Sasaki, Dan

Working Paper

Affirmative Priority Queueing

Quaderni - Working Paper DSE, No. 297

Provided in Cooperation with:
University of Bologna, Department of Economics

Suggested Citation: Sasaki, Dan (1997) : Affirmative Priority Queueing, Quaderni - Working Paper DSE, No. 297, Alma Mater Studiorum - Università di Bologna, Dipartimento di Scienze Economiche (DSE), Bologna, https://doi.org/10.6092/unibo/amsacta/5013

This Version is available at:
https://hdl.handle.net/10419/159140

Affirmative Priority Queueing

Dan Sasaki
Institute of Economics
University of Copenhagen
Studiestræde 6
1455 København K, Denmark
Dan.Sasaki@econ.ku.dk

July 1997 *

Abstract

Consider a first-come first-served queue where agents arrive randomly but their participation in the queue is voluntary and strategic. This paper shows that the introduction of priority-class discrimination (retaining first-come first-served *within* each class) unambiguously improves total welfare even if agents are *a priori* identical, i.e. agents have a fixed outside reservation utility and their unit cost of waiting (per period) is also homogeneous across agents. Furthermore, when agents have heterogeneous outside reservation utitlities, those who have low outside reservation utility should be given high priority in the queue for *total* welfare improvement, not only for equity.

Keywords : first-come first-served, participation, balking.

J.E.L. classification codes : C44, D61, J71.

---

* The author would like to thank the seminar participants in Bologna, June 1997, for comments and discussions. The author is responsible for remaining errors and therefore welcomes further comments.

I. Introduction

LAST-COME FIRST-SERVED queueing discipline, as has been shown by Hassin (1985) and Olson (1992), leads to a collectively efficient outcome when queueing agents have an exogenous reservation utility so that their participation in the queue (either to join or to leave) is voluntary and strategic. The intuition for this optimality result is essentially that, when the oldest agent in the queue makes the decision whether to balk (i.e., leave the queue), he or she creates no externality either to other agents in the queue or to new agents who are possibly arriving in the future, thus his or her individual optimisation coincides with social welfare optimisation. This intuition becomes clearer when contrasted with an ordinary first-come first-served queue, in which a new agent's decision to join or not to join the queue does indeed have externality to future agents.

In spite of this well-known result, last-come first-served queueing discipline is rarely practiced in reality. The reason seems twofold. An immediate skepticism to this inverse seniority discipline is, as has always been pointed out ever since the optimality result was published, the doubt cast against its implementability. Namely, whenever participation (joining and balking) is voluntary, older agents in an inverse seniority queue can always leave the queue temporarily and re-join the queue to restore the status of the newest agent. A number of methods to prevent such "cheating" have been discussed in the literature. Perhaps the most obvious among them is to directly prohibit such leaving and re-joining, which requires either that the administrator (the organiser of the queue) monitor the queue incessantly, or that the administrator be able to identify each individual agent. In fact, both Hassin (1985) and Olson (1992) explicitly discuss this implementability problem, mentioning those methods involving side payments. The idea of using side payments (tolls) to make up for strategic externalities in order to attain allocative efficiency is as old as Naor (1969), preceding the optimality result in favour of last-come first-served. Aside from possible transaction costs associated with such transfer payments, a potential complication stems from the fact that the calculation of optimal payments requires exact knowledge of agents' preferences (utility from the service, from the outside alternative, as well as waiting costs) and their arrival process. Practically speaking, implementation of a last-come first-served queue is far from a trivial task.

Apparently, the issue of implementability is not the only obstacle against last-come

first-served discipline. There seems a fairly strong *moral* objection to such an adverse queueing rule. Indeed, it has been proven that inverse seniority leads to a less egalitarian *ex post* utility distribution as opposed to the ordinary first-come first-served rule, even though the former yields higher *aggregate* efficiency than the latter. Besides, first-come first-served seems by far the most common queueing rule in every culture, which underscores our universal societal attitude that first-come first-served is somewhat "natural." Mathematical proof in support of last-come first-served queueing apparently does not by itself suffice to convince our society and the general public to change their widely accepted convention.

