Online Appendix:
"Optimal Queue Design"
(Not for Publication)

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

August 14, 2021

S.1 Axiomatic Foundation for Regular Service Rates

As defined in the text, for each $k \in \mathbb{N}$, μ_k represents the maximal service rate that any set of k agents may receive. By definition, μ_k is nondecreasing in k since if $\mu_k > \mu_m$ for $k < m$, we can simply redefine $\mu_m \triangleq \mu_k$. In this sense, we view $\mu = (\mu_k)_{k \in \mathbb{N}}$ as "effective" maximal service rates.[A.1]

Below we provide a more primitive definition of FCFS based on the concept that the priority must be assigned greedily to maximize the service rates for earlier arriving agents. Under regularity of μ, this definition will then produce the formula we presented in the main paper as the definition of FCFS.

[1]Department of Economics, Columbia University, USA. Email: yeonkooche@gmail.com.

[2]Department of Economics, Paris School of Economics, France. Email: tercieux@pse.ens.fr.

[A.1]Indeed, we can characterize μ as arising from more primitive service constraints. Say there are upper bound constraints $(c_k)_k$ for each group of k agents. We do not impose any condition on $(c_k)_k$, except that there exists $B > 0$ such that $c_k \leq B$ for all k and it is nondecreasing. The effective service rate μ_n for n agents can be defined as the value:

$$[C_n] \qquad \sup_{q \in \mathbb{R}^n_+} \sum_{i \in [n]} q_i$$

subject to

$$\sum_{i \in S} q_i \leq c_k, \forall k \in \mathbb{N}, \forall S \subset [n] \; s.t. \; |S| = k.$$

FCFS: Specifically, for each $k \in \mathbb{N}$, we define the service rates $(q^*_{k,1}, ..., q^*_{k,k}) \in \mathbb{R}^k_+$ that agents in the queue of k length receive under FCFS.

To begin, let $Q_0 \triangleq \mathbb{R}^k_+$, and consider a sequence of the following problems:

In step $j \in [k] \triangleq \{1, ..., k\}$, we choose

$$[C^*_j] \qquad Q_j = \arg \max_{q \in Q_{j-1}} \sum_{i \in [j]} q_i$$

subject to

$$\sum_{i \in S} q_i \leq \mu_m, \forall m \in [k], \forall S \subset [k] \; s.t. \; |S| = m.$$

In words, the first agent's service rate is maximized subject to the constraint that he can never receive more than μ_1 the maximal service rate any single agent can ever receive. Taking that as constraint, we next maximize first and second agents' service rate now only subject to μ_2 the maximal total service rate that any two agents can ever receive, and so on.

The FCFS service rates $(q^*_{k,1}, ..., q^*_{k,k})$ are then defined to be an optimal solution for step k—i.e., an element of Q_k. While it is in principle possible that Q_k has multiple elements, it is easily seen that Q_k is a singleton. We let μ^*_k denote the maximized value of $[C^*_k]$. Next, observe that, for any $i \leq k, k'$, we have $q^*_{k,i} = q^*_{k',i}$. Hence, we henceforth write q^*_i for $q^*_{k,i}$.

We now derive the optimal solution $(q^*_j)_{j \in [k]}$ explicitly. The resulting formula will resemble the one we defined for the service rates under FCFS.

Lemma S1. Fix k. The optimal value of $[C^*_j]$ is μ^*_j, where $\mu^*_1 = \mu_1$, and for $j = 2, ..., k$,

$$\mu^*_j = \mu^*_{j-1} + \min\{\mu_j - \mu^*_{j-1}, \mu^*_{j-1} - \mu^*_{j-2}\}.$$

Agent $j \in [k]$ receives service rates $q^*_j = \mu^*_j - \mu^*_{j-1}$, which is nonincreasing in j, for $j = 1, ...k$, where $\mu^*_0 \triangleq 0$.

Proof. The proof is inductive. First, it is trivial to note that $\mu^*_1 = \mu_1$ is indeed the value of $[C^*_1]$ and $q^*_1 = \mu^*_1 = \mu^*_1 - \mu^*_0$. Suppose next that μ^*_i is the value of $[C^*_i]$ for all $i = 1, ..., j-1$, and these steps pin down $q^*_i := \mu^*_i - \mu^*_{i-1}$. We make several observations: (i) Since μ^*_i is the value of $[C^*_i]$ for $i = j-2, j-1$, any $q \in Q_{j-1}$ has $q_\ell \leq \mu^*_{j-1} - \mu^*_{j-2} = q^*_{j-1}$ for all $\ell \geq j-1$. (Suppose to the contrary that $q_\ell > \mu^*_{j-1} - \mu^*_{j-2}$ for some $q \in Q_{j-1}$, then swapping q_{j-1} and q_ℓ between $j-1$ and ℓ is feasible and strictly improves the value of $[C^*_{j-1}]$, a contradiction.) (ii) By construction, we have $\mu^*_i \leq \mu_i$ for all $i = 1, ..., j-1$. (iii) By construction, we have $q^*_i \leq q^*_{i'}$ for $i' \leq i \leq j-1$ (which follows from the fact that $\mu^*_j - \mu^*_{j-1}$ is nonincreasing in j.).

Consider problem $[C^*_j]$. We will argue that its value is given by the formula $\mu^*_j = \mu^*_{j-1} + \min\{\mu_j - \mu^*_{j-1}, \mu^*_{j-1} - \mu^*_{j-2}\}$, and it pins down $q^*_j = \mu^*_j - \mu^*_{j-1}$. To this end, note first that the value μ^*_j of $[C^*_j]$ cannot exceed:

$$\mu^*_{j-1} + \min\{\mu_j - \mu^*_{j-1}, \mu^*_{j-1} - \mu^*_{j-2}\}.$$

To see this, simply observe that the above term can take two values, either μ_j or $\mu^*_{j-1} +$

$\mu^*_{j-1} - \mu^*_{j-2}$. Since, by definition of $[C^*_j]$, $\mu^*_j \le \mu_j$ the result holds in the former case. Since by (i) above $q_\ell \le \mu^*_{j-1} - \mu^*_{j-2}$ for all $\ell \ge j-1$ for any $q \in Q_{j-1}$, and since, by definition, the value of $[C^*_j]$ equals $\mu^*_{j-1} + q^*_j$, the result also holds in the latter case.

We next prove that the value is actually attained. Construct $\hat{q}$ such that $\hat{q}_i = q^*_i$ for all $i \le j-1$, $\hat{q}_j = \mu^*_j - \mu^*_{j-1}$ and $\hat{q}_i = 0$ for all $i \ge j+1$. Note that, since $\mu^*_i - \mu^*_{i-1}$ is nonincreasing in i, $\hat{q}_j \le \hat{q}_i$ for all $i \le j$. Take any $S \subset [k]$ such that $|S| = \ell < j$. Then,

$$\sum_{i \in S} \hat{q}_i \le \sum_{i \in [\ell]} \hat{q}_i = \mu^*_\ell \le \mu_\ell,$$

where the first inequality follows from (iii) and the second follows from (ii). Next, take any $S \subset [k]$ such that $|S| = k$. Then,

$$\sum_{i \in S} \hat{q}_i \le \sum_{i \in [j]} \hat{q}_i = \mu^*_j \le \mu_j,$$

where the first follows from (iii) and the fact that $q^*_j \le q^*_i$ for all $i \le j$, and the second follows from our prior observation that the value μ^*_j of $[C^*_j]$ must be smaller than μ_j. Lastly, it is trivial that $\sum_{i \in S} \hat{q}_i \le \mu_\ell$, for any S with $|S| = \ell$, where $\ell > j$. We thus conclude $\hat{q} \in Q_j$ and μ^*_j is the value of $[C^*_j]$. ∎

It is easy to verify that the optimal solution $(q^*_j)_{j \in [k]}$ is unique. More importantly, one can see that the solution coincides with the service rate we define for FCFS in the main text, provided that FCFS is work conserving. To see this note from Lemma S1 that $\sum_{j \in [k]} q^*_j = \mu^*_k$. Hence, if FCFS is work conserving, we must have $\mu^*_k = \mu_k$ for each k (since $\mu^*_k \le \mu_k$ for each k). In that case, we get $q^*_j = \mu_j - \mu_{j-1}$, precisely as we defined in the text.

