202. 2, 6, 9, 13, 24, 26

Hassin, R. (2016): *Rational Queueing*. CRC Press. 5

Hassin, R., and M. Haviv (1995): "Equilibrium strategies for queues with impatient customers," *Operations Research Letters*, 17, 41–45. 5

——— (2003): *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers. 5, 6, 12

Hassin, R., and A. Koshman (2017): "Profit maximization in the M/M/1 queue," *Operations Research Letters*, 45, 436–441. 6, 13

Haviv, M., and B. Oz (2016): "Regulating an observable M/M/1 queue," *Operations Research Letters*, 44, 196–198. 6

——— (2018): "Self-Regulation of an Unobservable Queue," *Management Science*, 64, 1–10. 6

Haviv, M., and Y. Ritov (2001): "Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions," *Queueing Systems*, 38, 495–508. 5

Kamenica, E., and M. Gentzkow (2011): "Bayesian persuasion," *American Economic Review*, 101, 2590–2615. 10

Kipp, M., C. Ryan, and S. Matt (2016): "The Slater Conundrum: Duality and Pricing in Infinite-Dimensional Optimization," *SIAM Journal on Optimization*, 1(26), 111–138. 32

Kittsteiner, T., and B. Moldovanu (2005): "Priority auctions and queue disciplines that depend on processing time," *Management Science*, 51, 236–248. 6

Larson, R. C. (1987): "Perspectives on queues: social justice and the psychology of queueing," *Operations Research*, 35, 895–905. 7

Leshno, J. (2019): "Dynamic matching in overloaded waiting lists," Discussion paper, SSRN Working Paper 2967011. 6, 15, 24, 25, 26, 27

Lingenbrink, D., and K. Iyer (2019): "Optimal signaling mechanisms in unobservable queues," *Operations Research*, 67, 1397–1416. 6, 13, 18

Mandelbaum, A., and N. Shimkin (2000): "A model for rational abandonments from invisible queues," *Queueing Systems*, 36, 141–173. 5, 9

Margaria, C. (2020): "Queueing to learn," Discussion paper, Boston University. 6

Myerson, R. B. (1981): "Optimal auction design," *Mathematics of Operations Research*, 6(1), 58–73. 7

Naor, P. (1969): "The regulation of queue size by levying tolls," *Econometrica*, 37, 15–24. 5, 6, 13, 24

Platz, T. T., and L. P. Østerdal (2017): "The curse of the first-in–first-out queue discipline," *Games and Economic Behavior*, 104, 165–176. 2, 6

Shaler Stidham, J. (2009): *Optimal Design of Queueing Systems*. CRC Press. 7

Shanthikumar, J. G., and U. Sumita (1987): "Convex ordering of sojourn times in single-server queues: extremal properties of FIFO and LIFO service disciplines," *Journal of Applied Probability*, 24, 737–748. 28

Sherzer, E., and Y. Kerner (2018): "Customers' abandonment strategy in an M/G/1 queue," *Queueing Systems*, 90, 65–87. 5

Simhon, E., Y. Hayel, D. Starobinski, and Q. Zhu (2016): "Optimal information disclosure policies in strategic queueing games," *Operations Research Letters*, 44, 1109–113. 6, 13

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. 2, 6, 24

Wolff, R. W. (1982): "Poisson Arrivals See Time Averages," *Operations Research*, 30, 223–231. 20

Appendix

A Proof of Theorem 1

Rewrite problem $[P']$ as:

$$[P'] \qquad \max_{p\in M} \sum_{k=0}^{\infty} p_k \left[\mu_k((1-\alpha)R + \alpha V) - \alpha Ck\right] \text{ s.t. } \sum_{k=0}^{\infty} p_k \left[\mu_k V - Ck\right] \geq 0,$$

where $M \triangleq \{p \in \Delta(\mathbb{Z}_+) : p \text{ satisfies } (B')\}$. (Recall our convention that, $\mu_0 = 0$).

Recall that an invariant distribution p is generated by a cutoff policy (x, y, z) with maximal length K if and only if $\text{supp}(p) = \{0, ..., K\}$ and (B') binds for all $k = 0, ..., K-2$ and holds for $k = K-1$ (with weak inequality). In the sequel, if a distribution p satisfies the latter feature, we will simply say that it exhibits a cutoff policy. Our goal in this section is to show that the above LP problem has an optimal solution which exhibits a cutoff policy.

Below we use a Langrangian characterization of the LP problem. Unlike finite dimensional LP problems, this characterization is not automatically valid in infinite dimensional LP problems.[A.45] In order to overcome the difficulty, we first study a finite dimensional truncation of $[P']$ where the state space contains finitely many states, say K, where K can potentially be "large". In this environment, we will show that an optimal solution p^K exhibits a cutoff policy (Appendix A.1). In a second step, we show that as K gets large, a limit point of $\{p^K\}$ is an optimal solution of $[P']$ and exhibits a cutoff policy. The proof of this second step, in essence, uses a continuity argument—and so uses fairly routine arguments. Hence it is sketched in Appendix A.2 but the formal argument is relegated to the online appendix Section S.3.

[A.45] Countably infinite linear programs (CILPs) are linear optimization problems with a countably infinite number of variables and a countably infinite number of constraints. It is well-known that many of the nice properties of finite dimensional linear programming may fail to hold in these problems. Indeed, while in finite dimensional LP problems, zero duality gap is ensured provided that the primal problem is feasible, necessary conditions for zero duality gap for CILPs are much more demanding and may often fail. See Kipp, Ryan, and Matt (2016) and references therein.

A.1 Finite dimensional analysis

In the sequel, we fix an integer $K \geq 0$. We consider the following "truncated" version of $[P']$, say $[P'_K]$

$$[P'_K] \qquad \max_{p \in M_K} \sum_{k=0}^{K} p_k \left[\mu_k((1-\alpha)R + \alpha V) - \alpha Ck\right] \text{ s.t. } \sum_{k=0}^{K} p_k \left[\mu_k V - Ck\right] \geq 0,$$

where $M_K \triangleq \{p \in \Delta(\{0,1,...,K\}) : p \text{ satisfies } (B')\}$.

Let us fix $\xi \geq 0$ and consider the problem $[\mathcal{L}_\xi]$

$$[\mathcal{L}_\xi] \qquad \max_{p \in M_K} \mathcal{L}(p, \xi)$$

where

$$\begin{aligned} \mathcal{L}(p,\xi) &\triangleq \sum_{k=0}^{K} p_k \left[\mu_k((1-\alpha)\, R + \alpha V) - \alpha Ck\right] + \xi \sum_{k=0}^{K} p_k \left[\mu_k V - Ck\right] \\ &= \sum_{k=0}^{K} p_k f(k;\xi), \end{aligned}$$

where $f(k;\xi) \triangleq \mu_k((1-\alpha)R + (\alpha+\xi)\, V) - (\alpha+\xi)\, Ck$.