Instead of pursuing yet another advanced mechanism to implement the last-come first-served rule, this paper seeks an alternative direction of discussion. In fact, a relatively simple mechanism can attain an outcome which is unambiguously more efficient than the outcome resulting from an ordinary first-come first-served mechanism. It shall be shown in the next Section that two-class priority queueing (first-come first-served *within* each of the two classes) attains higher efficiency compared with standard non-priority queueing even if agents are *a priori* identical. This is a situation where class discrimination is based upon an attribute of each agent which is in itself immaterial to his or her *economic* characteristics such as waiting costs and reservation utility from outside alternatives.

Obviously, if the discriminant is an unchangeable attribute, such as colour, geographic origin, or family names, which is *independent of each agent's tenure in the queue*, then there is no difficulty in implementing the priority queueing rule, unlike last-come first-served where agents have an incentive to balk and re-join the queue. Another advantage of this simple mechanism is the fact that its effectiveness does not rely upon the administrator's knowledge about the parameters (agents' preferences, arrival processes) of the system. (Note: To model the situation game-theoretically, it needs to be assumed in the paper that these parameters are commonly known, yet it is clear that the mechanism — priority-class discrimination — will stay effective in real practice where information is less than complete.)

The moral issue remains, however. For instance, priority-class discrimination may indeed be considered as illegal unless such discrimination enhances not only total welfare efficiency but also *fairness*. A commonly justified ground for discrimination is to argue that a group of agents have inferior outside alternatives (i.e. reservation utility) compared

with the remainder of the agents, so that the group should receive favourable treatment to compensate for their exogenous "underrepresentation." Once again, it suffices to know only the fact that a group is exogenously underrepresented, not the exact level of their reservation utility. Section IV of this paper shows that such "affirmative action" — allocating high priority to that group of agents whose exogenous reservation utility is low — does in fact improve total welfare, not only distributional fairness.

II. Arbitrary Priority Queue : the Basic Model

There is a queue, at the head of which a service arrives according to a Poisson process. The arrival rate is one per period on average. At the moment when a service arrives, if there is an agent waiting, then the agent consumes the service, receives the utility S and leaves the queue immediately. Assume for simplicity that consumption of the service by an agent is completed instantaneously. If there is no agent waiting when a service arrives, then the service is wasted. Namely, the service is not storable.

To the tail of the queue, *a priori identical* agents arrive according to a Poisson process independent of the arrival of the service. The average arrival rate is q agents per period. Upon arrival, each agent observes the number of other agents currently waiting in the queue, depending upon which the newly arrived agent decides whether to join the queue or not. If the agent decides to join, the cost of waiting in the queue is one per period. Namely, the resulting payoff of an agent who has joined the queue will be the utility of the service S minus the time spent in the queue. On the other hand, the payoff of an agent who decides not to join the queue is a fixed number U, where $S - U = \Delta \geq 1$.

The purpose of this simple model is to compare the equilibrium welfare levels resulting from the following two alternative queueing rules. Welfare is defined as the average expected payoff of a newly arrived agent.

1. The ordinary first-come first-served rule.

2. Two-class priority queueing. Agents are categorised into two *externally distinguishable* groups H and L. Their respective arrival rates are q_H and q_L, where obviously $q_H + q_L = q$. Upon arrival, each agent observes the numbers of H- and L-agents

currently waiting in the queue and decides whether to join the queue. When both H- and L-agents are waiting, priority is given to the former *irrespective of their tenure* in the queue. Within each of the two classes, the ordinary first-come first-served rule is in effect. Both H- and L-agents have identical preferences: unit waiting cost one per period, utility from the service S, and outside reservation utility U.

Theorem : *The welfare resulting from the two-class priority queue is higher than that from the ordinary first-come first-served queue.*

Proof : In the ordinary first-come first-served queue, the expected waiting time for a newly arrived agent, if the agent joins the queue, is $n+1$ when there are n other agents already waiting in the queue. Thus, a new agent should join the queue if $n \leq k-1$ and not join if $n \geq k$, where k is the largest integer not exceeding Δ.

Similarly, in the two-class priority queue, a new H-agent should join the queue if $n_H \leq k-1$ and not join if $n_H \geq k$. This rule does not automatically apply to L-class agents.

