Many service providers utilize priority queues. Many consumers revile priority queues. However, some form of priority service may be necessary to maximize social welfare. Consequently, it is useful to understand how the priority scheme chosen by a revenue-maximizing firm differs from the one a social planner would use. We examine this in a single server-queue with customers that draw their valuation from a continuous distribution and have a per-period waiting cost that is proportional to their realized valuation. The decision maker must post a menu offering a finite number of waiting time-price pairs. There are then three dimensions on which a revenue maximizer and social planner can differ: coverage (i.e., how many customer in total to serve), coarseness (i.e., how many classes of service to offer), and classification (i.e., how to map customers to priority levels).

We show that differences between the decision makers priority policies are all about classification. Both are content to offer very coarse schemes with just two priority levels, and they will have negligible differences in coverage. However, differences in classification are persistent. Further, a revenue maximizer may — relative to the social planner — have too few or too many high priority customers. Whether the revenue maximizer over- or under-stuffs the high priority class depends on a measure of consumer surplus that is captured by the mean residual life function of the valuation distribution. In addition we show that there is a large class of valuation distributions for which a move from first-in, first-out service to a priority scheme that places those with higher waiting costs at the front of the line reduces consumer surplus.

1. Introduction

In 2013 Walibi, a Belgian amusement park, introduced Speedy-Pass, a premium service that allowed purchasers to jump to the front of the line at park rides. These shorter waits did not come cheap; the service more than doubled the price of an adult ticket to the park. Perhaps unsurprisingly the announcement was met with an outpouring of opprobrium. An educator asked, “How in the name of God do you explain to a child that he has to wait in line in a long queue, while other children can go straight to the front, just because their parents have got more money?” Government ministers also chimed in denouncing the program [Flandersnews.be 2013].

Angst over priority queues is not limited to Europe. In the United States, travelers have petitioned the Transportation Security Administration not to allow airlines to profit from selling priority
access to airport security screening\(^1\) while the implementation of tolled express lanes – so called Lexus Lanes – has faced opposition in Georgia and Colorado (Walker Oct 10, 2012, Whaley Apr 21, 2015).

On the one hand, such outrage is puzzling. Many firms offer multiple products or service formats and consumers presumably benefit from the increased range of options. A customer who buys a Chevrolet instead of a Cadillac must believe that the former offers better value. On the other hand, one must acknowledge that capacity constrained service providers differ from firms selling physical goods. Queuing creates externalities between customers in different classes of service so that increased sales to one class can reduce the value obtained by customers of the other classes.

If General Motors sells more Cadillacs, Chevy buyers are unharmed, but the more Speedy-Passes Walibi sells, the worse service regular customers receive.

This, however, does not mean that priority schemes necessarily compromise social welfare. A social planner would use priorities if customers have different waiting costs. Consequently, the question is not whether a revenue-maximizing firm such as Walibi should use a priority scheme but whether Walibi’s implementation differs dramatically from what the social planner would do.

Such a comparison of the revenue maximizer’s and the social planner’s actions is the subject of this paper. We consider a service system modeled as an \(M/M/1\) queue. The decision maker may or may not offer multiple priority classes. Arriving customers all have the same average service time but differ in their valuations and waiting costs: they independently draw valuations for the service from a common, continuous distribution and have a per-unit-time waiting cost proportional to their valuation. The state of the queue is not observable to the customers, so they must choose which class of service to purchase based on a posted menu of expected delays and prices.

The decision maker, whether seeking to maximize revenue or social welfare, must make three decisions.

- **Coverage.** If the arrival rate of customers exceeds available capacity, the decision maker must turn away some customers. The revenue maximizer might choose to serve, in total, more or fewer customers than the social planner.

- **Coarseness.** We do not assume that the market is exogenously and \textit{a priori} divided into different classes. It is up to the decision maker to route arrivals into distinct priority levels. With valuations and waiting costs drawn from a continuous distribution, both types of decision makers would benefit from posting a continuous menu of prices and delays. In reality it is more common for service

providers to use coarse priority schemes that split arrivals into a finite number of discrete priority levels. A revenue maximizer might opt for a coarser or for a more refined division than the social planner.