The Lagrangian dual of problem $[P'_K]$ is taking the inf over $\xi \geq 0$ of the value of $[\mathcal{L}_\xi]$. Since M_K is a convex set, the problem constitutes a finite dimensional linear program, so strong duality applies. Hence, p^* is an optimal solution if and only if there is (a Lagrange multiplier) $\xi^* \geq 0$ such that (p^*, ξ^*) is a saddle point of the function $\mathcal{L}(\cdot,\cdot)$, i.e.,

$$\mathcal{L}(p, \xi^*) \leq \mathcal{L}(p^*, \xi^*) \leq \mathcal{L}(p^*, \xi)$$

for any $\xi \geq 0$ and $p \in M_K$. We fix a saddle point (p^*, ξ^*) of function $\mathcal{L}(\cdot,\cdot)$ and show that it exhibits a cutoff policy.

In this section, we will show a finite-dimensional version of Theorem 1 stated below.

Proposition A.1. If μ is regular, then there is an optimal solution for $[P'_K]$ which exhibits a cutoff policy. In addition, $p^*_k > 0$ for each $k \leq \min\{k^*, K\}$ where $k^* \triangleq \min \arg\max f(k;\xi^*)$.

In order to prove this proposition, we need to first establish several lemmas. To begin, we say a function $f : \mathbb{Z}_+ \to \mathbb{R}$ is *single-peaked* if $f(k-1) < f(k)$ for all $k \leq \min \arg\max_{k \in \mathbb{Z}_+} f(k)$ while $f(k) > f(k+1)$ for all $k \geq \max \arg\max_{k \in \mathbb{Z}_+} f(k)$. Our con-

vention is that if $\arg\max_{k\in\mathbb{Z}_+} f(k)$ is empty, then $\min\arg\max_{k\in\mathbb{Z}_+} f(k)$ is set to $+\infty$. We now show that the regularity of μ implies that $f(\cdot;\xi)$ is single-peaked.

Lemma A.3. If μ is regular, then for any $\xi \geq 0$, function $f(\cdot;\xi)$ is single-peaked.

Proof. Fix any $\xi \geq 0$. It is easily checked that $f(\cdot;\xi)$ is single-peaked if and only if $f(k;\xi) \geq (>) f(k+1;\xi)$ then $f(k';\xi) \geq (>) f(k'+1;\xi)$ for any $k' \geq k$. Assume that $f(k;\xi) \geq f(k+1;\xi)$, i.e.,

$$\mu_k((1-\alpha)\,R + (\alpha+\xi)V) - (\alpha+\xi)Ck \geq \mu_{k+1}((1-\alpha)\,R + (\alpha+\xi)V) - (\alpha+\xi)C(k+1).$$

Simple algebra shows that this is equivalent to

$$\mu_{k+1} - \mu_k \leq \frac{(\alpha+\xi)C}{(1-\alpha)\,R + (\alpha+\xi)V}.$$

Since μ is regular, $\mu_{k+1} - \mu_k$ is nonincreasing and so, for $k' \geq k$, we must have

$$\mu_{k'+1} - \mu_{k'} \leq \mu_{k+1} - \mu_k \leq \frac{(\alpha+\xi)C}{(1-\alpha)\,R + (\alpha+\xi)V}.$$

Hence, $f(k';\xi) \geq f(k'+1;\xi)$. The same argument holds to show that $f(k;\xi) > f(k+1;\xi)$ implies $f(k';\xi) > f(k'+1;\xi)$ for any $k' \geq k$. ∎

We will also use the following lemma.

Lemma A.4. Suppose

$$f(\ell;\xi^*) < f(\ell+1;\xi^*)$$

for some $\ell \leq K-1$. Then, $\lambda_\ell p^*_\ell = \mu_{\ell+1} p^*_{\ell+1}$.

Proof. Fix ℓ satisfying the properties of the lemma. Since p^* is an optimal solution of $[P'_K]$—and so satisfies (B')—we know that $\mu_{\ell+1}p^*_{\ell+1} \leq \lambda_\ell p^*_\ell$. Toward a contradiction, assume that $\mu_{\ell+1}p^*_{\ell+1} < \lambda_\ell p^*_\ell$. Now, simply consider $\hat{p}$ defined as

$$\hat{p}_k = \begin{cases} p^*_k + \varepsilon \text{ if } k = \ell+1 \\ p^*_k - \varepsilon \text{ if } k = \ell \\ p^*_k \text{ otherwise} \end{cases}$$

and note that we can choose $\varepsilon > 0$ so that $\mu_{\ell+1}\hat{p}_{\ell+1} = \lambda_\ell \hat{p}_\ell$ while ensuring $\hat{p}_\ell, \hat{p}_{\ell+1} \in (0,1)$.[A.46] Clearly, $\sum_{k=0}^K \hat{p}_k = 1$. Now, let us show that $\mu_{k+1}\hat{p}_{k+1} \leq \lambda_k \hat{p}_k, \forall k = 0, \ldots K-1$. Since these

[A.46] Indeed, at $\varepsilon = 0$, we have $\mu_{\ell+1}\hat{p}_{\ell+1} < \lambda_\ell \hat{p}_\ell$. In addition, for $\varepsilon = p_\ell > 0$ we have $\hat{p}_{\ell+1} = p_{\ell+1} + \varepsilon =$

inequalities holds at p^* (because p^* is an optimal solution of $[P'_K]$ and so satisfies (B')), by construction of $\hat{p}$, we only need to check this constraint for $k = \ell + 1$ and $k = \ell - 1$. For $k = \ell + 1$, we have

$$\mu_{\ell+2}\hat{p}_{\ell+2} = \mu_{\ell+2}p^*_{\ell+2} \leq \lambda_{\ell+1}p^*_{\ell+1} \leq \lambda_{\ell+1}\hat{p}_{\ell+1}.$$

Similarly, for $k = \ell - 1$,

$$\mu_{\ell}\hat{p}_{\ell} \leq \mu_{\ell}p^*_{\ell} \leq \lambda_{\ell-1}p^*_{\ell-1} = \lambda_{\ell-1}\hat{p}_{\ell-1}.$$

Now, we show that the value of the objective of $[\mathcal{L}_{\xi^*}]$ strictly increases when we replace solution p^* by $\hat{p}$. We have

$$\begin{aligned}\sum_{k=0}^{K}\hat{p}_k f(k;\xi^*) - \sum_{k=0}^{K}p^*_k f(k;\xi^*) &= \hat{p}_\ell f(\ell;\xi^*) - p^*_\ell f(\ell;\xi^*) + \hat{p}_{\ell+1}f(\ell+1;\xi^*) - p^*_{\ell+1}f(\ell+1;\xi^*) \\ &= -\varepsilon f(\ell;\xi^*) + \varepsilon f(\ell+1;\xi^*) = \varepsilon\left(f(\ell+1;\xi^*) - f(\ell;\xi^*)\right) > 0\end{aligned}$$

where the inequality comes from the assumption in the lemma. To conclude, we must have that $\mathcal{L}(\hat{p},\xi^*) > \mathcal{L}(p^*,\xi^*)$ which contradicts the fact that (p^*,ξ^*) is a saddle point of the function $\mathcal{L}(\cdot,\cdot)$. ∎

Finally, in the proof of Proposition A.1, we will need the following simple lemma which proof is relegated to Section S.2 of the online appendix.