Consider a pure strategy profile which is "symmetric" in that every H-agent applies the unique best strategy as aforementioned, and that every L-agent applies the strategy which is to join and stay in the queue if and only if the number of other agents *ahead*, i.e., the number of all other agents precedingly waiting in the queue plus the number of H-agents who has joined later, is $\ell - 1$ or less. Note first that $\ell \leq k$ is necessary for individual optimality. For, if an L-agent already has k agents ahead of him, then his utility from joining/staying in the queue is *at best* $S - k - 1 < U$, thus he should leave or not join the queue.

Also, when $\ell = k$ (whether this is an equilibrium or not), the stationary distribution of the queue length becomes identical to that of the first-come, first-served queue (see VI.i for exact distributions), and thus the resulting welfare efficiency also becomes identical between the two queueing systems.

As long as $\ell \leq k$, the expected payoff of an L-agent waiting in the queue with $\hat{n}$ other agents ahead, denoted by $V(\hat{n}|\ell)$ where $\hat{n} = 0, \cdots, \ell - 1$, has the following transition

equation:

$$V(\hat{n}|\ell) = \frac{1}{1+q_H} V(\hat{n}-1|\ell) + \frac{q_H}{1+q_H} V(\hat{n}+1|\ell) - \frac{1}{1+q_H} .$$

In words, the number of other agents ahead of the L-agent is currently $\hat{n}$, and will change as soon as either

- a service arrives, in which case the number decreases from $\hat{n}$ to $\hat{n}-1$, or
- a new H-agent arrives, in which case the number increases from $\hat{n}$ to $\hat{n}+1$.

The expected waiting time until one of these two events occurs is $\frac{1}{1+q_H}$, and the probabilities that each of these two events precedes the other are $\frac{1}{1+q_H}$ and $\frac{q_H}{1+q_H}$, respectively. The transition equation is solved as

$$V(\hat{n}|\ell) = S - \frac{1}{1-q_H}\left(\hat{n}+1+\frac{({q_H}^{\ell-\hat{n}} - {q_H}^{\ell+1})((1-q_H)\Delta-(\ell+1))}{1-{q_H}^{\ell+1}}\right),$$

where $V(\ell|\ell) = U$ and $V(-1|\ell) = S$ for notational consistency.

The next step of this proof is to look for an ℓ such that *all* L-agents' applying the cut-off rule ℓ is indeed an equilibrium. To be an equilibrium, ℓ must satisfy the participation condition

$$V(\hat{n}|\ell) \geq U \qquad \hat{n} = 0, \cdots, \ell-1$$

and the incentive compatibility condition

$$\frac{1}{1+q_H} V(\ell-1|\ell) + \frac{q_H}{1+q_H} U - \frac{1}{1+q_H} \leq U .$$

In words, the first condition assures that an L-agent has no incentive to *balk* when there are only $\hat{n} < \ell$ agents ahead, and all other L-agents are applying the same cut-off ℓ. The second condition is to assure that an L-agent has no incentive to *join* or *stay* in the queue when there are already ℓ agents ahead, and all other L-agents are applying the same cut-off ℓ.

Whenever the participation condition is satisfied, the relation

$$U = V(\ell|\ell) \leq V(\ell-1|\ell) \leq \cdots \leq V(0|\ell) \leq V(-1|\ell) = S$$

will hold. Conversely, the participation condition holds if $V(\ell-1|\ell) \geq U$, which is further equivalent to

$$(1-q_H)^2\Delta \geq (1-q_H)\ell - q_H(1-{q_H}^{\ell}) .$$

On the other hand, the incentive compatibility condition can be simplified as

$$(1-q_H)^2\Delta \le (1-q_H)(\ell+1) - q_H(1-{q_H}^{\ell+1})\,.$$

Clearly, there is only one ℓ satisfying these two conditions simultaneously (the only exceptional case is when either one of these two can be met with an equality, in which case there are two consecutive integer values of ℓ satisfying these two weak inequalities). Let ℓ^* denote such an ℓ.