- **Classification.** Given a level of coarseness, the decision maker must still determine how to classify customers into priority levels: how many customers should go into each class and who these customers are. Even if the revenue maximizer chooses the same coverage and coarseness as the social planner, social efficiency might suffer. The revenue maximizer might benefit from pursuing an *ultra-luxury strategy* where a smaller high priority class than is socially optimal is charged higher prices. In other scenarios, the revenue maximizer might benefit from pursuing a *mass-luxury strategy* with a high priority that is larger than socially optimal.

Our analysis shows that the loss of societal efficiency resulting from the revenue maximizers actions is largely a question of classification. We employ a limiting regime akin to Nazerzadeh and Randhawa (2015) in which the arrival rate of customers and the processing rate of the server are scaled up together. We show that both the social planner and revenue maximizer are content to use extremely coarse priority schemes; for either type of decision maker, the loss from using just two classes is negligible. Further, the level of coverage that both offer is essentially the same. Thus, revenue maximization is socially optimal as far as coverage and coarseness are concerned.

Differences, however, remain in classification. The revenue maximizer may opt for a mass-luxury strategy and admit more customers to the high priority class than is socially or for an ultra-luxury strategy and admit fewer customers to the high priority class. Which approach she chooses depends on how consumer surplus per admitted customer changes with the level of coverage. We show that this is related to the mean residual life (MRL) of the valuation distribution and that a convex MRL leads to a mass-luxury outcome while a concave MRL leads to an ultra-luxury outcome.

Intuition for why the revenue maximizer might have the high priority class over- or under-subscribed is gained from considering the problem of maximizing consumer surplus. Maximizing either revenue or social welfare calls for putting those with high valuations (and thus high waiting costs) at the front of the line. This is not necessarily the case with consumer welfare. If the valuation distribution has a decreasing failure rate, those with high valuations should be served first. However, with an increasing failure rate distribution, serving those with low valuations first maximizes consumer welfare. Note that this implies that consumers whose valuations follow an increasing failure rate distribution are better off under first-in, first-out service than under a priority scheme that favors those with high valuations. Further, a decreasing, convex MRL function implies an increasing failure rate, suggesting that the social planner is less aggressive than the revenue maximizer in routing customers to the high priority class in order to reduce the impact on consumer welfare.
2. Literature Review

The comparison of revenue maximization and social optimization in queues has been a topic of interest since Naor’s seminal paper (Naor 1969). Hassin and Haviv (2003) and Hassin (2016) provide excellent surveys. In this brief literature review, we focus on research that allows for a continuum, rather than an exogenously given number, of customer types.

Much of the work in this area builds on Kleinrock (1967), which considers customers who arrive to an unobservable queue and bid for priority. The customer bids are drawn from a common continuous distribution. The service provider offers, in turn, a continuum of priority levels and a customer is given priority over any customer who has bid less.

While the distribution of bids in Kleinrock (1967) is exogenously determined, subsequent work ties the distribution of bids to an equilibrium outcome between customers. Aféche and Mendelson (2004) consider customers who draw their valuation for service from a common distribution and who have a delay cost with two components: one that is proportional to the realized valuation and one that is independent of the valuation. They show that with a single first-in-first-out (FIFO) queue—that is, with uniform pricing—a revenue maximizer may offer greater or smaller coverage than socially optimal depending on the valuation distribution and the delay-cost structure. Studying priority auctions, they establish conditions for the revenue maximizer to achieve social efficiency. Katta and Sethuraman (2005) provide conditions under which the priority auctions in Aféche and Mendelson (2004) are in fact revenue maximizing.

Specializing the results of Aféche and Mendelson (2004) to our setting, we would have that the revenue maximizer admits fewer people than socially optimal under a uniform price but admits customers at the socially optimal rate under a preemptive auction. Under a non-preemptive auction the revenue maximizer does not achieve social efficiency.