Lemma A.5. Assume that p' stochastically dominates p. Let φ be a nondecreasing function. If there is κ such that

$$\sum_{k=\kappa}^{K}p'_k > \sum_{k=\kappa}^{K}p_k$$

and $\varphi(\kappa) > \varphi(\kappa-1)$ then

$$\sum_{k=0}^{K}p'_k\varphi(k) > \sum_{k=0}^{K}p_k\varphi(k).$$

Proof. See Section S.2 in the online appendix. ∎

Proof of Proposition A.1. Before proceeding, we make the following straightforward observations (1) $p^*_0 > 0$ (or else $p^*_k = 0$ for all k because, by construction of M_K, p satisfies (B'); this contradicts the assumption that p is a probability measure); (2) for all ξ, $f(0;\xi) = 0$. Using these two facts, we claim that Proposition A.1 holds whenever $f(k;\xi^*) = f(k';\xi^*)$ for all k, k' in the support of p^*. Indeed, since $p^*_0 > 0$, $f(k;\xi^*) = 0$ for all states k in the

$p_{\ell+1} + p_\ell \leq 1$ and $\mu_{\ell+1}\hat{p}_{\ell+1} > \lambda_\ell\hat{p}_\ell = 0$. Hence, by the Intermediate Value Theorem, there must exist $\varepsilon \in (0, p_\ell)$ so that $\mu_{\ell+1}\hat{p}_{\ell+1} = \lambda_\ell\hat{p}_\ell$ and $\hat{p}_\ell, \hat{p}_{\ell+1}$ are in $(0,1)$.

support of p^*. In that case, $\sup_p \mathcal{L}(p, \xi^*) = 0$. Thus, the value of the problem $[P'_K]$ is 0. Clearly, the distribution p corresponding to the Dirac measure on state 0 yields the same value and is a cutoff policy. Hence, in this very special case, Theorem 1 holds true. Thus, in the sequel, we assume that there is a pair of states k and k' in the support of p^* satisfying $f(k; \xi^*) \neq f(k'; \xi^*)$.

Let k^* be $\min \arg\max_k f(k; \xi^*)$ and k^{**} be $\max \arg\max_k f(k; \xi^*)$. Recall that k^* can be equal to $+\infty$. By Lemma A.3, we know that $f(k; \xi^*)$ is strictly increasing up to k^*. Hence, Lemma A.4 implies that $\mu_k p^*_k = \lambda_{k-1} p^*_{k-1}$ for each $k \leq \min\{k^*, K\}$. Note that (since $p^*_0 > 0$) this also implies that $p^*_k > 0$ for each $k \leq \min\{k^*, K\}$, as stated in Proposition A.1. If $K \leq k^*$, we are done. So assume from now on that $K > k^*$; note that this implies that $k^* < +\infty$. By means of contradiction, let us assume that p^* does not exhibit a cutoff policy. This means that there is $k_0 > k^*$ such that $\mu_{k_0} p^*_{k_0} < \lambda_{k_0-1} p^*_{k_0-1}$ and $p^*_{k_0+1} > 0$ (hence, $p^*_{k_0} > 0$).[A.47] Without loss, assume that for any $k < k_0$, we have $\mu_k p^*_k = \lambda_{k-1} p^*_{k-1}$. We consider two cases.

Case 1 : $p^*_k > 0$ for some $k > k^{**}$. Toward a contradiction, we construct a $\hat{p}$ that would achieve a strictly higher value than p^* in $[\mathcal{L}_{\xi^*}]$. Let $\hat{p}_k = p^*_k$ for $k \leq k_0 - 1$. For each $k \geq k_0$, build $\hat{p}$ inductively so that $\mu_{k_0} \hat{p}_{k_0} = \lambda_{k_0-1} \hat{p}_{k_0-1}$, $\mu_{k_0+1} \hat{p}_{k_0+1} = \lambda_{k_0} \hat{p}_{k_0}$... Since the total mass of $\hat{p}$ must be 1, this may be possible only up to a point $\hat{K}$ where, by construction, we will have $\mu_{\hat{K}} \hat{p}_{\hat{K}} \leq \lambda_{\hat{K}-1} \hat{p}_{\hat{K}-1}$. Finally, we set $\hat{p}_k = 0$ for all $k > \hat{K}$. In order to show that $\hat{p}$ lies in $\Delta(\{0, 1, ..., K\})$, we need to show that $\hat{K} \leq K$. By a simple induction argument, $\hat{p}_k \geq p^*_k$ for all $k \leq \hat{K} - 1$ and so we must have that $\hat{K} \leq K$. To recap, there is $\hat{K} \geq k_0$ (potentially equal to K) such that $\mu_k \hat{p}_k = \lambda_{k-1} \hat{p}_{k-1}$ for $k = 0, ..., \hat{K} - 1$, and $\hat{p}_k = 0$ for $k > \hat{K}$. One can show inductively that $\hat{p}_k > p^*_k$ for all $k = k_0, ..., \hat{K} - 1$ while, by construction, $\hat{p}_k = p^*_k$ for all $k \leq k_0 - 1$. We claim that distribution p^* stochastically dominates distribution $\hat{p}$. To see this, fix any $\kappa > \hat{K}$. Clearly, $\sum_{k=\kappa}^{K} \hat{p}_k = 0 \leq \sum_{k=\kappa}^{K} p^*_k$. Now, fix $\kappa \leq \hat{K}$.

$$\sum_{k=\kappa}^{K} \hat{p}_k = 1 - \sum_{k=0}^{\kappa-1} \hat{p}_k \leq 1 - \sum_{k=0}^{\kappa-1} p^*_k = \sum_{k=\kappa}^{K} p^*_k \tag{A.4}$$

where the inequality uses the fact that $\hat{p}_k \geq p^*_k$ for all $k = 0, ..., \kappa - 1$. Importantly, the above inequality is strict for all $\kappa \in \{k_0 + 1, ..., \hat{K}\}$ since $\hat{p}_k > p^*_k$ for all $k = k_0, ..., \hat{K} - 1$.[A.48] It is also strict for any $\kappa \geq \hat{K} + 1$ as long as $p^*_\kappa > 0$ since in that case the LHS is simply 0 while the RHS is strictly positive. In particular, given our assumption that $p^*_k > 0$ for some

[A.47] Indeed, given the above, by definition, p^* exhibits a cutoff policy if and only if $\mu_{k_0} p^*_{k_0} = \lambda_{k_0-1} p^*_{k_0-1}$ for all $k_0 = k^* + 1, \cdots K - 1$, i.e., (B') binds for all $k = 0, ..., K - 2$.

[A.48] Recall that, by construction, $k_0 + 1 \leq \hat{K}$.

$k > k^{**}$, it must be that $p^*_{k^{**}+1} > 0$. Consequently,

$$\sum_{k=\kappa}^{K} \hat{p}_k < \sum_{k=\kappa}^{K} p^*_k \tag{A.5}$$

for $\kappa = \max\{k_0 + 1, k^{**} + 1\}$.