If $\ell^* = k$, then the welfare resulting from the two-class priority queue is the same as the welfare in the ordinary first-come first-served queue (as explained before). Otherwise, if $\ell^* < k$, then

$$V(\hat{n}|\ell^*) > V(\hat{n}|k) \qquad \hat{n} = 0, \cdots, k-1\,,$$

so that the expected payoff for an L-agent is higher if the cut-off is ℓ^* than if it is k. Clearly, L-agents' actions have no externality to H-agents. Hence, the two-class priority queue welfare-dominates the first-come first-served queue in equilibrium. ■ Q.E.D. ■

Numerical Intuition : It is intuitively clear that:

- if the outside alternative U is extremely attractive, i.e. $\Delta \approx 1$, then $\ell = k = 1$ and thus the welfare W_1 in the first-come first-served queue and the welfare W_2 in the two-class priority queue become identical.
- if, on the other hand, the outside alternative is effectively non-existent, i.e. $\Delta = \infty$, then $\ell = k = \infty$, which again implies $W_1 = W_2$ as long as $q < 1$.

It is when the outside alternative is moderate, i.e. Δ takes an intermediate value, that agent's participation (either to join or to balk) becomes truly strategic. The following tables illustrate the performance of the two different queueing mechanisms.

Table 1 : $q = 0.5\,,\ q_H = q_L = 0.25$

Δ	1	2	3	6	10	20	≥ 50
$S - W_1$	1	1.428571	1.666667	1.937008	1.994138	1.999990	2
$S - W_2$	1	1.365079	1.588235	1.873283	1.976540	1.999817	2
$S - W_H$	1	1.238095	1.305882	1.332723	1.333330	1.333333	$1.\dot{3}$
$S - W_L$	1	1.492063	2.870588	2.413844	2.619749	2.666300	$2.\dot{6}$
$W_2 - W_1$	0	0.063492	**0.078431**	0.063725	0.017598	0.000173	0

Table 2 : $q = 0.9\,,\ q_H = q_L = 0.45$

Δ	1	2	10	20	50	200	≥ 500
$S - W_1$	1	1.630996	4.918627	7.406237	9.761822	10.00000	10
$S - W_2$	1	1.538180	3.900542	5.627203	8.588696	9.999157	10
$S - W_H$	1	1.394856	1.816309	1.818181	1.818182	1.818182	$1.\dot{8}\dot{1}$
$S - W_L$	1	1.681503	5.984776	9.436225	15.35921	18.18013	$18.\dot{1}\dot{8}$
$W_2 - W_1$	0	0.092816	1.018085	**1.779034**	1.173126	0.000843	0

Table 3 : $q = 0.99\,,\ q_H = q_L = 0.495$

Δ	1	10	100	260	500	2000	≥ 5000
$S - W_1$	1	5.813107	42.59649	79.57490	96.71970	100.0000	100
$S - W_2$	1	4.398337	24.95273	53.06612	78.61563	99.96131	100
$S - W_H$	1	1.974941	1.980198	1.980198	1.980198	1.980198	$1.\dot{9}80\dot{1}$
$S - W_L$	1	6.821733	47.92527	104.1520	155.2511	197.9424	$198.\dot{0}19\dot{8}$
$W_2 - W_1$	0	1.414771	17.64195	**26.50878**	18.10407	0.038690	0

When q is small and therefore the queue is expectedly short (Table 1), it is natural that the relevance of queueing rules is relatively low. As q becomes larger and thus the expected queue length grows longer (Tables 2 and 3), the welfare improvement made possible by two-class discrimination also becomes increasingly significant. In tables, numbers in boldface indicate the maximum welfare gain for each given (q, q_H).

As Δ grows, i.e. as the outside option becomes less attractive, the total welfare ultimately converges to a limit ($\Delta \uparrow \infty$) where both queueing rules (in fact *any* rule)

yield the same efficiency. However, in the two-class priority queue, expected utilities of H- and L-class agents show drastically different *rates of convergence*. W_L converges disproportionately slowly in contrast with W_H and W_1. In real-life queueing situations where $\Delta \ll \infty$, this difference in convergence rates can make the two-class priority rule significantly welfare-dominant over the ordinary first-come first-served rule.

When the queue is expectedly long and thus the two-class discrimination is particularly effective for welfare improvement, a natural question to investigate is how the two classes should be split. The split between q_H and q_L becomes particularly relevant when $q \geq 1$. As $\Delta \to \infty$, the welfare gain from two-class discrimination grows to infinity when $q_H < 1$ (Table 4), whereas the limit welfare gain is finite when $q_H > 1$ (Table 5). Note in particular that choice of $q_H < 1$ can enhance *Pareto* improvement as well as total welfare increase. Slanted numbers in Table 4 indicate those cases where $W_L > W_1$, i.e. even L-class agents *benefit from being discriminated against*. This seemingly counterintuitive phenomenon reflects the fact that, when $q \geq 1$, the total welfare *diverges* to $-\infty$ as Δ tends to infinity.