Axiomatic Characterization: We now prove that regularity of μ is a necessary and sufficient condition for FCFS to be work-conserving, i.e., $\sum_{i \in [k]} q^*_{k,i} = \mu_k$ for all k.

Theorem S1. FCFS is work-conserving if and only if μ is regular.

Proof. By Lemma S1, for all $k \in \mathbb{N}$, $\sum_{i \in [k]} q^*_i = \mu^*_k$, and by feasibility $\mu^*_k \le \mu_k$. Hence, FCFS is work-conserving if and only if $\mu^*_k = \mu_k$ for all $k \in \mathbb{N}$. Thus, it suffices to prove that μ is regular if and only if $\mu^*_k = \mu_k$ for all k.

To prove the "only if" direction, suppose μ is regular. We argue inductively that $\mu^*_k = \mu_k$ for all k. First, by definition, $\mu^*_1 = \mu_1$. Suppose $\mu^*_i = \mu_i$ for all $i \in [k-1]$. Then,

$$\begin{aligned} \mu^*_k &= \mu^*_{k-1} + \min\{\mu_k - \mu^*_{k-1}, \mu^*_{k-1} - \mu^*_{k-2}\} \\ &= \mu_{k-1} + \min\{\mu_k - \mu_{k-1}, \mu_{k-1} - \mu_{k-2}\} = \mu_k, \end{aligned}$$

where the first equality is by definition of μ^*_k, the second follows from the induction hypothesis, and the last follows from the regularity.

The converse, the "if" direction, follows from the fact that $\mu_k - \mu_{k-1} = \mu^*_k - \mu^*_{k-1} = q^*_k$ and q^*_k is nonincreasing in k by Lemma S1. ∎

We have focused only on FCFS, but the LCFS can be defined analogously, and a similar result is obtained.

S.2 Proof of Lemma A.5

We have

$$\begin{aligned}
\sum_{k=0}^{K} p'_k \varphi(k) &= \varphi(K) - \sum_{L=0}^{K-1}\left(\sum_{k=0}^{L} p'_k\right)(\varphi(L+1)-\varphi(L)) \\
&= \varphi(K) - \sum_{L=0}^{K-1}\left(1-\sum_{k=L+1}^{K} p'_k\right)(\varphi(L+1)-\varphi(L)) \\
&> \varphi(K) - \sum_{L=0}^{K-1}\left(1-\sum_{k=L+1}^{K} p_k\right)(\varphi(L+1)-\varphi(L)) \\
&= \varphi(K) - \sum_{L=0}^{K-1}\left(\sum_{k=0}^{L} p_k\right)(\varphi(L+1)-\varphi(L)) \\
&= \sum_{k=0}^{K} p_k \varphi(k),
\end{aligned}$$

where the first and the last equalities hold by Abel's formula for summation by parts while the strict inequality uses the fact that (1) p' stochastically dominates p; (2) φ is a nondecreasing function and (3) there is $\kappa \geq 1$ such that

$$\sum_{k=\kappa}^{K} p'_k > \sum_{k=\kappa}^{K} p_k \text{ and } \varphi(\kappa) > \varphi(\kappa-1).$$

S.3 Remaining Proof of Theorem 1

In this section we prove Proposition A.2 which completes the proof of Theorem 1.

S.3.0.1 Existence of a solution in the infinite-dimensional problem

Our problem $[P']$ can be written as

$$[P'] \qquad \max_{p\in M'} \sum_{k=0}^{\infty} p_k \left[\mu_k((1-\alpha)R+\alpha V) - \alpha C k\right]$$

where $M' \triangleq \{p \in \Delta(\mathbb{Z}_+) : \sum_{k=0}^{\infty} p_k [\mu_k V - Ck] \geq 0, \lambda_k p_k \geq \mu_{k+1} p_{k+1}, \forall k\}$. We prove the following result.

Proposition S1. The set of optimal solutions of $[P']$ is nonempty.

We start by showing that the objective of the optimization problem is upper semi-continuous (Proposition S2). We endow $\mathbb{Z}_+$ with the discrete topology and $\Delta(\mathbb{Z}_+)$ with the weak topology. Since $\mathbb{Z}_+$ endowed with the discrete topology is a (separable) metric space, $\Delta(\mathbb{Z}_+)$ is metrizable by Prokhorov's Theorem. We next show that set M' is compact (Proposition S3). This enough for our purpose. Indeed, by the Extreme Value Theorem for upper semi-continuous functions, optimization problem $[P']$ has an optimal solution.

Proposition S2. The function

$$\sum_{k=0}^{\infty} p_k \left[\mu_k((1-\alpha)R + \alpha V) - \alpha C k\right]$$

is upper semi-continuous in $p \in \Delta(\mathbb{Z}_+)$.

Proof. Consider a sequence $\{p^n\}$ in $\Delta(\mathbb{Z}_+)$ converging to p^*. Since the function $k \mapsto \mu_k((1-\alpha)R+\alpha V)-\alpha Ck$ is continuous (in the discrete topology) and upper bounded,[A.2] by Portmanteau's Theorem, $\limsup \sum_{k=0}^{\infty} p_k^n \left[\mu_k((1-\alpha)R + \alpha V) - \alpha C k\right] \leq \sum_{k=0}^{\infty} p_k^* \left[\mu_k((1-\alpha)R + \alpha V) - \alpha C k\right]$ and so we get the upper semi-continuity of our function. ∎

Proposition S3. Set M' is compact.

Proof. The proof is based on the two lemmas proved below.

Lemma S2. The set M' is tight.

Proof. We need to show that for any $\varepsilon > 0$, there is n large enough so that any probability measure $p \in M'$ has $\sum_{k=n+1}^{\infty} p_k < \varepsilon$. Suppose to the contrary that there is $\varepsilon > 0$ and a sequence $\{p^n\}_n$ in M' (which satisfies $\sum_{k=0}^{\infty} p_k^n \left[\mu_k V - Ck\right] \geq 0$) such that $\sum_{k=n+1}^{\infty} p_k^n > \varepsilon$ for all n. This implies

$$\begin{aligned}
\sum_{k=0}^{\infty} p_k^n(\mu_k V - Ck) &= V \sum_{k=0}^{\infty} p_k^n \mu_k - C \sum_{k=0}^{\infty} p_k^n k \\
&\leq \sup_k \mu_k V - C \sum_{k=n+1}^{\infty} p_k^n k \\
&\leq \sup_k \mu_k V - C(n+1) \sum_{k=n+1}^{\infty} p_k^n \\
&\leq \sup_k \mu_k V - C(n+1)\varepsilon.
\end{aligned}$$

Note that for n large enough, using our assumption that $\sup_k \mu_k < +\infty$, the above term must be strictly negative. This contradicts the fact that $\sum_{k=0}^{\infty} p_k^n(\mu_k V - Ck) \geq 0$ for all n. ∎

[A.2] Recall our assumption that μ_k is uniformly bounded.

Lemma S3. The set M' is closed.

Proof. To show that M' is closed, we need to show that it contains all its limit points. Recall that since $\Delta(\mathbb{Z}_+)$ is a metric space, $p \in \Delta(\mathbb{Z}_+)$ is a limit point of M' if and only if there is a sequence of points in $M' \backslash \{p\}$ converging to p. Take any sequence $\{p^n\}_n$ in M' converging to p^*. We need to show that (1) $\sum_{k=0}^{\infty} p_k^*(\mu_k V - Ck) \geq 0$ and (2) for all k, $\lambda_k p_k^* \geq \mu_{k+1} p_{k+1}^*$.

(1) $\sum_{k=0}^{\infty} p_k^*(\mu_k V - Ck) \geq 0$. Proceed by contradiction and assume that $\sum_{k=0}^{\infty} p_k^*(\mu_k V - Ck) < 0$. By Portmanteau's Theorem, since the function $k \mapsto \mu_k V - Ck$ is bounded above (and trivially continuous in the discrete topology), we must have that $\limsup \sum_{k=0}^{\infty} p_k^n(\mu_k V - Ck) \leq \sum_{k=0}^{\infty} p_k^*(\mu_k V - Ck)$. Hence, since, by assumption, $\sum_{k=0}^{\infty} p_k^*(\mu_k V - Ck) < 0$, it must be that for n large enough, $\sum_{k=0}^{\infty} p_k^n(\mu_k V - Ck) < 0$, a contradiction with the fact that $p^n \in M'$.