Now, we show that the value of the objective in $[\mathcal{L}_{\xi^*}]$ strictly increases when we replace solution p^* by $\hat{p}$. We have to show that

$$\sum_{k=0}^{K} \hat{p}_k f(k; \xi^*) > \sum_{k=0}^{K} p^*_k f(k; \xi^*).$$

Since $\hat{p}_k = p^*_k$ for all $k \leq k_0 - 1$, this is equivalent to showing

$$\sum_{k=k_0}^{K} \hat{p}_k f(k; \xi^*) > \sum_{k=k_0}^{K} p^*_k f(k; \xi^*) \tag{A.6}$$

Now, define a function $\varphi : \mathbb{Z}_+ \to \mathbb{R}$ as follows

$$\varphi(k) = \begin{cases} f(k_0; \xi^*) \text{ if } k \leq k_0 - 1 \\ f(k; \xi^*) \text{ if } k \geq k_0. \end{cases}$$

Since $k_0 > k^*$, by Lemma A.3, this function is weakly decreasing and it is strictly decreasing from k to $k + 1$ for any $k \geq \max\{k_0, k^{**}\}$. Thus, $\varphi(\kappa - 1) > \varphi(\kappa)$ for $\kappa = \max\{k_0 + 1, k^{**} + 1\}$. Now, we know that p^* stochastically dominates $\hat{p}$, that inequality (A.5) holds at $\kappa = \max\{k_0 + 1, k^{**} + 1\}$. and that $\varphi(\kappa - 1) > \varphi(\kappa)$. Applying Lemma A.5,

$$\sum_{k=0}^{K} (\hat{p}_k - p^*_k)\, \varphi(k) > 0.$$

Since $\hat{p}_k = p^*_k$ for all $k \leq k_0 - 1$, this is equivalent to Equation (A.6). To conclude, $\mathcal{L}(\hat{p}, \xi^*) > \mathcal{L}(p^*, \xi^*)$ which contradicts the fact that (p^*, ξ^*) is a saddle point of $\mathcal{L}(\cdot, \cdot)$.

Case 2 : $p^*_k = 0$ for all $k > k^{**}$. Recall our assumption that there is a pair of states k and k' in the support of p^* satisfying $f(k; \xi^*) \neq f(k'; \xi^*)$. Hence, because $f(\cdot; \xi^*)$ is single-peaked, f must be weakly increasing on the support of p^* and strictly increasing from k to $k + 1$ for all $k < k^*$. In particular, this holds at $k = 0$, and so we have $f(0; \xi^*) < f(1; \xi^*)$ and $p^*_0 > 0$. Recall that k_0 is the smallest k in $\{k^* + 1, ..., k^{**} - 1\}$ such that $\mu_k p^*_k < \lambda_{k-1} p^*_{k-1}$

and $p^*_{k+1} > 0$. We now construct a measure $\hat{p}$ as follows

$$\hat{p}_k = \begin{cases} p^*_k/Z_1 \text{ if } k \leq k_0 - 1 \\ p^*_k + Z_2 \text{ if } k = k_0 \\ p^*_k \text{ if } k \geq k_0 + 1, \end{cases}$$

where $Z_1 > 1$ and $Z_2 \triangleq \sum_{k=0}^{k_0-1}(p^*_k - \hat{p}_k)$ so that $\hat{p}$ sums up to 1. We pick Z_1 small enough so that $\hat{p}_{k_0}$ remains between 0 and 1 for each k. We show that, for $Z_1 > 1$ small enough, for each $k \leq K$, $\mu_k \hat{p}_k \leq \lambda_{k-1}\hat{p}_{k-1}$. To see this, first fix $k \leq k_0 - 1$ and note that

$$\mu_k \hat{p}_k = \mu_k p^*_k/Z_1 \leq \lambda_{k-1}p^*_{k-1}/Z_1 = \lambda_{k-1}\hat{p}_{k-1}$$

where the inequality follows from the fact that p^* is a feasible solution of $[P'_K]$. Next,

$$\mu_{k_0}\hat{p}_{k_0} = \mu_{k_0}\left(p^*_{k_0} + Z_2\right) \leq \lambda_{k_0-1}p^*_{k_0-1}/Z_1 = \lambda_{k_0-1}\hat{p}_{k_0-1}$$

where the inequality holds if Z_1 is small enough since, by assumption, $\mu_{k_0}p^*_{k_0} < \lambda_{k_0-1}p^*_{k_0-1}$ (and Z_2 vanishes as Z_1 goes to 1).[A.49] Now, for $k = k_0 + 1$, we have

$$\mu_{k_0+1}\hat{p}_{k_0+1} = \mu_{k_0+1}p^*_{k_0+1} \leq \lambda_{k_0}p^*_{k_0} \leq \lambda_{k_0}(p^*_{k_0} + Z_2) = \lambda_{k_0}\hat{p}_{k_0}.$$

Finally, by construction, for any $k > k_0 + 1$, $\mu_k\hat{p}_k \leq \lambda_{k-1}\hat{p}_{k-1}$ must hold since p^* and $\hat{p}$ coincide.

Now, we show that the value of the objective in $[\mathcal{L}_{\xi^*}]$ strictly increases when we replace solution p^* by $\hat{p}$. To see this, observe first that $\hat{p}$ must stochastically dominate p^*. Indeed, fix any $\kappa > k_0$. Clearly, since $\hat{p}_k = p^*_k$ for all $k \geq k_0 + 1$, $\sum_{k=\kappa}^{K}\hat{p}_k = \sum_{k=\kappa}^{K}p^*_k$. Now, fix $\kappa \leq k_0$.

$$\sum_{k=\kappa}^{K}\hat{p}_k = 1 - \sum_{k=0}^{\kappa-1}\hat{p}_k > 1 - \sum_{k=0}^{\kappa-1}p^*_k = \sum_{k=\kappa}^{K}p^*_k \tag{A.7}$$

where the inequality uses the fact that $\hat{p}_k = p^*_k/Z_1 < p^*_k$ for all $k = 0, ..., \kappa - 1$ (since $Z_1 > 1$ and $p^*_k > 0$ for such k). Now, we show that the value of the objective in $[\mathcal{L}_{\xi^*}]$ strictly increases when we replace solution p^* by $\hat{p}$, i.e.,

$$\sum_{k=0}^{K}\hat{p}_k f(k;\xi^*) > \sum_{k=0}^{K}p^*_k f(k;\xi^*).$$

[A.49] Indeed, by construction, for each $k \leq k_0 - 1$, $\hat{p}_k \to p^*_k$ as $Z_1 \to 1$. Since $Z_2 = \sum_{k=0}^{k_0-1}(p^*_k - \hat{p}_k)$, Z_2 converges to 0 as $Z_1 \to 1$.