Table 4 : $q = 2$, $q_H = 0.9$, $q_L = 1.1$

Δ	1	2	8	9	100	1000	∞	Order
$S - W_1$	1	1.857143	7.516634	8.509286	99.50000	999.5000	∞	$\approx \Delta$
$S - W_2$	1	1.766298	6.055775	6.718763	58.82461	553.5001	∞	$\approx 0.55\Delta$
$S - W_H$	1	1.630996	4.237758	4.587118	9.997583	10.00000	10	
$S - W_L$	1	1.876999	7.543243	*8.462836*	*98.77399*	*998.1819*	∞	$\approx \Delta$
$W_2 - W_1$	0	0.090845	1.460859	1.790523	40.67539	445.9999	∞	$\approx 0.45\Delta$

Table 5 : $q = 2$, $q_H = 1.1$, $q_L = 0.9$

Δ	1	2	10	100	≥ 200
$S - W_1$	1	1.857143	9.505129	99.50000	$\Delta - 0.5$
$S - W_2$	1	1.788520	8.132577	95.00357	$\Delta - 5$
$S - W_H$	1	1.697885	6.795979	90.91575	$\Delta - 9.\dot{0}\dot{9}$
$S - W_L$	1	1.899295	9.766196	99.99980	Δ
$W_2 - W_1$	0	0.068623	1.372553	4.496427	4.5

To be complete, explicit algebraic calculation of equilibrium expected welfare resulting from the two alternative queueing rules is listed in VI.ii.

III. Tenure Independent Rationing : a Limit Result

It has been proven in Section II that, even when agents are *a priori* homogeneous, arbitrarily imposed two-class discrimination will yield an unambiguous welfare improvement. It is straightforward to prove that a similar result can be extended to more general multi-class priority queueing. In particular, the limit case is where there are infinitely many priority classes, so that the priority ranking among waiting agents is solely defined by their classes *regardless of how long each agent has been waiting*.

More concretely, compare the equilibrium welfare levels between the following two alternative queueing rules.

1. The ordinary first-come first-served rule.
2. Multi-class priority queueing. Each agent is assigned with a predetermined "class" indexed by a real number, which is *externally observable*. Whenever a service arrives, priority is given to the agent whose class is the highest. If there is more than one agent with an identical class, then the ordinary first-come first-served rule is effective among them.

In either case, all agents have identical preferences: unit waiting cost one per period, utility from the service S, and outside reservation utility U when not joining the queue. Note that these two rules become identical when the class distribution in the latter is degenerate. The multi-class rule is reduced into a two-class rule if the class distribution is binary. It is when the class distribution is *atomless* that the multi-class priority rule becomes purely tenure independent.

The following can be viewed as a generalisation of the previous Theorem.

Corollary : *The welfare in the multi-class priority queue is higher than that in the ordinary first-come first-served queue.*

Supplementary Proof : Compare between the two-class queue as in Section II, and a three-class queue where agents of the top, middle and bottom classes arrive at rates q_T , q_M and q_L , respectively, where $q_T + q_M = q_H$. In other words, the H-class in the two-class queue is subdivided into T- and M-classes in the three-class queue.

In either queueing rule, the presence of the bottom class (population q_L) has no externality toward higher-class agents. Therefore, by directly applying the Theorem, the equilibrium welfare of higher-class agents is higher in the three-class queue (the population-weighted sum of T- and M-class agents' utility) than in the two-class queue (the utility of H-class agents).

Note also that, in Section II, the queue length distribution in the two-class queue was shorter than in the single-class first-come first-served queue in first-order stochastic dominance (see VI.i). Thus, the combined queue length of top two classes in the three-class queue is shorter than the queue length of H-class in the two-class queue. This directly implies that the equilibrium welfare of the bottom class is also higher in the three-class queue than in the two-class queue.

The same logic can be recursively applied to any arbitrary multi-class priority queue.

■ Q.E.D. ■

IV. Affirmative Priority Queue : an Extension

Despite its superior welfare performance, it is highly presumable that imposition of priority-class discrimination would encounter many objections, especially those concerning fairness and equity. For a large part, these objections stem from the arbitrarity of discrimination.