In Aféche and Mendelson (2004) as well as most other work in this vein, types are never pooled, so small differences in waiting costs may result in absolute differences in priorities. Depending on the waiting cost structure, the revenue maximizer may pool some types together, impose a common price and offer the same expected wait (Katta and Sethuraman 2005). Additionally, Aféche and Pavlin (2016) show that a revenue maximizer facing customers with a utility structure different than ours may use a complex service discipline that may pool customers or exclude some with intermediate valuations or impose strategic delay. Note that in these papers, the service provider still offers a continuum of priority levels despite pooling some types: customers who are not pooled are still given distinct priority levels despite small differences in waiting costs.

A limited set of papers consider how to map a continuum of customers to a coarse set of priority levels. Ghanem (1975) considers customers that differ only on their per-period waiting costs drawn
from a common distribution, and examines how they should be classified into a predetermined number of priority classes in order to minimize total delay costs. [Gilland and Warsing (2009)] consider a similar problem from the perspective of revenue maximization but restrict their analysis to uniformly distributed delay costs and assume that all customers must be served.

In [Gavirneni and Kulkarni (2014)], customers differ only in their waiting costs which follow a Burr Type XII distribution and the service provider only offers two levels of priority. They present examples in which the revenue maximizer classifies more customers as high priority than the social planner would.

[Doroudi et al. (2015)] consider the same valuation and cost structure as we do; arriving customers draw a valuation from a common distribution and have waiting costs that are proportional to their realized valuations. The bulk of their analysis focuses on offering a continuum of priorities but they do demonstrate numerically that a coarse priority scheme with a limited number of priority classes performs very well.

Finally, [Nazerzadeh and Randhawa (2015)] examine how a revenue-maximizing service provider should manage coverage, coarseness, and classification when customer draw valuations from a common distribution and have per-period waiting costs that are a function of their realized valuations. They use an asymptotic analysis to show that a very coarse priority scheme is sufficient; two levels of priority capture nearly all of the possible system value. In examining coarseness, we take a similar approach. Indeed, their results partly provide the revenue maximizer side of our comparison. We not only expand the analysis to the social planner but also strengthen it. To demonstrate the near optimality of two classes for the revenue maximizer (as done in [Nazerzadeh and Randhawa (2015)]), it suffices to identify one solution (among possibly many solutions). In order to compare the decisions (i.e., the menus) of the revenue maximizer and social planner, however, we must assure the uniqueness of the asymptotically optimal prescriptions.

3. Model Formulation

We consider a service modeled as a queuing system. There is a (potential) arrival stream that is Poisson with rate $\Lambda$. The queue is served by a single server, and the service time is exponential with rate $\mu$ and independent across customers. Without loss of generality, we fix $\mu$ to 1. We further assume that $\Lambda \geq 1$ so that not all customers can be served. How much of the market the decision maker chooses to cover is then a non-trivial question.

Customer valuation for the service is drawn from a distribution $F$ with support $(a, b)$ with $0 \leq a < b$. Valuations are independent across customers. Customers are also adverse to delay. A customer’s delay cost is linear in her waiting time (which includes the delay in the queue and the service time)
with a coefficient that is proportional to her valuation: a customer with service valuation \( v \) incurs a cost of \( \alpha v \) per unit of delay where \( \alpha < 1 \). This specification provides heterogeneity in both the cost coefficient and the valuation.

\( F \) is assumed to be differentiable with differentiable density \( f \). Let \( \bar{F}(v) = 1 - F(v) \), let \( h(v) = f(v)/\bar{F}(v) \) be the distribution’s failure rate and \( g(v) = vh(v) \) be the generalized failure rate of \( F(v) \). We make the following assumption on \( F \) and \( g \).

**Assumption 1**: \( F \) has a finite mean.

**Assumption 2**: \( F \) has an increasing generalized failure rate, i.e., \( g'(v) \geq 0 \) on \((a,b)\).

The IGFR assumption is satisfied by many common distributions. It implies that \( f \) is strictly positive on \((a,b)\) so both \( F \) and \( \bar{F} \) are invertible.

Let \( MRL \) be the mean residual life of \( F \), i.e.,

\[
MRL(v) = E[X - v | X \geq v] = \int_v^\infty \frac{tf(t)dt}{\bar{F}(v)} - v.
\]

In the reliability literature the mean residual life represents the expected remaining life of a component given that it has survived to a given age. Here, \( MRL(v) \) represents the expected value beyond a base level \( v \) created by serving a customer conditional on that customer having a value of at least \( v \). Given that \( F \) has a finite mean, \( MRL(v) \) is defined for all \( v \) in \((a,b)\).