We know that $\hat{p}$ stochastically dominates p^*, that inequality (A.7) holds at $\kappa = 1$ and that $f(0;\xi^*) < f(1;\xi^*)$. In addition, $f(\cdot;\xi^*)$ is nondecreasing on the support of p^* and $\hat{p}$. Hence, this follows from Lemma A.5. ∎

A.2 Infinite dimensional analysis

Let us consider the sequence $\{p^K\}_K$ where for each K, p^K is an optimal solution of problem $[P'_K]$. If μ is regular, we assume each p^K exhibits a cutoff policy which is well-defined by Proposition A.1. For each K, we see p^K as a point in $\mathbb{R}^{\mathbb{Z}_+}$ with value 0 on states weakly greater than $K+1$. We will be interested in the limit points of sequence $\{p^K\}_K$. Together with the result showing that $[P']$ has an optimal solution, the following statement implies Theorem 1.

Proposition A.2. Assume μ is regular. Sequence $\{p^K\}_K$ has a subsequence which converges to a distribution p^* which is an optimal solution to $[P']$ and exhibits a cutoff policy. Further, it satisfies $p^*_k > 0$ for each $k \leq \min \arg\max_k \mu_k V - Ck$.

This result is shown in the online appendix Section S.3 through the following steps. First, we show that the infinite-dimensional problem $[P']$ admits an optimal solution (Proposition S1). Then, we show that the set of feasible distributions of $[P']$ exhibiting a cutoff-policy is sequentially compact, which in turn implies that (when μ is regular) $\{p^K\}_K$ has a subsequence converging to a point which exhibits a cutoff policy (Proposition S4). Finally, we argue that any limit point of $\{p^K\}_K$ must be an optimal solution of $[P']$ (Proposition S5).

B Proofs from Section 5: FCFS with No Information

B.1 Proof of Lemma 1

The expected waiting time satisfies the following recursion. The agent in the first position has expected waiting time

$$\tau_1^* = (q_1^* dt)dt + [1 - q_1^* dt](\tau_1^* + dt) + o(dt),$$

since he waits for dt period with probability $q_1^* dt$ and for $\tau_1^* + dt$ periods with the remaining probability. Letting $dt \to 0$, we get

$$\tau_1^* = 1/q_1^* = 1/\mu_1.$$

More generally, the agent in queue position ℓ waits for

$$\tau_\ell^* = (q_\ell^* dt)dt + \left[1 - \sum_{j=1}^{\ell} q_j^* dt\right](\tau_\ell^* + dt) + \left(\sum_{j=1}^{\ell-1} q_j^* dt\right)(\tau_{\ell-1}^* + dt) + o(dt),$$

since he is served in dt period with probability $q_\ell^* dt$, in $\tau_\ell^* + dt$ periods with probability $1 - \sum_{j=1}^{\ell} q_j^* dt$ (when nobody before him is served), and in $\tau_{\ell-1}^* + dt$ periods with probability $\sum_{j=1}^{\ell-1} q_j^* dt$ (when somebody before him is served).[B.50]

The recursion equations yield a unique solution:

$$\tau_\ell^* = \frac{\ell}{\sum_{j=1}^{\ell} q_j^*} = \frac{\ell}{\mu_\ell},$$

where the last equality follows from feasibility.

Part (ii) of regularity implies that q_ℓ^* is nonincreasing in ℓ. Therefore, for each ℓ

$$\tau_{\ell+1}^* - \tau_\ell^* = \frac{\sum_{j=1}^{\ell} q_j^* - \ell q_{\ell+1}^*}{(\sum_{j=1}^{\ell} q_j^*)(\sum_{j=1}^{\ell+1} q_j^*)} \geq 0.$$

Hence, it follows that τ_ℓ^* is nonincreasing in ℓ. Further, if $2\mu_1 > \mu_2$, then $q_1^* > q_2^* \geq q_\ell^*$ for all $\ell \geq 2$. Then, the above inequality becomes strict for all ℓ, which proves the last statement. ∎

B.2 Proof of Lemma 2

We let $\bar{K}$ be the largest state in the support of p^* (which can potentially be infinite). We first study the dynamics for the case with $\bar{K} < \infty$. For $\bar{K} = \infty$, we show that the dynamics can be approximated by the dynamics for $\bar{K} < \infty$ when $\bar{K}$ goes to infinity. While it requires some care, the argument for $\bar{K} = \infty$ essentially relies on the case with $\bar{K} < \infty$. Hence, we defer the proof to online appendix Section S.5. In the sequel, we assume that $\bar{K} < \infty$.

Using (2), we write for each such $\ell \geq 2$,

$$r_\ell^{t+dt} = \frac{\tilde{\gamma}_\ell^{t+dt}}{\tilde{\gamma}_{\ell-1}^{t+dt}} = \frac{(1 - \mu_\ell dt)\tilde{\gamma}_\ell^t + \mu_\ell dt \tilde{\gamma}_{\ell+1}^t}{(1 - \mu_{\ell-1} dt)\tilde{\gamma}_{\ell-1}^t + \mu_{\ell-1} dt \tilde{\gamma}_\ell^t} + o(dt) = \frac{1 - \mu_\ell dt + \mu_\ell dt r_{\ell+1}^t}{(1 - \mu_{\ell-1} dt)\frac{1}{r_\ell^t} + \mu_{\ell-1} dt} + o(dt).$$

[B.50] Again, the probability that multiple agents are served during $[t, t + dt)$ has a lower order of magnitude denoted by $o(dt)$.

Rearranging, we get

$$\frac{r_\ell^{t+dt} - r_\ell^t}{dt} = \frac{\mu_{\ell-1} - \mu_\ell - \mu_{\ell-1} r_\ell^t + \mu_\ell r_{\ell+1}^t}{(1-\mu_{\ell-1}dt)\frac{1}{r_\ell^t} + \mu_{\ell-1}dt} + o(dt)/dt.$$

Letting $dt \to 0$, we obtain

$$\dot{r}_\ell^t = r_\ell^t \left(\mu_{\ell-1} - \mu_\ell - \mu_{\ell-1} r_\ell^t + \mu_\ell r_{\ell+1}^t\right). \tag{B.8}$$

(B.8) forms a system of ordinary differential equations. The boundary condition is defined as follows. Recall that the effective arrival rate be $\tilde{\lambda}_k \triangleq \lambda_k x_k^*$ for each k. For $\ell \leq \bar{K}$,

$$r_\ell^0 = \frac{\tilde{\gamma}_\ell^0}{\tilde{\gamma}_{\ell-1}^0} = \frac{p_\ell^* \mu_\ell}{p_{\ell-1}^* \mu_{\ell-1}} = \frac{\tilde{\lambda}_{\ell-1}}{\mu_{\ell-1}}, \tag{B.9}$$

where the second equality uses the fact that $\tilde{\gamma}_\ell^0 = p_\ell^* \mu_\ell \setminus \sum_{i=1}^\infty p_i^* \mu_i$ for each ℓ, while the third one uses (B) whereby $\frac{p_\ell^*}{p_{\ell-1}^*} = \frac{\tilde{\lambda}_{\ell-1}}{\mu_\ell}$.[B.51] It is routine to see that the system of ODEs (B.8) together with the boundary condition (B.9) admits a unique solution $(r_\ell^t)_\ell$ for all $t \geq 0$.[B.52]

We first claim that $\dot{r}_\ell^0 \leq 0$ for all $\ell = 2, ..., \bar{K}$. It follows from (B.8) that, for $\ell = 2, ..., \bar{K}$, $\dot{r}_\ell^0 \leq 0$ if and only if

$$\mu_{\ell-1} - \mu_\ell \leq \mu_{\ell-1} r_\ell^0 - \mu_\ell r_{\ell+1}^0. \tag{B.10}$$

Consider any $\ell = 2, ..., \bar{K}$. Substituting (B.9) into (B.10), the condition simplifies to:

$$\mu_{\ell-1} - \mu_\ell \leq \tilde{\lambda}_{\ell-1} - \tilde{\lambda}_\ell,$$

which holds by regularity of (λ, μ) and the fact that x_k^* is nonincreasing in k.