(2) For all k, $\lambda_k p_k^* \geq \mu_{k+1} p_{k+1}^*$. By contradiction, assume that for some k, $\lambda_k p_k^* < \mu_{k+1} p_{k+1}^*$. Since p_k^n and p_{k+1}^n converge pointwise to p_k^* and p_{k+1}^*, for n large enough we have $\lambda_k p_k^n < \mu_{k+1} p_{k+1}^n$ which contradicts the fact that p^n is in M'. ∎

Since M' is closed and tight, by Prokhorov Theorem, M' must be sequentially compact. Since $\Delta(\mathbb{Z}_+)$ is a metric space, this implies that M' is compact, as claimed. ∎

S.3.0.2 Completion of the proof of Proposition A.2

Let M'' be the set of p's in M' which exhibits a cutoff policy. That is any $p \in M''$ satisfies for some $\hat{K}$, $\lambda_k p_k = \mu_{k+1} p_{k+1}, \forall k = 0, ..\hat{K} - 1$ and $p_k = 0$ for all $k \geq \hat{K} + 1$. We define the sequence $\{p^K\}_K$ where, for each K, p^K is an optimal solution of $[P_K']$. If μ is regular, we assume that p^* exhibits a cutoff policy which is well-defined by Proposition A.1. In addition, for each K, we see p^K as a point in $\mathbb{R}^{\mathbb{Z}_+}$ with $p_k^K = 0$ for all $k \geq K + 1$. Clearly $\{p^K\}_K$ is a sequence in M''. In the next proposition we show that M'' is (sequentially) compact. This will show that $\{p^K\}_K$ must have a subsequence converging to a point that exhibits a cutoff policy. In the sequel, we assume that μ is regular.

Proposition S4. $\{p^K\}_K$ must have a subsequence converging to a feasible point p^* of $[P']$ that exhibits a cutoff policy. In addition, $p_k^* > 0$ for each $k \leq \min \arg\max_k \mu_k V - Ck$.

Proof. For the first part of the statement, it suffices to show that M'' is (sequentially) compact. Since M'' is a subset of M' which is compact (Proposition S3), we only need to show that M'' is closed. Consider a sequence $\{p^n\}$ in M'' converging to p^*. We show that $p^* \in M''$. Since M' is (sequentially) compact, we already know that $p^* \in M'$. Letting $\hat{K}$ be the largest state in the support of p^* (which is potentially $+\infty$ if the support is unbounded), we proceed by contradiction and assume that there exists $k_0 < \hat{K}$ such that $\lambda_{k_0-1} p_{k_0-1}^* > \mu_{k_0} p_{k_0}^*$. Now, simply pick n large enough so that (1) $p_k^n > 0$ for all $k = 0, ..., k_0+1$ and (2) $\lambda_{k_0-1} p_{k_0-1}^n > \mu_{k_0} p_{k_0}^n$. This contradicts the assumption that p^n is in M''. We thus conclude that $p^* \in M''$.

We now show the second part of the statement. We just proved that $\{p^K\}_K$ must have a subsequence converging to a feasible point p^* of $[P']$. We show that p^* satisfies

$p_k^* > 0$ for each $k \leq \min\arg\max_k \mu_k V - Ck$. First, we simply observe that for any $\xi \geq 0$, $\min\arg\max_k \mu_k V - Ck \leq \min\arg\max f(k;\xi)$.[A.3] Now, we proceed by contradiction and assume that there is $k_0 \leq \min\arg\max_k \mu_k V - Ck$ such that $p_{k_0}^* = 0$. Let us assume that k_0 is the smallest state satisfying this property, so, in particular, $p_{k_0-1}^* > 0$. This implies that $p_{k_0}^* \mu_{k_0} < p_{k_0-1}^* \lambda_{k_0-1}$. Since $\{p^K\}_K$ converges to p^*, for K large enough, $p_{k_0}^K \mu_{k_0} < p_{k_0-1}^K \lambda_{k_0-1}$. Since $k_0 \leq \min\arg\max f(k;\xi_K^*)$, using single-peakedness of $f(\cdot;\xi_K^*)$, we must have $f(k_0-1;\xi_K^*) < f(k_0;\xi_K^*)$ (where we use the notation (p^K, ξ_K^*) for the saddle point of the Lagrangian in $[P_K']$). This contradicts Lemma A.4. ∎

Finally, we complete the proof of Proposition A.2 via the following proposition.

Proposition S5. Take any subsequence of $\{p^K\}_K$ converging to a limit p^*. Then, p^* must be an optimal solution of $[P']$.

Proof. In the sequel, we let p^* be the limit of an arbitrary converging subsequence $\{p^K\}_K$. We proceed by contradiction and assume that p^* is not a solution to the infinite dimensional problem. By Proposition S1, we know that there is a solution to this problem. Let us call it $\bar{p}$. By assumption,

$$\sum_{k=0}^{\infty} \bar{p}_k \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] > \sum_{k=0}^{\infty} p_k^* \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right]. \tag{A.1}$$

Now, let us note by $\bar{p}^K$ the distribution $\bar{p}$ conditional on $\{0,...,K\}$, i.e., $\bar{p}_k^K = 0$ for all $k \geq K+1$ while $\bar{p}_k^K = \bar{p}_k \Big/ \sum_{k=0}^{K} \bar{p}_k$ for all $k \leq K$. We claim that

$$\lim \sum_{k=0}^{\infty} \bar{p}_k^K \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] = \sum_{k=0}^{\infty} \bar{p}_k \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right].$$

Indeed, by construction, for each K,

$$\sum_{k=0}^{\infty} \bar{p}_k^K \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] = \sum_{k=0}^{K} \bar{p}_k \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] \Bigg/ \sum_{k=0}^{K} \bar{p}_k .$$

Taking limits on both sides as $K \to \infty$ (and using the fact that $\lim_{K\to\infty} \sum_{k=0}^{K} \bar{p}_k = 1$), we obtain

$$\lim_{K\to\infty} \sum_{k=0}^{\infty} \bar{p}_k^K \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] = \sum_{k=0}^{\infty} \bar{p}_k \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right],$$

as claimed.

[A.3] Straightforward algebra show that $\mu_{k+1} V - C(k+1) > \mu_k V - Ck$ if and only if $\mu_{k+1} - \mu_k > C/V$. Similarly, given $\xi \geq 0$, we have that $f(k+1;\xi) > f(k;\xi)$ if and only if $\mu_{k+1} - \mu_k > C/[(1-\alpha)/(\alpha+\xi) + V]$. Hence, whenever $\mu_k V - Ck$ is strictly increasing from k to $k+1$, so is $f(k;\xi)$. Since by Lemma A.3, these functions are single-peaked, we must have $\min\arg\max_k \mu_k V - Ck \leq \min\arg\max f(k;\xi)$.

Now, using Equation (A.1), for K large enough, we must have

$$\sum_{k=0}^{\infty} \bar{p}_k^K \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] > \sum_{k=0}^{\infty} p_k^* \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] + \varepsilon \tag{A.2}$$

for some $\varepsilon > 0$. Now, since $\{p^K\}_K$ converges weakly to p^*, by Proposition S2,

$$\lim_{K\to\infty} \sup \sum_{k=0}^{\infty} p_k^K \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] \leq \sum_{k=0}^{\infty} p_k^* \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right].$$

Hence, we must have that for K large enough,

$$\sum_{k=0}^{\infty} p_k^* \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] + \varepsilon > \sum_{k=0}^{\infty} p_k^K \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right]. \tag{A.3}$$

Using Equation (A.2) and (A.3), we conclude that for K large enough,

$$\sum_{k=0}^{\infty} \bar{p}_k^K \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right] > \sum_{k=0}^{\infty} p_k^K \left[\mu_k \left((1-\alpha) R + \alpha V\right) - \alpha C k\right].$$

This contradicts the fact that p^K is an optimal solution of $[P'_K]$ since $\bar{p}^K$ is feasible in this problem. ∎

S.4 Proof that FCFS with no information satisfies (IC_0).

Recall that policy (q^*, I^*) stands for FCFS queueing rule and the no information (beyond recommendations) rule.

Lemma S4. The queueing/information policy (q^*, I^*) satisfies (IC_0).