**Customer actions.** We assume that the service provider offers \( K \in \mathbb{Z}_+ \) different bundles of price and waiting time, \((p_i, W_i)\). A customer must consequently choose whether to join the queue and, if she does, at which grade of service. A customer with valuation \( v \) that chooses menu item \( i \) obtains the utility

\[
U(v; i) = v - \alpha v W_i - p_i.
\]

The service provider must pair higher prices with shorter waiting times as no customer would buy a bundle that charges a higher price and imposes a longer delay than another menu item. We must, in particular, have \( W_i \neq W_j \) and \( p_i \neq p_j \) for \( i \neq j \).

Customers who select a particular menu item have valuations that fall within an interval:

**Lemma 1.** Suppose that a customer with valuation \( \tilde{v} \) optimally chooses menu item \((p_i, W_i)\) and another customer with valuation \( \hat{v} \) optimally chooses menu item \((p_j, W_j)\). If \( W_i > W_j \) and \( p_i < p_j \), then \( \tilde{v} \leq \hat{v} \). Therefore if two customers with valuations \( v < u \) choose the same menu item \((p_k, W_k)\), a customer with valuations \( w \in (v, u) \) must choose the same menu item.

Thus, a price-delay menu segments the valuation space into intervals and we may, without loss of generality, number the offerings such that a higher index corresponds to a higher price and a
shorter wait. Customers with the highest valuations thus choose item $K$. Since the set of customer valuations for a given menu item is an interval and $F$ is strictly increasing, there exists a unique cutoff valuation $v_i$ such that $U(v_i; i) = U(v_i; i - 1)$, i.e., a customer with the valuation $v_i$ is indifferent between menu items $i$ and $i - 1$. Let $v$ be the vector of cut off values. The fraction of all customers choosing menu item $i \in \{1, \ldots, K\}$ is given by $\bar{F}(v_i) - \bar{F}(v_{i+1})$ (we define $v_{K+1} \equiv \infty$) and the number of such customers is

$$\lambda_i(v) = \Lambda(\bar{F}(v_i) - \bar{F}(v_{i+1})).$$

We write $\lambda(v)$ (dropping the subscript) for the vector of arrival rates. Let

$$\bar{\lambda}_i(v) = \sum_{j=i}^{K} \lambda_j(v) = \Lambda F(v_i)$$

be the number of customers that choose menu items $i$ or higher. With this notation $\bar{\lambda}_1(v) = \Lambda F(v_1)$ is the service provider’s coverage: the volume of customers who enter the system per time unit. For stability this volume must be strictly smaller than the service rate—$\bar{\lambda}_1(v) < 1$—so that no customer with valuation less than

$$\bar{v} = \bar{F}^{-1}\left(\frac{1}{\Lambda}\right),$$

enters the service (i.e., $v_1 > \bar{v}$).

The menu prices are uniquely determined by the cutoffs and the waiting times. Indeed, since the customer with valuation $v_i$ is indifferent between $i$ and $i - 1$, we have

$$U(v_i; i - 1) = v_i - p_{i-1} - \alpha v_i W_{i-1} = v_i - p_i - \alpha v_i W_i = U(v_i; i),$$

so that

$$p_i - p_{i-1} = \alpha v_i (W_{i-1} - W_i). \quad (2)$$

The lowest-valuation customer who patronizes the service is indifferent between joining and not joining. Assuming that all customers have an outside option of zero, we then have

$$v_1 - W_1 \alpha v_1 - p_1 = 0 \implies p_1 = v_1 (1 - \alpha W_1). \quad (3)$$

$p_1$ can be interpreted as an entrance fee to the system – anyone that enters has to pay at least $p_1$. A customer with valuation $v > v_1$ enters the system and obtains a strictly positive utility since

$$U(v; i) \geq v(1 - \alpha W_1) - p_1 > v_1 (1 - \alpha W_1) - p_1 = U(v_1; 1) = 0.$$ 