Having established that $\dot{r}_\ell^0 \leq 0$ for each $\ell = 2, ..., \bar{K}$, we next prove that $\dot{r}_\ell^t \leq 0$ for all $t > 0$. To this end, suppose this is not the case. Then, there exists

$$\ell \in \arg\min_{\ell'=2,...,\bar{K}} T_{\ell'},$$

where

$$T_{\ell'} \triangleq \inf\{t' : \dot{r}_{\ell'}^{t'} > 0\}$$

[B.51] One can obtain the expression for $\tilde{\gamma}_\ell^0$ as follows. The optimality of the cutoff policy means $x_k^* = 1$ for all $k = 0, ..., K^* - 2$, $x_k^* = 0$ for all $k > K^* - 1$, and $y_{k,\ell}^* = z_{k,\ell}^* = 0$ for all (k, ℓ). Substituting these into (B), one obtains the expression by rewriting (1).

[B.52] This follows from the observation that the RHS of (B.8) is locally Lipschitzian in r (a fact implied by the continuous differentiability of RHS in r_ℓ^t's). See Hale p. 18, Theorem 3.1, for instance.

if the infimum is well defined, or else $T_{\ell'} \triangleq \infty$. Let $t = T_\ell < \infty$, by the hypothesis. Then, we must have

$$\dot{r}^t_\ell > 0; \dot{r}^t_{\ell'} \leq 0, \forall \ell' \neq \ell; \text{ and } \dot{r}^t_\ell = 0.$$

Differentiating (B.8) on both sides, we obtain

$$0 < \ddot{r}^t_\ell = \dot{r}^t_\ell \left(\mu_{\ell-1} - \mu_\ell - \mu_{\ell-1} r^t_\ell + \mu_\ell r^t_{\ell+1}\right) - r^t_\ell(\mu_{\ell-1}\dot{r}^t_\ell - \mu_\ell \dot{r}^t_{\ell+1}) = r^t_\ell \mu_\ell \dot{r}^t_{\ell+1} \leq 0,$$

a contradiction. We thus conclude that $\dot{r}^t_\ell \leq 0$, for all $\ell = 2, ..., \bar{K}$, for all $t \geq 0$.

C Proof of Theorem 3

Fix a queuing rule q which differs from FCFS. We consider the information policy that provides no information (beyond the recommendations) for all $t \geq 0$. This is without loss since, if a queueing rule q fails (IC_t), for some $t \geq 0$, under no information, it would fail (IC_t) under *any* information policy.

Recall that we have fixed the service rate μ. While arrival rate λ is yet to be fixed, for each λ, we can choose parameters V, C and α to ensure that the optimal outcome (x^*, y^*, z^*, p^*) (i) involves a maximal length $K^* = 2$ (i.e., $x^*_2 = 0$ or $z^*_{2,1} + z^*_{2,2} = 1$), (ii) no rationing at $k = 1$ (i.e., $x^*_1 = 1$ and $z^*_{1,1} = 0$), and (iii) (IR) is binding at p^*.[C.53] Importantly, assumption (ii) implies that $y^*_{k,\ell}$ are all zeros.[C.54] In the sequel, we fix such an outcome (x^*, y^*, z^*, p^*). Note that $x^*_2 > 0$ implies that $z^*_{2,1} + z^*_{2,2} = 1$ and since the values of $z^*_{2,1}$ and $z^*_{2,2}$ are irrelevant when $x^*_2 = 0$, without loss, we will assume that $z^*_{2,1} + z^*_{2,2} = 1$. While the variables we study below do depend on μ and λ, for simplicity, we omit the dependence in notations.

We then study an agent's expected utility with elapse of time $t \geq 0$ on the queue:

$$U(t) \triangleq S(t)V - W(t)C. \tag{C.11}$$

[C.53] If $V/C = \frac{2\lambda+\mu}{(\lambda+\mu)\mu}$ and $\alpha = 0$, one can easily show that there is a unique optimal solution p to $[P']$ and any outcome (x, y, z) implementing p satisfies (i), (ii) and (iii).

[C.54] Indeed, in that case, $x^*_0 = x^*_1 = 1$ and $\sum_{\ell=1}^0 z^*_{0,\ell} = \sum_{\ell=1}^1 z^*_{1,\ell} = 0$. Further, (x^*, y^*, z^*, p^*) satisfies (B), i.e., for each k

$$p^*_k \lambda_k x^*_k (1 - \sum_{\ell=1}^{k} z^*_{k,\ell}) = p^*_{k+1} (\sum_{\ell=1}^{k+1} y^*_{k+1,\ell} + \mu_{k+1}).$$

From the above equation, it is easily checked that if $x^*_k = 1$ and $\sum_{\ell=1}^k z^*_{k,\ell} = 0$, given that $p^*_k \lambda_k \leq p^*_{k+1}\mu_{k+1}$ since p^* satisfies (B'), we must have that $y^*_{k+1,\ell} = 0$ for each ℓ. Thus, we must have that $y^*_{1,\ell} = y^*_{2,\ell} = 0$ for each ℓ.

$W(t)$ stands for the residual waiting time, conditional on having spent time $t \geq 0$ on the queue, i.e.,

$$W(t) \triangleq \gamma^t_{1,1}\tau_{1,1} + \gamma^t_{2,1}\tau_{2,1} + \gamma^t_{2,2}\tau_{2,2}$$

where $\gamma^t = (\gamma^t_{1,1}, \gamma^t_{2,1}, \gamma^t_{2,2})$ is the belief an agent has about alternative states (k, ℓ) and $\tau = (\tau_{1,1}, \tau_{2,1}, \tau_{2,2})$ are his expected waiting times at alternative states, both under the queueing rule q. Similarly, $S(t)$ is the probability of eventually getting served and writes as:

$$S(t) \triangleq \gamma^t_{1,1}\sigma_{1,1} + \gamma^t_{2,1}\sigma_{2,1} + \gamma^t_{2,2}\sigma_{2,2}$$

where $\sigma = (\sigma_{1,1}, \sigma_{2,1}, \sigma_{2,2})$ are the probabilities of an agent getting eventually served at alternative states (k, ℓ), again under the queueing rule q. (Throughout, we suppress the dependence on q for notational ease.)