Proof. Recall the optimality of the cutoff policy means $x_k^* = 1$ for all $k = 0, ..., K^* - 2$ and $x_k^* = 0$ for all $k > K^* - 1$, and $y_{k,\ell}^* = z_{k,\ell}^* = 0$ for all (k, ℓ). Substitute these into (*B*). Use the resulting equations to rewrite (1):

$$\tilde{\gamma}_\ell^0 = \frac{p_\ell^* \mu_\ell}{\sum_{i=1}^{K^*} p_i^* \mu_i}, \forall \ell = 1, ..., K^*.$$

An agent's expected payoff when joining the queue after being recommended to do so is:

$$\begin{aligned} V - C \sum_{k=1}^{K^*} \tilde{\gamma}_k^0 \cdot \tau_k^* =& V - C \frac{\sum_{k=1}^{K^*} p_k^* \mu_k \cdot \tau_k^*}{\sum_{i=1}^{K^*} p_i^* \mu_i} \\ =& V - C \frac{\sum_{k=1}^{K^*} p_k^* k}{\sum_{i=1}^{K^*} p_i^* \mu_i} \end{aligned}$$

$$= \left(\frac{1}{\sum_{i=1}^{K^*} p_i^* \mu_i} \right) \sum_{k=1}^{K^*} p_k^* \left(\mu_k V - kC \right),$$

where the first equality is from the preceding observation and the second equality follows from Lemma 1. Since $\sum_{i=1}^{K^*} p_i^* \mu_i > 0$, (IC_0) holds if and only if (IR) holds. ∎

S.5 Proof of Lemma 2 when $\bar{K} = \infty$.

We first derive the infinite system of ODEs in terms of agents' belief of occupying queue position $\ell = 1, ..., \infty$ at time t. It follows from (2), together with $q_i^* = \mu_i - \mu_{i-1}$, that

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

$$\begin{aligned}
\frac{\tilde{\gamma}_k^{t+dt} - \tilde{\gamma}_k^t}{dt} &= \frac{(1-\mu_k dt)\tilde{\gamma}_k^t + \mu_k dt \tilde{\gamma}_{k+1}^t}{dt \sum_{i=1}^{\infty} \tilde{\gamma}_i^t (1 - q_i^* dt)} - \frac{\tilde{\gamma}_k^t}{dt} + \frac{o(dt)}{dt} \\
&= \frac{(1-\mu_k dt)\tilde{\gamma}_k^t + \mu_k dt \tilde{\gamma}_{k+1}^t}{dt \left[1 - dt \sum_{i=1}^{\infty} \tilde{\gamma}_i^t q_i^*\right]} - \frac{\tilde{\gamma}_k^t}{dt} + \frac{o(dt)}{dt} \\
&= \frac{(1-\mu_k dt)\tilde{\gamma}_k^t + \mu_k dt \tilde{\gamma}_{k+1}^t}{dt \left[1 - dt \sum_{i=1}^{\infty} \tilde{\gamma}_i^t (\mu_i - \mu_{i-1})\right]} - \frac{\tilde{\gamma}_k^t}{dt} + \frac{o(dt)}{dt} \\
&= \frac{(1-\mu_k dt)\tilde{\gamma}_k^t + \mu_k dt \tilde{\gamma}_{k+1}^t - \tilde{\gamma}_k^t \left[1 - dt \sum_{i=1}^{\infty} \tilde{\gamma}_i^t (\mu_i - \mu_{i-1})\right]}{dt \left[1 - dt \sum_{i=1}^{\infty} \tilde{\gamma}_i^t (\mu_i - \mu_{i-1})\right]} + \frac{o(dt)}{dt} \\
&= \frac{-\mu_k \tilde{\gamma}_k^t + \mu_k \tilde{\gamma}_{k+1}^t + \tilde{\gamma}_k^t \left[\sum_{i=1}^{\infty} \tilde{\gamma}_i^t (\mu_i - \mu_{i-1})\right]}{\left[1 - dt \sum_{i=1}^{\infty} \tilde{\gamma}_i^t (\mu_i - \mu_{i-1})\right]} + \frac{o(dt)}{dt}.
\end{aligned}$$

Letting $dt \to 0$, we obtain: for all $k \in \mathbb{N}$,

$$\dot{\tilde{\gamma}}_k^t = -\mu_k \tilde{\gamma}_k^t + \mu_k \tilde{\gamma}_{k+1}^t + \tilde{\gamma}_k^t \left[\sum_{i=1}^{\infty} \tilde{\gamma}_i^t (\mu_i - \mu_{i-1})\right] \triangleq f_k(\tilde{\gamma}^t), \tag{A.4}$$

and let $f \triangleq (f_k)_{k \in \mathbb{N}}$. The following proposition states that, given an initial condition, this system of ODEs has a unique solution.

Proposition S6. For any initial condition in $\Delta(\mathbb{N})$, there is a unique solution to the system of ODEs given by (A.4).

Proof. Let $\mathbf{X}$ be the set of sequences in ℓ^1-space endowed with ℓ^1-norm. As is well-known, this is a Banach space. Clearly, $\Delta(\mathbb{N}) \subseteq \mathbf{X}$. Further, we can see that f maps from $\mathbf{X}$ to $\mathbf{X}$.

Indeed, for any $\tilde{\gamma}^t \in \mathbf{X}$:

$$
\begin{aligned}
||f(\tilde{\gamma}^t)|| &= \sum_{k=1}^{\infty} \left| f_k(\tilde{\gamma}_k^t) \right| \\
&= \sum_{k=1}^{\infty} \left| -\mu_k \tilde{\gamma}_k^t + \mu_k \tilde{\gamma}_{k+1}^t + \tilde{\gamma}_k^t \left[\sum_{i=1}^{\infty} \tilde{\gamma}_i^t (\mu_i - \mu_{i-1}) \right] \right| \\
&\leq \sum_{k=1}^{\infty} \left| -\mu_k \tilde{\gamma}_k^t \right| + \sum_{k=1}^{\infty} \left| \mu_k \tilde{\gamma}_{k+1}^t \right| + \sum_{k=1}^{\infty} \left| \tilde{\gamma}_k^t \left[\sum_{i=1}^{\infty} \tilde{\gamma}_i^t (\mu_i - \mu_{i-1}) \right] \right| \\
&\leq \bar{\mu} \sum_{k=1}^{\infty} \left| \tilde{\gamma}_k^t \right| + \bar{\mu} \sum_{k=1}^{\infty} \left| \tilde{\gamma}_{k+1}^t \right| + \bar{\mu} \left(\sum_{k=1}^{\infty} \left| \tilde{\gamma}_k^t \right| \right) \left(\sum_{i=1}^{\infty} \left| \tilde{\gamma}_i^t \right| \right) < \infty
\end{aligned}
$$

where we recall that $\bar{\mu} \triangleq \sup_k \mu_k < \infty$ and use the fact that $\tilde{\gamma}^t \in \mathbf{X}$. Hence, we have $f(\tilde{\gamma}^t) \in \mathbf{X}$.

Lemma S5. Consider the restriction of f defined as follows $f : U \to \mathbf{X}$ where $U \triangleq \{\{x_k\}_{k \geq 1} \in \mathbf{X} : \sum_{k=1}^{\infty} |x_k| < 1 + \varepsilon\} \subset \mathbf{X}$, for some $\varepsilon > 0$, is an open set containing $\Delta(\mathbb{N})$. Mapping f (restricted to U) is Lipschitz continuous.