Thus, given cutoff values $v_1 < v_2 < \ldots < v_K$ and waiting times $W_1 > W_2 > \cdots > W_K$, the prices menu is uniquely determined by $[2]$ and $[3]$ regardless of whether the decision maker seeks to maximize revenue or social welfare. The objective functions of the two providers will determine how the vectors $v$ and $W$ are set.
The social planner’s problem. For a fixed number $K$ of classes, the social planner must choose a vector $v$ of cut-off values and a vector $W$ of waiting times to maximize social welfare, i.e., the aggregate utility of arriving customers. The expected utility of a class-$i$ customer (a customer that chooses menu item $i$) with valuation $v$ is $v - \alpha v W_i$. The social welfare following from a given $v$ and $W$ can be written as

$$S_K(v, W) = \Lambda \sum_{i=1}^{K} \left( \int_{v_i}^{v_{i+1}} u f(u) du \right) (1 - \alpha W_i)$$

$$= \Lambda V(v_1) - \alpha \sum_{i=1}^{K} \lambda_i(v) \left( \frac{V(v_i) - V(v_{i+1})}{F(v_i) - F(v_{i+1})} \right) W_i$$

where $V(x) := \int_x^\infty u f(u) du$ so that $(V(v_i) - V(v_{i+1}))/\left( \hat{F}(v_i) - \hat{F}(v_{i+1}) \right)$ is the average contribution to social welfare ignoring delay costs of a customer conditional on her selecting class $i$. Similarly, $c^S_i(v) := \alpha \frac{V(v_i) - V(v_{i+1})}{F(v_i) - F(v_{i+1})}$ is the average cost of delay among class-$i$ customers. In maximizing $S_K(v, W)$, the constraints on the social planner’s actions are that the cut-off values $v$ are increasing while the waiting-time vector $W$ must be decreasing and feasible given the induced arrival rates. We allow the use of preemptive priority schemes; see § 6 for the case of non-preemptive policies. Let $\mathcal{W}(v)$ be the set of feasible waiting times given $v$ and the resulting arrival rates. $\mathcal{W}(v)$ is determined by the achievable region; see e.g. Theorem A.10 of [Stidham 2009].

The social planner’s problem can then be written as

$$S^*_K = \max_{v \uparrow, W \in \mathcal{W}} S_K(v, W) \quad \text{s.t. } W \in \mathcal{W}(v)$$

Because $F$ is strictly increasing, $V(v_i) - V(v_{i+1}) > 0$ for all $i$ so that work conservation is optimal (rather than, say, inserting strategic delay). Given $v$, the cost coefficients $c^S_i(v)$ are increasing in $i$ so that it is optimal to preemptively prioritize customers in decreasing order of $c^S_i$ and the social planner’s problem reduces to a search over cutoff values. We thus have (recall that $\lambda_i(v) = \Lambda(\hat{F}(v_i) - \hat{F}(v_{i+1}))$)

$$S^*_K = \max_{v \uparrow} S_K(v) := \Lambda V(v_1) - \alpha \Lambda \sum_{i=1}^{K} (V(v_i) - V(v_{i+1})) W_i(\lambda(v))$$

where, for each $i = 1, \ldots, K$

$$W_i(\lambda(v)) = \frac{1}{(1 - \lambda_{i+1}(v))(1 - \lambda_i(v))} = \frac{1}{(1 - \Lambda \hat{F}(v_i))(1 - \Lambda \hat{F}(v_{i+1}))}$$

is the preemptive static priority waiting time of class $i$ with arrival vector $\lambda(v)$ and service rate equal to one.
The revenue maximizer’s problem. As with the social planner, the revenue maximizer sets cutoff values \( v \) and expected waits \( W \). The relationship (2) and (3) map prices and waits to cutoffs and waits. The firm’s revenue is then given by

\[
R_K(v, W) = \sum_{i=1}^{K} \lambda_i(v) p_i = \Lambda p_1 F(v_1) + \Lambda \sum_{i=1}^{K} (p_{i+1}(v) - p_i(v)) F(v_{i+1}) \\
= \Lambda v_1 F(v_1) (1 - \alpha W_1) + \alpha \Lambda \sum_{i=1}^{K} (W_i - W_{i+1}) v_{i+1} F(v_{i+1}) \\
= \Lambda \rho(v_1) - \alpha \sum_{i=1}^{K} \lambda_i(v) \frac{\rho(v_i) - \rho(v_{i+1})}{F(v_i) - F(v_{i+1})} W_i \\
= \Lambda \rho(v_1) - \sum_{i=1}^{K} \lambda_i(v) c^R_i(v) W_i,
\]