Since $U(0) = 0$ (as implied by a binding (IR)), it suffices to show that $U(t)$ decreases strictly in the neighborhood of $t = 0$ which will then prove that q fails (IC_t) for some small $t > 0$. We establish this for a sufficiently small value $\lambda > 0$.[C.55] Specifically, we focus on $\dot{U}(0)$—the change in utility "right after joining the queue"—as $\lambda \to 0$. As it turns out, $\dot{U}(0) \to 0$ as $\lambda \to 0$. Hence, one must consider how "slowly" $\dot{U}(0)$ converges to 0, or more precisely, the limit behavior of $\dot{U}(0)/\lambda$ as $\lambda \to 0$.

Hence, we will show that $\dot{U}(0)/\lambda$ converges to a strictly negative number as $\lambda \to 0$. For our purpose, it is enough to show that, as λ vanishes, $S'(0)/\lambda$ converges to 0 while $W'(0)/\lambda$ converges to a strictly positive number. To this end, it is necessary to characterize the limit behaviors of $(\tau_{k,\ell}), (\sigma_{k,\ell})$ and $(\dot{\gamma}^0_{k,\ell})$. We do this first.

Limit behavior of $(\tau_{k,\ell})$. The expected waiting time $\tau_{1,1}$ must satisfy:

$$\tau_{1,1} = (\mu dt)\, dt + \lambda dt\, (dt + \tau_{2,1}) + (1 - \mu dt - \lambda dt)\, (dt + \tau_{1,1}) + o(dt),$$

since, for a small time increment dt, the sole agent in the queue waits for time dt if he is served during $[t, t + dt)$ (which occurs with probability μdt), for $dt + \tau_{2,1}$ if another agent arrives during $[t, t + dt)$ (which occurs with probability λdt), and for $dt + \tau_{1,1}$ if neither event arises (which occurs with probability $1 - \mu dt - \lambda dt$). By a similar reasoning, we have:

$$\tau_{2,1} = \left(q_{2,1}dt + \lambda x^*_2 z^*_{2,1} dt\right) dt + q_{2,2}dt(dt + \tau_{1,1}) + \left(1 - \mu dt - \lambda x^*_2 z^*_{2,1} dt\right) (dt + \tau_{2,1}) + o(dt)$$

[C.55] Recall we adjust the values of C, V and α so as to ensure that (IR) is binding at the optimal cutoff policy that solves $[P']$.

and

$$\tau_{2,2} = \left(q_{2,2}dt + \lambda x_2^* z_{2,2}^* dt\right) dt + q_{2,1}dt(dt+\tau_{1,1}) + \lambda x_2^* z_{2,1}^* dt\,(dt+\tau_{2,1}) + (1-\mu dt - \lambda x_2^* dt)\,(dt+\tau_{2,2}) + o(dt).$$

Letting $dt \to 0$ and simplifying, we obtain:

$$(\mu+\lambda)\,\tau_{1,1} = \lambda\tau_{2,1}+1,\ \ \left(\mu+\lambda x_2^* z_{2,1}^*\right)\tau_{2,1} = q_{2,2}\tau_{1,1}+1 \text{ and } \ (\mu+\lambda x_2^*)\,\tau_{2,2} = \lambda x_2^* z_{2,1}^*\tau_{2,1} + q_{2,1}\tau_{1,1}+1.$$

Thus, we have that, as $\lambda \to 0$,

$$\tau_{1,1} \to \frac{1}{\mu}, \tau_{2,1} \to \frac{q_{2,2}}{\mu}\frac{1}{\mu} + \frac{1}{\mu} \text{ and } \tau_{2,2} \to \frac{q_{2,1}}{\mu}\frac{1}{\mu} + \frac{1}{\mu} \tag{C.12}$$

where we abuse notations and simply note $q_{2,2}$ for the limit as λ vanishes of $q_{2,2}$ (and similarly for $q_{2,1}$). We assume here that this limit is well-defined and take a subsequence of our vanishing sequence of λ if necessary.

Limit behavior of $(\sigma_{k,\ell})$. We have

$$\sigma_{1,1} = \mu dt + \lambda dt\sigma_{2,1} + (1-\mu dt - \lambda dt)\sigma_{1,1} + o(dt)$$

since, for a small time increment dt, the sole agent in the queue is served with probability μdt; the agent is eventually served with probability $\sigma_{2,1}$ if another agent arrives (which occurs with probability λdt), and the agent is served with probability $\sigma_{1,1}$ if neither event arises (which occurs with probability $1-\mu dt - \lambda dt$). Similar reasoning yields the following expressions for $\sigma_{2,1}$ and $\sigma_{2,2}$

$$\sigma_{2,1} = q_{2,1}dt + (1-\mu dt - \lambda x_2^* dt)\sigma_{2,1} + q_{2,2}dt\sigma_{1,1} + \lambda x_2^* dt z_{2,2}^*\sigma_{2,1} + o(dt),$$

and

$$\sigma_{2,2} = q_{2,2}dt + (1-\mu dt - \lambda x_2^* dt)\sigma_{2,2} + q_{2,1}dt\sigma_{1,1} + \lambda x_2^* dt z_{2,1}^*\sigma_{2,1} + o(dt).$$

We obtain

$$\begin{aligned}(\mu+\lambda)\sigma_{1,1} &= \mu + \lambda\sigma_{2,1}\\ (\mu+\lambda x_2^*(1-z_{2,2}^*))\sigma_{2,1} &= q_{2,1} + q_{2,2}\sigma_{1,1}\\ (\mu+\lambda x_2^*)\sigma_{2,2} &= q_{2,2} + q_{2,1}\sigma_{1,1} + \lambda x_2^* z_{2,1}^*\sigma_{2,1}.\end{aligned}$$

Hence, we obtain that

$$\sigma_{1,1}, \sigma_{2,1}, \sigma_{2,2} \to 1 \text{ as } \lambda \to 0. \tag{C.13}$$

Limit behavior of $(\dot{\gamma}^0_{k,\ell})$. We study the dynamics of beliefs. An agents' beliefs evolve during $[t, t+dt)$ according to Bayes rule. For instance, for state $(k,\ell) = (1,1)$, we obtain

$$\gamma^{t+dt}_{1,1} = \frac{\gamma^t_{1,1}\left[1-\mu dt-\lambda dt\right] + \gamma^t_{2,2}\left[q_{2,1}dt\right] + \gamma^t_{2,1}\left[q_{2,2}dt\right]}{\gamma^t_{1,1}\left[1-\mu dt\right] + \gamma^t_{2,1}\left[1-q_{2,1}dt-\lambda x^*_2 z^*_{2,1}dt\right] + \gamma^t_{2,2}\left[1-q_{2,2}dt-\lambda x^*_2 z^*_{2,2}dt\right]} + o(dt)$$

where the numerator is the probability that the agent's state is $(k,\ell) = (1,1)$ after staying in the queue for length $t+dt$ of time. This event occurs if either (i) the agent is already in state $(1,1)$ in the queue at time t, the agent is not served and no agent arrives in the queue during time increment dt; or (ii) his state is $(2,2)$ or $(2,1)$ at t and the other agent in the queue is served by $t+dt$. The denominator in turn gives the probability that the agent has not been served or removed from the queue by time $t+dt$. Hence, given that an agent has not been served or removed from the queue by t, the above expression gives the conditional belief that his state is $(1,1)$ at time $t+dt$.