Fortunately, there is an additional aspect of reality that has not yet been exploited in the simple analysis presented in the previous two sections. Namely, agents may have heterogeneous outside alternatives. If agents can be categorised into two (or more) *externally recognisable* groups, and if there is any reason to believe that different groups of agents have different reservation utilities, then it is arguable that the queueing rule should favour those whose outside alternatives are less attractive, in order for fairness.

The purpose of this section is to examine the welfare performance of such "affirmative" discrimination.

Revisit the basic model presented in Section II *except* that now H- and L-class agents are assumed to have different exogenous reservation utilities U_H and U_L, respectively, where $S - U_H = \Delta_H \geq 1$ and $S - U_L = \Delta_L \geq 1$. Theorem in Section 2 implies the following.

Corollary : *When $U_H < U_L$, the welfare resulting from the two-class priority queue is strictly higher than that from the ordinary first-come first-served queue.*

Supplementary Proof : When the queueing rule does not discriminate between H and L types, an H-agent joins the queue if and only if $n \leq k_H - 1$ and an L-agent joins if and only if $n \leq k_L - 1$, where n is the number of other agents already in the queue, and $k_H = \text{int}(\Delta_H)$, $k_L = \text{int}(\Delta_L)$.

When the queueing rule favours H-agents, an H-agent joins the queue if and only if $n_H \leq k_H - 1$, where n_H is the number of other H-agents already in the queue. Now, suppose that the decision rule for an L-agent is to join and to stay in the queue if and only if $\ell - 1$ agents are ahead in the queue.

If $\ell = k_L$, the stationary distribution of the *total* number of agents in the queue is identical between the two queueing rules. However, in the two-class priority queue, if an H-agent arrives when there are already k_L agents in the queue, and if these k_L agents include at least one L-agent, then the newest L-agent has to leave the queue and is replaced by the new H-agent. When $\Delta_L \geq 1$, this event takes place with a strictly positive probability.

Namely, in the priority queue, *more L-agents and less H-agents end up taking outside alternatives* compared with the ordinary first-come first-served queue. This clearly improves the total welfare when $U_H < U_L$.

Analogously to the Theorem, L-class agents can only be *better* off than this if their optimal decision rule is $\ell \neq k_L$ instead of $\ell = k_L$. Hence both when $\ell = k_L$ and when

$\ell \neq k_L$, the two-class priority rule *strictly* welfare-dominates the first-come first-served rule. ■ Q.E.D. ■

Economic Implication : Once again, the result can directly be extended to more general multi-class priority rules. In the limit, the welfare implication offered by the two Corollaries is such that the priority relation among waiting agents should be determined entirely by their exogenous reservation utilities (the lower the reservation is, the higher the priority should be) *irrespective of their tenure in the system.*

V. Concluding Remark

In a queueing situation where participation is voluntary, priority-class discrimination can increase total (sometimes even Pareto) welfare as opposed to an ordinary first-come first-served queue. Moreover, this priority-class rule is easy to implement in the following ways. First, unlike the famous last-come first-served rule, agents have no incentive to balk and rejoin the queue to gain extra priority. Second, the rule is simple and easy to understand even for those agents who are not theoretical experts.

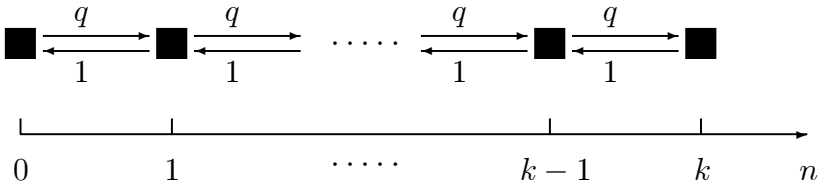
Finally, when agents have heterogeneous reservation utility, the "affirmative" priority-class discrimination which favours those "disadvantaged" agents whose outside alternatives are relatively less attractive will improve *total* welfare, not only redistribution of utility across different types of agents. Conveniently, installation of such a discriminatory queueing rule is not ethically objectionable, since it enhances both efficiency and equity at the same time.

VI. Mathematical Notes

VI. i. Queue Length Distributions

The stationary distribution of the first-come first-served queue length is illustrated in Figure 1.