where \( \rho(v) := v F(v) \) and \( \rho(v_{K+1}) = 0 \). The coefficient \( c^R_i(v) = \alpha \frac{\rho(v_i) - \rho(v_{i+1})}{F(v_i) - F(v_{i+1})} \), captures the discount given to customers of class \( i \) to compensate them for their delay. Using \( \lambda_i(v) = \Lambda F(v_i) \), these can also be written as \( c^R_i(v) = \alpha \frac{\rho(\lambda_i(v)) - \rho(\lambda_{i+1}(v))}{\lambda_i(v) - \lambda_{i+1}(v)} \), where \( \rho(\lambda) := \rho(\bar{F}^{-1}(\lambda/\Lambda)) \) for \( \lambda \geq 0 \). If \( \rho \) is decreasing and \( \bar{\rho} \) is convex then \( c^R_i(v) \) is positive and increasing in \( i \). Then much like the social planner, work conservation is optimal and the revenue maximizer will want to preemptively prioritize customers in decreasing order of \( c^R_i \). In that case, we can state the revenue maximizer’s problem as choosing cutoff values knowing that customers will have expected waits given by \( W_i(\lambda(v)) \).

\[
R_K^* = \max_{v_1} R_K(v) := \Lambda \rho(v_1) - \alpha \Lambda \sum_{i=1}^{K} (\rho(v_i) - \rho(v_{i+1})) W_i(\lambda(v)),
\]

which should be contrasted with (4). The relationship between \( V(v_i) - V(v_{i+1}) \) and \( \rho(v_i) - \rho(v_{i+1}) \) will play a key role in our results.

Notice that if customers were not delay sensitive (i.e., \( \alpha = 0 \)) the revenue maximizer’s problem reduces to choosing the cutoff that maximizes \( \Lambda v F(v) = \Lambda \rho(v) \) or equivalently an arrival rate that maximizes \( \lambda p(\lambda) \) where \( p(\lambda) = \bar{F}^{-1}(\lambda/\Lambda) \). If \( \rho \) is a decreasing function, the revenue can be read as the delay-insensitive revenue minus a delay-discount of \( \alpha (\rho(v_i) - \rho(v_{i+1})) \) for class-\( i \) customers.

We conclude this section with a lemma that, among other things, allows us to conclude that, indeed, \( c^R_i(v) > 0 \) and increasing in \( i \) within a certain range of valuations.

**Lemma 2.** Given Assumptions 1 and 2:

1. \( v^*_0 := \inf \{ v : g(v) \geq 1 \} \) is finite.
2. \( \rho(v) \) is maximized at \( v^*_0 \). It is increasing and concave for \( v < v^*_0 \) and decreasing for \( v \geq v^*_0 \).
3. \( \bar{\rho}(\lambda) \) is maximized at \( \lambda_0^* := \Lambda F(v^*_0) \). For \( \lambda < \lambda_0^* \), \( \bar{\rho}(\lambda) \) is increasing and concave.
4. $\varepsilon(\lambda) := -p(\lambda)/(\lambda p'(\lambda))$ is increasing in $\lambda$.

In what follows we will assume that the system is capacity constrained in the sense that

$$v_0^* < \bar{v},$$

where recall that $\bar{v} = F^{-1}(\frac{1}{K})$. In words, the revenue maximizer can not admit as many customers as she would want to if customer were not delay sensitive. This restricts us to the range of $\lambda$ such that $\tilde{\rho}(\lambda)$ is increasing and concave in $\lambda$ and the range of $v$ such that $\rho(v)$ is decreasing in $v$.