Proof. Indeed, for any $\tilde{\gamma}$ and $\tilde{\gamma}'$ in U,

$$
\begin{aligned}
\|f(\tilde{\gamma}') - f(\tilde{\gamma})\| &= \sum_{k=1}^{\infty} |f_k(\tilde{\gamma}') - f_k(\tilde{\gamma})| \\
&\leq \sum_{k=1}^{\infty} |-\mu_k \tilde{\gamma}'_k + \mu_k \tilde{\gamma}_k| + \sum_{k=1}^{\infty} |\mu_k \tilde{\gamma}'_{k+1} - \mu_k \tilde{\gamma}_{k+1}| \\
&\quad + \sum_{k=1}^{\infty} \left| \tilde{\gamma}'_k \left[\sum_{i=1}^{\infty} \tilde{\gamma}'_i (\mu_i - \mu_{i-1}) \right] - \tilde{\gamma}_k \left[\sum_{i=1}^{\infty} \tilde{\gamma}_i (\mu_i - \mu_{i-1}) \right] \right| \\
&\leq \sum_{k=1}^{\infty} \mu_k |\tilde{\gamma}'_k - \tilde{\gamma}_k| + \sum_{k=1}^{\infty} \mu_k |\tilde{\gamma}'_{k+1} - \tilde{\gamma}_{k+1}| \\
&\quad + \max\{\sum_{i=1}^{\infty} |\tilde{\gamma}'_i| (\mu_i - \mu_{i-1}), \sum_{i=1}^{\infty} |\tilde{\gamma}_i| (\mu_i - \mu_{i-1})\} \sum_{k=1}^{\infty} |\tilde{\gamma}'_k - \tilde{\gamma}_k| \\
&\leq \bar{\mu} \|\tilde{\gamma}' - \tilde{\gamma}\| + \bar{\mu} \|\tilde{\gamma}' - \tilde{\gamma}\| + (1 + \varepsilon) \bar{\mu} \|\tilde{\gamma}' - \tilde{\gamma}\| .
\end{aligned}
$$

Thus, f restricted to U is Lipschitz continuous with Lipschitz constant equal to $\bar{\mu}(3 + \varepsilon)$. ∎

In order to complete the proof of Proposition S6, let us consider the system of ODEs given by (A.4) where the vector field f is the mapping from $\mathbf{X}$ to $\mathbf{X}$. Since f is bounded and, by Lemma S5, Lipschitz continuous on $\Delta(\mathbb{N})$ and $\Delta(\mathbb{N})$ is positively invariant, existence and uniqueness of a solution for our system of ODEs with initial condition in $\Delta(\mathbb{N})$ follows from Picard-Lindelöf Theorem on Banach spaces.[A.4] ∎

In the sequel, we consider solutions to the system of ODEs when the entry rule x^* is "truncated" to x^K, i.e., where $x_k^K = x_k^* = 1$ for all $k \leq K$ and $x_k^K = 0$ otherwise. We show that solutions to the system of ODEs under the truncated cutoff policy (x^K, y^*, z^*) approximate solutions to the system under the original cutoff policy (x^*, y^*, z^*). More specifically, we let $\tilde{\gamma}^K(t) = (\tilde{\gamma}_k^K(t))_{k\in\mathbb{N}}$ denote a solution to the system given by (A.4)

$$\dot{\tilde{\gamma}}^t = f(\tilde{\gamma}^t),$$

when $\tilde{\gamma}^0 = \tilde{\gamma}^K(0) \triangleq (\tilde{\gamma}_k^K(0))_{k\in\mathbb{N}}$ where $\tilde{\gamma}_k^K(0)$ is an agent's belief of entering the queue with position k at $t = 0$ under the truncated cutoff policy.[A.5] Meanwhile, $\tilde{\gamma}^\infty(t) = (\tilde{\gamma}_k^\infty(t))_{k\in\mathbb{N}}$ denotes a solution to this system of ODEs when $\tilde{\gamma}^0 = \tilde{\gamma}^\infty(0) \triangleq (\tilde{\gamma}_k^\infty(0))_{k\in\mathbb{N}}$ where $\tilde{\gamma}_k^\infty(0)$ is an agent's belief of entering the queue with position k at $t = 0$ under the original cutoff policy. We show that solution $\tilde{\gamma}^K$ converges to solution $\tilde{\gamma}^\infty$ when K goes to infinity.

Lemma S6. The solution $\tilde{\gamma}^K$ converges pointwise to the solution $\tilde{\gamma}^\infty$, i.e., for each $t > 0$,

$$\lim_{K\to\infty} ||\tilde{\gamma}^K(t) - \tilde{\gamma}^\infty(t)|| = 0.$$

Proof. The following two steps prove the lemma.

Step 1. $\left\|\tilde{\gamma}^K(0) - \tilde{\gamma}^\infty(0)\right\| \to 0$ as $K \to \infty$.

Proof. We know that for all $\ell = 2, ..., K : \tilde{\gamma}_\ell^K(0) = \prod_{i=2}^\ell r_i^0 \tilde{\gamma}_1^K(0) = \prod_{i=2}^\ell \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}} \tilde{\gamma}_1^K(0)$, where we used (B.9), while $\tilde{\gamma}_\ell^K(0) = 0$ for $\ell \geq K + 1$. In addition, we know that $\tilde{\gamma}_\ell^\infty(0) = \prod_{i=2}^\ell \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}} \tilde{\gamma}_1^\infty(0)$ and $\sum_{k=1}^\infty \prod_{i=2}^k \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}} \tilde{\gamma}_1^\infty(0) = 1$ where our convention is that $\prod_{i=2}^1 \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}} \triangleq 1$. Thus,

$$\tilde{\gamma}_1^\infty(0) = \frac{1}{\sum_{k=1}^\infty \prod_{i=2}^k \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}}}.$$

Note that this implies that $\sum_{k=1}^\infty \prod_{i=2}^k \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}} < \infty$. Similar computation yields

$$\tilde{\gamma}_1^K(0) = \frac{1}{\sum_{k=1}^K \prod_{i=2}^k \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}}}.$$

Note that $\left|\tilde{\gamma}_1^K(0) - \tilde{\gamma}_1^\infty(0)\right| \to 0$ as K increases. We have

$$\left\|\tilde{\gamma}^K(0) - \tilde{\gamma}^\infty(0)\right\| = \sum_{k=1}^\infty \left|\tilde{\gamma}_k^K(0) - \tilde{\gamma}_k^\infty(0)\right|$$

[A.4] Recall that a subset S of $\mathbf{X}$ is positively invariant if no solution starting inside S can leave S in the future.

[A.5] Note that $\tilde{\gamma}_k^K(0) = 0$ for all $k > K$.

$$= \sum_{k=1}^{K} \left| \tilde{\gamma}_k^K(0) - \tilde{\gamma}_k^\infty(0) \right| + \sum_{k=K+1}^{\infty} \left| \tilde{\gamma}_k^\infty(0) \right|$$

$$= \sum_{k=1}^{K} \left| \prod_{i=2}^{k} \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}} \tilde{\gamma}_1^K(0) - \prod_{i=2}^{k} \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}} \tilde{\gamma}_1^\infty(0) \right| + \sum_{k=K+1}^{\infty} \left| \tilde{\gamma}_k^\infty(0) \right|$$

$$= \left| \tilde{\gamma}_1^K(0) - \tilde{\gamma}_1^\infty(0) \right| \sum_{k=1}^{K} \prod_{i=2}^{k} \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}} + \sum_{k=K+1}^{\infty} \left| \tilde{\gamma}_k^\infty(0) \right|.$$

Since $\left| \tilde{\gamma}_1^K(0) - \tilde{\gamma}_1^\infty(0) \right| \to 0$ as $K \to \infty$, $\sum_{k=1}^{\infty} \prod_{i=2}^{k} \frac{\tilde{\lambda}_{i-1}}{\mu_{i-1}} < \infty$, and $\sum_{k=K+1}^{\infty} \left| \tilde{\gamma}_k^\infty(0) \right|$ goes to 0 as $K \to \infty$, the result follows. ∎

Step 2. For each $t > 0$,

$$\lim_{K \to \infty} \sum_{k=1}^{\infty} \left| \tilde{\gamma}_k^K(t) - \tilde{\gamma}_k^\infty(t) \right| = 0.$$

Proof. By Grönwall's inequality,

$$\left\| \tilde{\gamma}^K(t) - \tilde{\gamma}^\infty(t) \right\| \le e^{Ct} \left\| \tilde{\gamma}^K(0) - \tilde{\gamma}^\infty(0) \right\|,$$

where, by Lemma S5, $C \triangleq \bar{\mu}(3 + \varepsilon)$ is the Lipschitz constant for the Lipschitz continuous function f restricted to open set $U = \{\{x_k\}_{k \ge 1} \in \mathbf{X} : \sum_{k=1}^{\infty} |x_k| < 1 + \varepsilon\}$. The result then follows from Step 1. ∎

∎

We now complete the proof of Lemma 2 when $\bar{K} = \infty$ with the following lemma.

Lemma S7. $\dot{r}_\ell^\infty(t) \le 0$ for all $\ell \ge 2$ and t, where $r_\ell^\infty(t) = \tilde{\gamma}_\ell^\infty(t) / \tilde{\gamma}_{\ell-1}^\infty(t)$ for all $\ell \ge 2$.