Figure 1 : Transition of queue length.



In the diagram, black squares indicate feasible states and arrows indicate conditional transition frequencies. The condition for stationarity can be summarised as

$$\text{Prob}(n) = q \cdot \text{Prob}(n-1) \qquad\qquad n = 1, \cdots, k$$

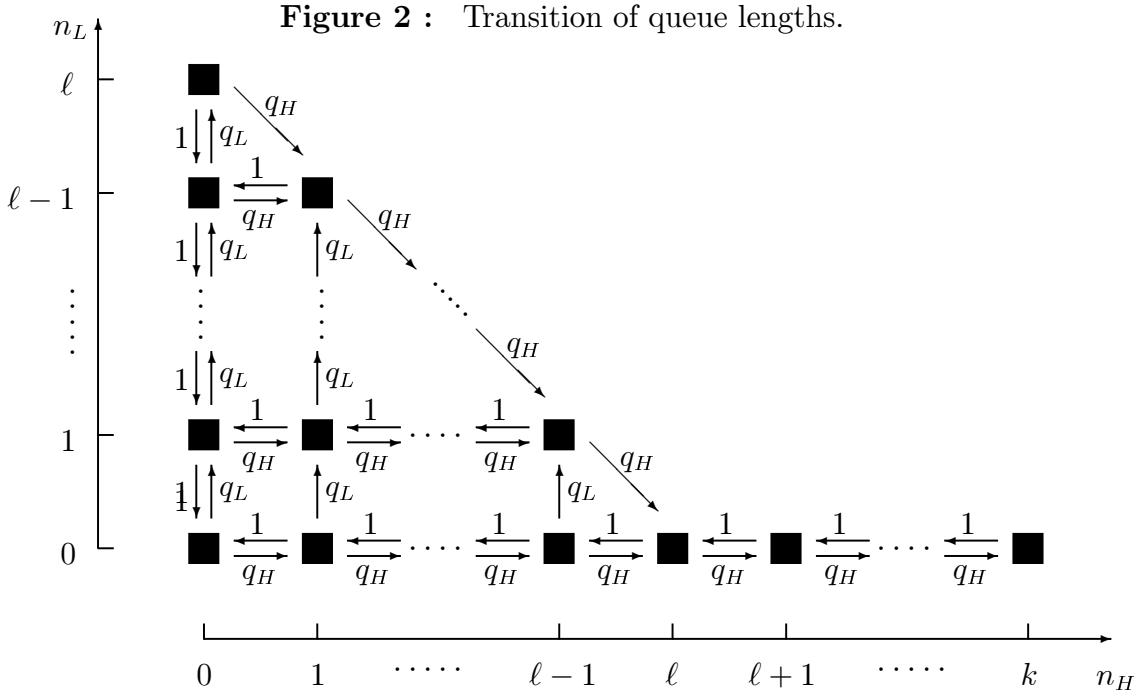
where $\text{Prob}(n)$ is the stationary probability of n agents waiting in the queue. The stationary distribution of the queue length is therefore

$$\text{Prob}(n) = \frac{1-q}{1-q^{k+1}}\, q^n \qquad\qquad n = 0, \cdots, k\,.$$

Without further notice, this Section presents explicit computation for $q \neq 1$ only. The analysis in the case of $q = 1$ involves no theoretical or qualitative departure from the generic case $q \neq 1$.

Likewise, the stationary distribution of the two-class priority queue length is illustrated in Figure 2.

Figure 2 : Transition of queue lengths.



From conditional transition frequencies, the transition of the total queue length $n = n_H + n_L$ can be identified as

$$\Pr(n) = \begin{cases} q \cdot \Pr(n-1) & n = 1, \cdots, \ell\,, \\ q_H \cdot \Pr(n-1) & n = \ell+1, \cdots, k\,. \end{cases}$$

At the same time, the number of H-agents n_H follows the transition equation

$$\Pr_H(n_H) = q \cdot \Pr_H(n_H - 1) \qquad\qquad n_H = 1, \cdots, k$$

and thus is distributed according to

$$\Pr_H(n_H) = \frac{1-q_H}{1-{q_H}^{k+1}}\,{q_H}^{n_H} \qquad\qquad n_H = 0, \cdots, k\,.$$