4. **The tension in social planning**

There is, of course, a link – or perhaps more accurately, a gap – between the objective of the social planner and that of the revenue maximizer. The social planner’s objective is the sum of the firm’s revenue and the consumers’ surplus, i.e.,

$$S_K(v, W) = R_K(v, W) + C_K(v, W)$$

where $C_K(v, W)$ is the consumer surplus with $K$ classes. It is written as the surplus from admitting all customers with valuations greater than $v_1$ less the waiting costs they incur:

$$C_K(v, W) = \Lambda CS(v_1) - \sum_{i=1}^{K} \alpha_i (CS(v_i) - CS(v_{i+1})) W_i$$

$$= \Lambda CS(v_1) - \alpha \sum_{i=1}^{K} \lambda_i(v) c_i^{CS}(v) W_i,$$

where $CS(v) = V(v) - \rho(v)$ for $v \geq 0$ and $c_i^{CS}(v) = \frac{CS(v_i) - CS(v_{i+1})}{\Lambda(F(v_i) - F(v_{i+1}))}$ and for a vector $v$.

The tension between $R_K$ and $C_K$ is important to the results of this paper and, as it turns out, the nature of this tension depends to a great extent on the failure rate of the distribution.

In Figure 1 we plot, for two classes, both the revenue and consumer surplus for different distributions and service disciplines. The total height of a bar corresponds to the social welfare under a particular priority scheme while the shading indicates how much is captured by the firm versus by customers. We compare FIFO service with a priority scheme in which those with higher valuations – and hence higher waiting costs – are served first.

In the case of a decreasing failure rate (DFR) distribution, the interests of the service provider and its customers are aligned; both are better off when priorities are used and the increase in social welfare comes from making both parties better off. The case of increasing failure rate (IFR), however, presents a different story. While social welfare and revenue again increase under priorities,
consumer surplus now falls. Customers are hurt if priorities are implemented, and social welfare only increases because the service provider’s revenue rises by more than consumer welfare drops.

This is a generalizable phenomenon. There is a large class of valuation distributions for which consumer surplus falls if those with higher valuations are given priority. It is straightforward that \( CS(v_i) - CS(v_{i+1}) > 0 \) so that work conservation is optimal. As with our analysis of the social planner and the revenue maximizer, if one wants to maximize consumer surplus it is be optimal to preemptively prioritize customers in decreasing orders of \( c_{SV_i}^i(v) \). However, the values \( c_{SV_i}^i(v) \) may not be ordered as one expects.

**Lemma 3.** \( c_{SV_i}^i(v) \leq c_{SV_{i+1}}^i(v) \) if the failure rate \( h(\cdot) \) is decreasing and \( c_{SV_i}^i(v) \geq c_{SV_{i+1}}^i(v) \) if it is increasing. With constant failure rate, \( c_{SV_i}^i(v) = c_{SV_{i+1}}^i(v) \).

When the valuation distribution has a decreasing failure rate, the priority scheme aligns with that used by the social planner or the revenue maximizer: Customers with high valuations (and thus high waiting costs) are placed at the front of the line while those with low valuations (and thus low waiting costs) are relegated to the rear. However, the situation is reversed when valuations are governed by an increasing failure rate (IFR) distribution. Now it is optimal to let those with low waiting costs enjoy shorter waits. This explains the phenomenon illustrated in Figure 1.

2 There are IGFR distributions such as the log-normal that have non-monotone failure rates. However, one cannot easily characterize how a non-monotone failure rates impacts the ordering of \( c_{SV_i}^i(v) \). It may be that all change points of \( h(v) \) fall below \( \hat{v} \) so the \( h(v) \) is monotone over the range. Otherwise, we cannot rule out that the value of \( c_{SV_i}^i(v) \) may not be monotone.
Weibull distribution is DFR for shape parameters less than one but IFR shape parameters greater than one. Hence, a move from FIFO to prioritizing those with high valuations when \( k > 1 \) reduces consumer welfare.

Prioritizing customers with low valuations may not be implementable if valuations are private information. The prices and waiting times must satisfy (2) and (3) and a menu where customers with low valuations wait less cannot be incentive compatible. Yet, the decision maker could still maximize consumer surplus by using a service discipline such as FIFO that is independent of customer valuations.