Proof. Recall from Equation (B.8) in the main text that the system of ODEs is given by

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

for all $\ell \ge 2$. Suppose to the contrary that $\dot{r}_\ell^\infty(t) > 0$ for some ℓ and t. We already proved in Appendix B.2 (in the main text) where $\bar{K} < \infty$ that

$$\dot{r}_\ell^K(t) = r_\ell^K(t) \left(\mu_{\ell-1} - \mu_\ell - \mu_{\ell-1} r_\ell^K(t) + \mu_\ell r_{\ell+1}^K(t) \right) \le 0$$

for all $K < \infty$, ℓ and t. To show a contradiction, it is enough to prove that

$$r_\ell^K(t) \left(\mu_{\ell-1} - \mu_\ell - \mu_{\ell-1} r_\ell^K(t) + \mu_\ell r_{\ell+1}^K(t) \right) \to r_\ell^\infty(t) \left(\mu_{\ell-1} - \mu_\ell - \mu_{\ell-1} r_\ell^\infty(t) + \mu_\ell r_{\ell+1}^\infty(t) \right)$$

as $K \to \infty$. To this end, it suffices to show that $r_\ell^K(t)$ and $r_{\ell+1}^K(t)$ converge respectively to $r_\ell^\infty(t)$ and $r_{\ell+1}^\infty(t)$. It follows from Lemma S6 that for each k :

$$\lim_{K\to\infty} \tilde{\gamma}_k^K(t) = \tilde{\gamma}_k^\infty(t).$$

By assumption $\tilde{\gamma}_k^\infty(0) > 0$ for all k, so

$$\lim_{K\to\infty} r_\ell^K(t) = \lim_{K\to\infty} \frac{\tilde{\gamma}_\ell^K(t)}{\tilde{\gamma}_{\ell-1}^K(t)} = \frac{\tilde{\gamma}_\ell^\infty(t)}{\tilde{\gamma}_{\ell-1}^\infty(t)} = r_\ell^\infty(t)$$

Similarly,

$$\lim_{K\to\infty} r_{\ell+1}^K(t) = r_{\ell+1}^\infty(t),$$

which completes the argument. ∎

S.6 Generalization of Naor (1969)

In this section, we generalize Naor (1969)'s classic result (obtained for the $M/M/1$ queue) to our more general Markov process: *agents would have excess incentives to queue under FCFS with full information.* Since the designer can simply stop excessive queueing, this means that FCFS with a full information rule, denoted I^{FI}, can be used to achieve the optimal cutoff policy.

Proposition S7. Suppose $\alpha = 1$ and μ is regular. Then, FCFS with full information, I^{FI}, implements the optimal cutoff outcome (x^*, y^*, p^*).

Proof. Consider FCFS with full information. We need to show that (IC_t) holds for all $t \geq 0$. With the full information rule, we only need to show that (IC_0) holds. By Lemma 1, condition (IC_0) can be written as:

$$V - C\frac{k}{\mu_k} \geq 0 \iff \mu_k V - Ck \geq 0 \tag{A.5}$$

for all $k \leq K^*$. In the sequel, we let K^{FI} be the largest integer satisfying (A.5). We know that, by regularity of μ, $k \mapsto \mu_k V - Ck$ is single-peaked (by Lemma A.3 for $\alpha = 1$ and $\xi = 0$). Hence, K^{FI} is well-defined (i.e., finite) given our assumption that μ_k is uniformly bounded. In addition, Equation (A.5) holds at state k if and only if $k \leq K^{FI}$. Hence, it is enough for our purpose to show that $K^* \leq K^{FI}$.

Proceed by contradiction and assume that the optimal cutoff policy p^*, which we recall solves $[P']$, puts strictly positive weight on $k > K^{FI}$. Note that, using again the fact that $k \mapsto \mu_k V - Ck$ is single-peaked, for any such k, $\mu_k V - Ck < 0$. Now, build p' such that $p'_k = 0$ for all $k > K^{FI}$ and $p'_k = Zp^*_k$ for all $k \leq K^{FI}$ where $Z > 1$ is set so that the sum of p'_k is equal to 1. Given that p^* satisfies (B') and given that, by construction, $p'_k/p'_{k-1} = p^*_k/p^*_{k-1}$ for all $k \leq K^{FI}$, we must have that p' also satisfies (B'). Compared to p^*, distribution p'

removes all weight on negative values and, for each positive value, increases its weight. This must strictly increase the value of the objective. It remains to show that p' satisfies (IR). The value of the objective must be positive under p^* (recall that the dirac mass on 0 brings a value of the objective of 0), and so the value of the objective must be positive under p' as well. Given that $\alpha = 1$, this implies that (IR) is satisfied. ∎

S.7 Formal Arguments for Dynamic Matching with Overloaded Lists

In this section, we explain how our results can be obtained in the setting with overloaded waiting-lists as proposed by Leshno (2019). Consider an infinite discrete time horizon model where at each period a number of agents are waiting on a wait-list. Each period t begins with arrival of an item which can be of two types, either A with probability μ_A or B with probability $\mu_B = 1 - \mu_A$ independently across periods. Period t ends when the item is assigned to an agent. Agents can be either type α or β each with probability μ_α and $\mu_\beta = 1 - \mu_\alpha$. Type α agents prefer A items to B items while type β agents prefer B over A. Agents' non-preferred item is referred to as a mismatched item. As in our main setting, all agents are infinitely lived, risk neutral, and incur a common linear waiting cost $C > 0$ per period until they are assigned. Opting out of the waiting-list is assumed to be equivalent to never getting assigned and entails a utility of $-\infty$. An agent's value of being assigned an item is V if the item is his most preferred and value 0 if assigned a mismatched item. As we already mentioned in the main text, since agents prefer to receive a mismatched item over never being assigned, taking a mismatched item for an agent could simply be intrepreted as choosing an outside option.

A mechanism decides at each date, to which agent waiting on the list is assigned the arriving item. As in Leshno (2019), we restrict our attention to *buffer-queue mechanisms* where a separate buffer-queue is held for each item. A (B) items arriving are assigned to the agents waiting on the A (B) buffer-queue if it is non-empty and to agents on the waiting-list otherwise. Those, agents from the wait-list can either accept the A (B) item or refuse. If they refuse a A (B) item, these agents are identified as β (α) type agents and enter the B (A) buffer-queue. The buffer-queue mechanism specifies the queueing discipline (i.e., how to prioritize agents within the buffer-queue) and so uses positions within the buffer-queue to decide who gets assigned the arriving item. In addition, a buffer-queue mechanism specifies the maximum number of agents in each buffer queue, say K_A (K_B) for the A (B) buffer-queue. Any buffer-queue mechanism induces a stochastic process over the number of agents in each buffer-queue. Note that at each date, one of the two buffer-queues must be empty. Hence, we can think of the state space as $\{-K_B, ..., -1, 0, 1, ..., K_A\}$ where $k \geq 0$ means that the B buffer-queue is empty and that there are k agents in the A buffer-queue. The invariant distribution of this process is denoted by $p = (p_{-K_B}, ..., p_{-1}, p_0, p_1, ..., p_{K_A})$ and characterized in Leshno (2019). The stochastic processes induced by buffer-queue mechanisms are not birth-death processes. Hence, our results do not directly apply to the environment under study. However, in the sequel, we explain how these can be adapted.

The social planner's goal is to allocate items to maximize total utility. We follow Leshno (2019) and assume that the waiting-list is "overloaded", i.e., no mechanism will ever exhaust the waiting-list. In this context, any allocation reduces total waiting costs by the same amount. Hence, the social planner can ignore waiting costs when comparing different allocations and his goal boils down to minimizing misallocations under incentive constraints.

So far the setup is the same as Leshno (2019). However, Leshno (2019) assumes that upon entering a buffer-queue, agents are informed of the length of the buffer queue and, hence, perfectly know their positions in that queue at all subsequent periods. We, however, depart from the full information rule. We allow similar information policy as in previous sections and, hence, impose similar obedience constraint, i.e., we require that conditional on the information released to agents, any agent recommended to join a buffer-queue or to stay in that queue must have an incentive to follow that recommendation. Consistently with what we proved in previous sections, we will show that FCFS with no information is optimal. Under FCFS with no information, the obedience constraints for the two buffer-queues can be written as follows. For agents recommended to enter the A buffer-queue:

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

where $\tilde{\gamma}_\ell^t$ stands for an agent's belief on having position ℓ in the A buffer-queue after spending t periods on this buffer-queue while τ_ℓ^* is the expected waiting time induced by the policy of an agent having position ℓ in the A buffer-queue. A similar condition applies to agents recommended to join the B buffer-queue.