Note also that

$$\Pr(n) = \Pr_H(n) \qquad\qquad n = \ell+1, \cdots, k\,.$$

The stationary distribution of the queue length is therefore

$$\Pr(n) = \begin{cases} \dfrac{1-{q_H}^{\ell+1}}{1-{q_H}^{k+1}} \cdot \dfrac{1-q}{1-q^{\ell+1}}\,q^n & n = 0, \cdots, \ell\,, \\ \dfrac{1-q_H}{1-{q_H}^{k+1}}\,{q_H}^n & n = \ell+1, \cdots, k\,. \end{cases}$$

Obviously, when $\ell = k$, this distribution coincides with the distribution of the first-come first-served queue length. Since both H- and L-agents have homogeneous reservation utility and waiting costs, identical queue length distributions automatically lead to identical total welfare between the two queueing rules.

Otherwise, if $\ell < k$, this two-class priority queue is shorter than the first-come first-served queue in first-order stochastic dominance.

VI. ii. Equilibrium Welfare

The equilibrium welfare under the first-come, first-served rule is

$$\begin{aligned} W_1 &= \sum_{n=0}^{k-1} \mathrm{Prob}(n) \cdot (S - (n+1)) + \mathrm{Prob}(k) \cdot U \\ &= S - \frac{1}{1-q^{k+1}} \left(\frac{1-q^k}{1-q} + q^k((1-q)\Delta - k) \right), \end{aligned}$$

where $k = \mathrm{int}(\Delta)$.

Analogously, the equilibrium welfare for the H class in the two-class priority queue is computed as

$$\begin{aligned} W_H &= \sum_{n_H=0}^{k-1} \Pr_H(n_H) \cdot (S - (n_H + 1)) + \Pr_H(k) \cdot U \\ &= S - \frac{1}{1 - {q_H}^{k+1}} \left(\frac{1 - {q_H}^k}{1 - q_H} + {q_H}^k((1 - q_H)\Delta - k) \right). \end{aligned}$$

On the other hand, the equilibrium expected payoff of a newly arriving L-agent is calculated as

$$\begin{aligned} W_L &= \sum_{n=0}^{\ell^*-1} \Pr(n) \cdot V(n) + \sum_{n=\ell^*}^{k} \Pr(n) \cdot U \\ &= S - \frac{1}{1 - {q_H}^{k+1}} \left(\frac{1 - {q_H}^{\ell^*+1}}{(1-q)(1-q_H)} - {q_H}^{k+1}\Delta + (q^{\ell^*+1} - {q_H}^{\ell^*+1}) \frac{(1-q)\Delta - (\ell^* + 1)}{(1 - q^{\ell^*+1})(q - q_H)} \right) \end{aligned}$$

(for the definition of ℓ^*, see Section II). Hence, the equilibrium welfare under the two-class priority queueing rule (that is, the population-weighted average between an H-agent's expected payoff and that of an L-agent) becomes

$$\begin{aligned} W_2 &= \frac{q_H W_H + q_L W_L}{q} \\ &= S - \frac{1}{(1 - q_H)q} + \frac{1}{(1 - {q_H}^{k+1})q} \left[1 - (1 - {q_H}^{\ell^*+1}) \left(\frac{1}{1-q} - \frac{1}{1-q_H} \right) \right. \\ &\qquad \left. - {q_H}^{k+1} \Big((1-q)\Delta - k \Big) - (q^{\ell^*+1} - {q_H}^{\ell^*+1}) \frac{(1-q)\Delta - (\ell^* + 1)}{1 - q^{\ell^*+1}} \right]. \end{aligned}$$

Note that $W_2 = W_1$ when $\ell^* = k$.

References

DeGroot, M. H., *Optimal Statistical Decisions*, McGraw-Hill (1970).

Hassin, R., "On the Optimality of First Come Last Served Queues," *Econometrica*, Vol. 53 (1985), pp. 201 - 02.

Jaiswal, N. K., *Priority Queues*, Academic Press (1968).

Kleinrock, L., *Queueing Systems*, John Wiley (1975).

Naor, P., "The Regulation of Queue Size by Levying Tolls," *Econometrica*, Vol. 37 (1969), pp. 15 - 24.

Olson, M., "Queues When Balking Is Strategic," University of Arizona, Discussion Paper 92-11 (1992).