Similar reasoning yields the following expressions for the evolution of beliefs for state $(2,1)$ and $(2,2)$

$$\gamma^{t+dt}_{2,1} = \frac{\gamma^t_{2,1}\left[\lambda x^*_2 z^*_{2,2}dt + 1 - \mu dt - \lambda x^*_2 dt\right] + \gamma^t_{2,2}\left[\lambda x^*_2 z^*_{2,1}dt\right] + \gamma^t_{1,1}\left[\lambda dt\right]}{\gamma^t_{1,1}\left[1-\mu dt\right] + \gamma^t_{2,1}\left[1-q_{2,1}dt-\lambda x^*_2 z^*_{2,1}dt\right] + \gamma^t_{2,2}\left[1-q_{2,2}dt-\lambda x^*_2 z^*_{2,2}dt\right]} + o(dt)$$

and

$$\gamma^{t+dt}_{2,2} = \frac{\gamma^t_{2,2}\left[1-\mu dt-\lambda x^*_2 dt\right]}{\gamma^t_{1,1}\left[1-\mu dt\right] + \gamma^t_{2,1}\left[1-q_{2,1}dt-\lambda x^*_2 z^*_{2,1}dt\right] + \gamma^t_{2,2}\left[1-q_{2,2}dt-\lambda x^*_2 z^*_{2,2}dt\right]} + o(dt).$$

From these, we can derive ODEs that describe belief evolutions:

$$\begin{aligned}\dot{\gamma}^t_{1,1} &= -\gamma^t_{1,1}\left[\mu+\lambda\right] + \gamma^t_{2,2}\left[q_{2,1}\right] + \gamma^t_{2,1}\left[q_{2,2}\right] + \left(\gamma^t_{1,1}\right)^2\left[\mu\right] \\ &\quad + \gamma^t_{1,1}\gamma^t_{2,1}\left[q_{2,1}+\lambda x^*_2 z^*_{2,1}\right] + \gamma^t_{1,1}\gamma^t_{2,2}\left[q_{2,2}+\lambda x^*_2 z^*_{2,2}\right],\end{aligned}$$

$$\begin{aligned}\dot{\gamma}^t_{2,1} &= -\gamma^t_{2,1}\left[\mu+\lambda x^*_2(1-z^*_{2,2})\right] + \gamma^t_{2,2}\left[\lambda x^*_2 z^*_{2,1}\right] + \gamma^t_{1,1}\left[\lambda\right] \\ &\quad + \gamma^t_{2,1}\gamma^t_{1,1}\left[\mu\right] + \left(\gamma^t_{2,1}\right)^2\left[q_{2,1}+\lambda x^*_2 z^*_{2,1}\right] + \gamma^t_{2,1}\gamma^t_{2,2}\left[q_{2,2}+\lambda x^*_2 z^*_{2,2}\right]\end{aligned}$$

and

$$\begin{aligned}\dot{\gamma}^t_{2,2} &= -\gamma^t_{2,2}\left[\mu+\lambda x^*_2\right]+\gamma^t_{2,2}\gamma^t_{1,1}\left[\mu\right]\\ &\quad +\gamma^t_{2,2}\gamma^t_{2,1}\left[q_{2,1}+\lambda x^*_2 z^*_{2,1}\right]+\left(\gamma^t_{2,2}\right)^2\left[q_{2,2}+\lambda x^*_2 z^*_{2,2}\right]\end{aligned}$$

with a boundary condition at $t=0$ satisfying $\gamma^0_{2,1}=0$ and

$$\gamma^0_{1,1}=\frac{\lambda p_0}{\lambda p_0+\lambda p_1+\lambda x^*_2 p_2\left(z^*_{2,1}+z^*_{2,2}\right)}=\frac{1}{1+\frac{\lambda}{\mu}+x^*_2\left(\frac{\lambda}{\mu}\right)^2},$$

and

$$\gamma^0_{2,2}=\frac{\lambda p_1+\lambda x^*_2 p_2\left(z^*_{2,1}+z^*_{2,2}\right)}{\lambda p_0+\lambda p_1+\lambda x^*_2 p_2\left(z^*_{2,1}+z^*_{2,2}\right)}=\frac{\frac{\lambda}{\mu}+x^*_2\left(\frac{\lambda}{\mu}\right)^2}{1+\frac{\lambda}{\mu}+x^*_2\left(\frac{\lambda}{\mu}\right)^2},$$

where we used the fact that $p_1\mu=\lambda p_0$ and $p_2\mu=\lambda p_1=\lambda\frac{\lambda}{\mu}p_0$ at the invariant distribution together with $z^*_{2,1}+z^*_{2,2}=1$ since state $k\geq 3$ have mass 0 at the invariant distribution. (Recall that we assumed, wlog, that $z^*_{2,1}+z^*_{2,2}=1$).

Observe that

$$\frac{\gamma^0_{1,1}}{\lambda}-\frac{1}{\lambda}\to-\frac{1}{\mu},\ \frac{\gamma^0_{2,2}}{\lambda}\to\frac{1}{\mu}\text{ and }\frac{\gamma^0_{2,1}}{\lambda}=0$$

In addition,

$$\frac{\dot{\gamma}^0_{1,1}}{\lambda}\to-1<0,\ \frac{\dot{\gamma}^0_{2,1}}{\lambda}\to 1>0\text{ and }\frac{\dot{\gamma}^0_{2,2}}{\lambda}\to 0. \tag{C.14}$$

Completion of the proof of Theorem 3. As we already mentioned, for our purpose, it is enough to show that as λ vanishes, $S'(0)/\lambda$ converges to 0 while $W'(0)/\lambda$ converges to a strictly positive number. We have that

$$\frac{W'(0)}{\lambda}=\frac{\dot{\gamma}^t_{1,1}}{\lambda}\tau_{1,1}+\frac{\dot{\gamma}^0_{2,1}}{\lambda}\tau_{2,1}+\frac{\dot{\gamma}^0_{2,2}}{\lambda}\tau_{2,2}\to-\frac{1}{\mu}+\left(\frac{q_{2,2}}{\mu}\frac{1}{\mu}+\frac{1}{\mu}\right)=\left(\frac{q_{2,2}}{\mu}\right)\frac{1}{\mu}>0$$

where the limit result comes from (C.12) and (C.14) while the strict inequality holds given our assumption that q differs from FCFS and so $q_{2,2}>0$. Further, we have

$$\frac{S'(0)}{\lambda}=\frac{\dot{\gamma}^t_{1,1}}{\lambda}\sigma_{1,1}+\frac{\dot{\gamma}^0_{2,1}}{\lambda}\sigma_{2,1}+\frac{\dot{\gamma}^0_{2,2}}{\lambda}\sigma_{2,2}\to 0$$

where the limit result comes from (C.13) and (C.14). Thus, as claimed, $\dot{U}(0)/\lambda$ converges to a strictly negative number as $\lambda\to 0$. ∎