Intuitively, if one wants to minimize misallocations, the problem of deriving the optimal buffer-queue mechanism reduces to finding the maximal size of an incentive-compatible buffer-queue mechanism. These maximal sizes K_A^* and K_B^*, for buffer-queues A and B respectively, are identified in Leshno (2019). In the sequel, we show that under FCFS with no information, when the size of the buffer-queues are set to these maximal sizes, obedience constraints are satisfied at all $t \geq 0$.

Theorem S2. Assume that the maximal sizes of buffer-queues are given by K_A^* and K_B^*. FCFS with no information satisfies the obedience constraints.

In the full information context, Leshno (2019) proves that FCFS is not optimal among incentive compatible buffer-queue mechanisms and, further, that it can be dominated by SIRO. SIRO is not incentive compatible when the maximal sizes of buffer-queues are set to K_A^* and K_B^*.[A.6] Hence, Theorem S2 not only shows that FCFS becomes optimal with a no information policy but it also shows that FCFS under no information outperforms SIRO under full information. Further, one can show that, under the no information policy, SIRO may violate obedience constraints for $t > 0$.

[A.6] SIRO is not optimal in general but it is optimal within belief-free incentive compatible mechanisms. We observe that, trivially, FCFS is belief-free incentive compatible in Leshno (2019)'s terminology.

We prove that FCFS with these sizes of buffer-queues is incentive compatible under the no information policy. Our argument parallels that of Theorem 2. Indeed, we first prove that the obedience constraints hold at $t = 0$. In a second step, we show that the likelihood ratio of beliefs about being in queue position ℓ cersus being in queue position $\ell - 1$ after spending time t on the queue.declines as t increases, meaning one's belief about getting served improves over time under FCFS with no information proving that the obedience constraints hold at all $t \geq 0$.

In order to prove both of these results, the lemma below stating that the likelihood ratio r^0_ℓ does not depend on ℓ is helpful. Leshno (2019) defines an extended Markov chain to describe the evolution of the number of agents in the buffer-queues both across periods and within a period and characterizes the invariant distribution of the process. Upon entering the A buffer-queue at a date t, an agent knows that a B item arrived and holds some beliefs over the number of agents who are ahead of him (including those who entered before him at the current date t). Conditional on a B item arriving, Leshno (2019)'s characterization states that the ratio of the likelihood of having ℓ agents over the likelihood of having $\ell - 1$ agents in the A-queue is a constant equal to $\frac{\mu_\alpha}{\mu_A}$. This yields the lemma below whose proof is provided for completeness.[A.7]

Lemma S8. Under FCFS with no information, we have $r^0_\ell = \frac{\tilde{\gamma}^0_\ell}{\tilde{\gamma}^0_{\ell-1}} = \frac{\mu_\alpha}{\mu_A}$ for all $\ell = 2, ..., K_A$.

Proof. Recall that our goal here is to show that the likelihood ratio r^ℓ of beliefs of agents entering the A-buffer queue does not depend on ℓ. A symmetric argument clearly holds for agents entering into the B-buffer queue.

Let $\mathbf{M}$ be the size of the main queue at $t = 0$ (i.e., the date at which the agent enters into the A buffer-queue) and recall that there are caps K^A and K^B on the A-buffer and B-buffer queues. For a given α-type agent, we compute the probability $\tilde{\gamma}^0_\ell$ that he gets position ℓ in the A-buffer queue conditional on the agent entering into the A-buffer queue,

$$\tilde{\gamma}^0_\ell = \frac{\sum_{\ell'=0}^{\ell} p_{\ell'} \mu_B \frac{1}{\mathbf{M}} \mu_\alpha^{\ell-\ell'}}{\sum_{\ell=1}^{K} \sum_{\ell'=0}^{\ell} p_{\ell'} \mu_B \frac{1}{\mathbf{M}} \mu_\alpha^{\ell-\ell'}}$$

for all $\ell = 1, ..., K_A$. The probability that there are ℓ' agents in the A-queue at the begining of the period is $p_{\ell'}$. Given that there are ℓ' agents in the A-queue, the event that the agent enters in the A-queue and has position ℓ in this buffer-queue corresponds to the joint event that (1) a B-item arrived (which occurs with probability μ_B), (2) the agents has queue position exactly $\ell - \ell'$ in the main queue and all agents ahead of him in the A-queue and

[A.7] We simply focus on the likelihood ratio of beliefs of agents entering the A-buffer queue. A symmetric argument holds for agents entering the B buffer-queue.

himself are of α-type (which occurs with probability $(1/\mathbf{M}) \times \mu_\alpha^{\ell-\ell'}$).[A.8] This explains the numerator. The denominator is simply the sum over ℓ of probabilities in the numerator.

Hence,

$$\tilde{\gamma}_\ell^0 = \frac{\mu_B \frac{1}{\mathbf{M}} \left(p_0 \mu_\alpha^\ell + \sum_{\ell'=1}^{\ell} p_{\ell'} \mu_\alpha^{\ell-\ell'} \right)}{\sum_{\ell=1}^{K} \sum_{\ell'=0}^{\ell} p_{\ell'} \mu_B \frac{1}{\mathbf{M}} \mu_\alpha^{\ell-\ell'}} = \frac{\mu_B \frac{1}{\mathbf{M}} p_0 \left(\mu_\alpha^\ell + \sum_{\ell'=1}^{\ell} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'} \right)}{\sum_{\ell=1}^{K} \sum_{\ell'=0}^{\ell} p_{\ell'} \mu_B \frac{1}{\mathbf{M}} \mu_\alpha^{\ell-\ell'}}$$

where the second equality is obtained by Lemma 2 in Leshno (2019) where it is proved that $p_\ell = \mu_B \left(\frac{\mu_\alpha}{\mu_A}\right)^\ell p_0$ for all $\ell = 1, ..., K_A$.

To prove the lemma, we note that for all $\ell = 2, ..., K_A$,

$$\begin{aligned}
r_\ell^0 &= \frac{\tilde{\gamma}_\ell^0}{\tilde{\gamma}_{\ell-1}^0} \\
&= \frac{\mu_\alpha^\ell + \sum_{\ell'=1}^{\ell} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'}}{\mu_\alpha^{\ell-1} + \sum_{\ell'=1}^{\ell-1} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'-1}} \\
&= \frac{\mu_\alpha^\ell + \left(\frac{\mu_\alpha}{\mu_A}\right) \mu_B \mu_\alpha^{\ell-1} + \sum_{\ell'=2}^{\ell} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'}}{\mu_\alpha^{\ell-1} + \sum_{\ell'=1}^{\ell-1} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'-1}} \\
&= \frac{\mu_\alpha^{\ell-1} (\mu_\alpha + \left(\frac{\mu_\alpha}{\mu_A}\right) \mu_B) + \sum_{\ell'=2}^{\ell} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'}}{\mu_\alpha^{\ell-1} + \sum_{\ell'=1}^{\ell-1} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'-1}} \\
&= \frac{\mu_\alpha^{\ell-1} \frac{\mu_\alpha}{\mu_A} (\mu_A + \mu_B) + \sum_{\ell'=2}^{\ell} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'}}{\mu_\alpha^{\ell-1} + \sum_{\ell'=1}^{\ell-1} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'-1}}
\end{aligned}$$

[A.8] Note that these agents have never been offered any item. Otherwise, they would have been matched or they would be in a buffer queue. So, wlog, we can consider that we are drawing their types and positions only at the current period. Further, we are assuming that each agent believes that the position he holds in the main queue is a uniform draw over all possible positions.

$$= \frac{\mu_\alpha}{\mu_A} \frac{\mu_\alpha^{\ell-1} + \sum_{\ell'=2}^{\ell} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'-1} \mu_B \mu_\alpha^{\ell-\ell'}}{\mu_\alpha^{\ell-1} + \sum_{\ell'=1}^{\ell-1} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'-1}}$$

$$= \frac{\mu_\alpha}{\mu_A} \frac{\mu_\alpha^{\ell-1} + \sum_{\ell'=1}^{\ell-1} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'-1}}{\mu_\alpha^{\ell-1} + \sum_{\ell'=1}^{\ell-1} \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell'} \mu_B \mu_\alpha^{\ell-\ell'-1}} = \frac{\mu_\alpha}{\mu_A}.$$

∎

Now, we are in a position to prove the first step of our argument, i.e., that obedience constraints hold at $t = 0$. (Recall that the maximal sizes of the buffer-queues compatible with the obedience constraints are denoted K_A^* and K_B^*, for buffer-queues A and B respectively).

Proposition S8. Assume that the maximal sizes of buffer-queues are given by K_A^* and K_B^*. FCFS with no information satisfies the obedience constraints at $t = 0$.

Proof. In the sequel, we simply focus on the obedience constraint at $t = 0$ for agents recommended to join the A buffer-queue. A symmetric argument holds for the other obedience constraint. The maximal sizes of the buffer-queues satisfying the obedience constraints are identified in Leshno (2019). In case $\mu_\alpha = \mu_A$, it is equal to $\lfloor 2\mu_A \frac{V}{C} \rfloor - 1 \equiv K_A^*$ for the A buffer-queue. In case $\mu_\alpha \neq \mu_A$, K_A^* is equal to $\sup\{K \in \mathbb{Z}_+ \mid K + \frac{\mu_A}{\mu_A - \mu_\alpha} + \frac{K}{\left(\frac{\mu_\alpha}{\mu_A}\right)^K - 1} \leq \frac{V}{C}\mu_A\}$.

First, it is clear that $\tau_\ell^* = \ell/\mu_A$. Indeed, an agent in position ℓ in the A-buffer-queue will have to wait the arrival of ℓ items A to get matched. Since, at each date, the likelihood that an item A arrives is μ_A, the expected waiting time for this agent must be ℓ/μ_A.

Case 1: $\mu_\alpha = \mu_A$. Lemma S8 implies that $\tilde{\gamma}_\ell^0 = \frac{1}{K_A^*}$ for all $\ell = 1, ..., K_A^*$. Hence, we can rewrite the obedience constraint at $t = 0$ as follows

$$\begin{aligned} V - C \sum_{\ell=1}^{K_A^*} \tilde{\gamma}_\ell^0 \tau_\ell^* &= V - C \frac{1}{K_A^*} \sum_{\ell=1}^{K_A^*} \frac{\ell}{\mu_A} \\ &= V - \frac{1}{\mu_A} C \frac{K_A^* + 1}{2} \geq 0. \end{aligned}$$

Note that this inequality holds true if and only if $K_A^* \leq 2\frac{V}{C}\mu_A - 1$. Since K_A^* is an integer, this is equivalent to $K_A^* \leq \lfloor 2\mu_A \frac{V}{C} \rfloor - 1$ which holds by definition of K_A^*.

Case 2: $\mu_\alpha \neq \mu_A$. Lemma S8 implies that $\tilde{\gamma}_\ell^0 = \tilde{\gamma}_1^0 \left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell-1}$ for all $\ell = 1, ..., K_A^*$ and

$\tilde{\gamma}_1^0 = \frac{1}{\sum_{\ell=1}^{K_A^*}\left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell-1}}$. Hence, we can rewrite the obedience constraint at $t = 0$ as follows

$$
\begin{aligned}
V - C\sum_{\ell=1}^{K_A^*}\tilde{\gamma}_\ell^0\tau_\ell^* &= V - C\frac{1}{\mu_A}\frac{\sum_{\ell=1}^{K_A^*}\left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell-1}\ell}{\sum_{\ell=1}^{K_A^*}\left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell-1}} \\
&= V - C\frac{1}{\mu_A}\frac{\sum_{\ell=1}^{K_A^*}\left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell}\ell}{\sum_{\ell=1}^{K_A^*}\left(\frac{\mu_\alpha}{\mu_A}\right)^{\ell}} \\
&= 1 - C\frac{1}{\mu_A}\frac{\mu_A}{\mu_A-\mu_\alpha}\frac{\mu_\alpha}{\mu_A}\frac{1-(K_A^*+1)\left(\frac{\mu_\alpha}{\mu_A}\right)^{K_A^*}+K_A^*\left(\frac{\mu_\alpha}{\mu_A}\right)^{K_A^*+1}}{\left(\frac{\mu_\alpha}{\mu_A}\right)-\left(\frac{\mu_\alpha}{\mu_A}\right)^{K_A^*+1}} \\
&= V - C\frac{1}{\mu_A}[K_A^* + \frac{\mu_A}{\mu_A-\mu_\alpha} + \frac{K_A^*}{\left(\frac{\mu_\alpha}{\mu_A}\right)^{K_A^*}-1}] \geq 0
\end{aligned}
$$

where the third equality uses basic properties of power series. Note that this inequality holds true since $K_A^* = \sup\{K \in \mathbb{Z}_+ \mid K + \frac{\mu_A}{\mu_A-\mu_\alpha} + \frac{K}{\left(\frac{\mu_\alpha}{\mu_A}\right)^K-1} \leq \frac{V}{C}\mu_A\}$ and since, as proved in Leshno (2019), the function $K \mapsto K + \frac{\mu_A}{\mu_A-\mu_\alpha} + \frac{K}{\left(\frac{\mu_\alpha}{\mu_A}\right)^K-1}$ is monotonically increasing in K and goes to infinity when K grows large.

∎

Finally, we need to show that the obedience constraints hold for all $t > 0$. In order to do so, we simply show that the likelihood ratio decreases over time.

Proposition S9. $r_\ell^t \leq r_\ell^0$ for all $t \geq 0$ for all $\ell \in \{2, ..., K_A\}$.

Proof. Note that one can write $\{\tilde{\gamma}_\ell^{t+1}\}_\ell$ as a function of $\{\tilde{\gamma}_\ell^t\}_\ell$ as follows

$$\tilde{\gamma}_\ell^{t+1} = \frac{\tilde{\gamma}_\ell^t\mu_B + \tilde{\gamma}_{\ell+1}^t\mu_A}{\tilde{\gamma}_1^t\mu_B + \sum_{i=2}^K\tilde{\gamma}_i^t}$$

for $\ell = 1, ..., K_A$ where we recall that $\tilde{\gamma}_{K_A+1}^t = 0$. Indeed, the numerator is the probability that the agent's queue position is ℓ after staying in the queue for $t+1$ periods. This event occurs either if (1) the agent has already $\ell - 1$ agents ahead of him in the queue at time t and none of them as well as himself are served at time t; or (2) if there are ℓ agents ahead of him at t and at least one agent ahead of him is served at t. The denominator in turn gives the probability that the agent has not been served by time $t+1$. Hence, given that an agent has not been served when period $t+1$ starts, the above expression gives the conditional belief that his position in the queue is ℓ when period $t+1$ starts.

Thus, we obtain

$$\begin{aligned} r_\ell^{t+1} &= \frac{\tilde{\gamma}_\ell^{t+1}}{\tilde{\gamma}_{\ell-1}^{t+1}} \\ &= \frac{\tilde{\gamma}_\ell^t \mu_B + \tilde{\gamma}_{\ell+1}^t \mu_A}{\tilde{\gamma}_{\ell-1}^t \mu_B + \tilde{\gamma}_\ell^t \mu_A} \\ &= \frac{\mu_B + r_{\ell+1}^t \mu_A}{\frac{1}{r_\ell^t}\mu_B + \mu_A} \end{aligned}$$

for $\ell = 2, ..., K_A$ where we recall that $r_{K_A+1}^t = 0$. Hence, we have a mapping from $r^t \equiv \{r_\ell^t\}_\ell$ to $r^{t+1} \equiv \{r_\ell^{t+1}\}_\ell$. Clearly, this mapping is increasing (in the product order). Thus, if we can show that $r_\ell^1 \le r_\ell^0$ for $\ell = 2, ..., K_A$, then the sequence $\{r_\ell^t\}_t$ will be decreasing for each $\ell = 2, ..., K_A$, and so the proof will be complete.

Fix any $\ell = 2, ..., K_A$, we want to show that $r_\ell^1 \le r_\ell^0$, i.e.,

$$\frac{\mu_B + r_{\ell+1}^0 \mu_A}{\frac{1}{r_\ell^0}\mu_B + \mu_A} \le r_\ell^0$$

which in turn is equivalent to

$$r_{\ell+1}^0 \le r_\ell^0.$$

The above holds true by Lemma S8 (we recall that $r_{K_A+1}^0 = 0$). ∎