This observation speaks to our earlier examples of customers protesting the implementation of priority schemes. If valuations follow an IFR distribution, an implementable priority mechanism lowers consumer surplus even though social welfare increases. However, this assertion implicitly assumes that moving from FIFO to priorities does not increase coverage. Once coverage is in play, consumer surplus is potentially pulled in two ways. Priorities may harm customers but expanded coverage would benefit them.

**Implications for classification:** With IFR valuation distributions the social planner faces a tension as it seeks to maximize the sum of firm revenue and consumer surplus: Revenue maximization dictates prioritizing customers with higher valuations while consumer surplus is compromised by such priority. We will see that the social planner will alleviate some of this tension by having a smaller high priority class than the revenue maximizer; see Remark 1 further below.

### 5. Coverage, Coarseness and Classification

If the providers are restricted to use a single menu item – an admission fee and a delay to go with it, the revenue maximizer’s problem reduces to

\[
R^*_1 = \max_{v > \bar{v}} \Lambda \rho(v) (1 - \alpha W_1(\lambda(v)))
\]  

(7)

where \( W_1(\lambda(v)) = (1 - \Lambda F(v))^{-1} \) is the delay in the \( M/M/1 \) case. The social planner’s problem is

\[
S^*_1 = \max_{v > \bar{v}} \Lambda V(v)(1 - \alpha W_1(\lambda(v))),
\]  

(8)

The function \( (1 - W_1(\lambda(v))) \) is concave increasing and \( V(v) \) is concave and decreasing in \( v \). The social planner’s objective function is thus strictly concave and has a unique maximizer \( v^*_S \). By Lemma 2, \( \rho(v) \) is concave and decreasing for all \( v > \bar{v} \) so that the revenue has a unique maximizer \( v^*_R \).

These are special instances of the problems in Afêche and Mendelson (2004). It follows from their Proposition 1 and our Lemma 2 that with an IGFR distribution \( v^*_S \leq v^*_R \) so that \( \lambda_1(v^*_S) \geq \lambda_1(v^*_R) \).
The revenue maximizer offers a smaller coverage than the social planner. If $F$ has a constant generalized failure rate (i.e., it is a Pareto distribution), then $\lambda_1(v^*_R) = \lambda_1(v^*_S)$.

Contrast this with a setting in which the providers can tailor a different price and delay to each valuation $v$. The welfare from a customer with valuation $v$ is $v - \alpha v W_c(v)$. The higher the customer’s valuation, the higher her priority so that a customer with valuation $v$ has an expected waiting time $W_c(v) = (1 - \Lambda \bar{F}(v))^{-2}$ [Kleinrock 1967, Theorem 2] which is a continuous segmentation analogue of (5). The social planner solves the problem

$$S^*_\infty = \max_{v_S \geq \bar{v}} \int_{v_S}^{\infty} (v - \alpha v W_c(v)) f(v) dv.$$  \hspace{1cm} (9)

The revenue maximizer chooses similarly an admission valuation $v^*_R$ such that a customer enters if and only if her valuation is $v \geq v^*_R$. The price is determined so that the first customer to enter is indifferent between entering or not, i.e., $p(v^*_R) = v^*_R - \alpha v^*_R W_c(v^*_R)$. All other entering customers pay the entering price plus a premium for shorter delays, i.e.,

$$p(v) = p(v^*_R) - \alpha \int_{v^*_R}^{v} u W'_c(u) du.$$  

Since $W'_c(v) \leq 0$ the premium is positive. The revenue maximizer’s problem reduces then to

$$R^*_\infty = \max_{v_R \geq \bar{v}} \left( -p(v_R) \bar{F}(v_R) + \int_{v_R}^{\infty} p'(v) \bar{F}(v) dv \right).$$  \hspace{1cm} (10)

Problems (9) and (10) are instances of the priority auctions described in [Afeche and Mendelson 2004]. By Proposition 3 there, $v^*_R = v^*_S$ (i.e., the two providers choose the same entry cutoff) provided that $\epsilon(\lambda)$ (recall Lemma 2) is increasing in $\lambda$.

In summary, with a single class, for strictly IGFR valuation distributions, the revenue and social optimizers choose different coverage levels and the social planner serves more customers. With a continuum of classes, however, they make identical decisions. We turn to study the intermediate and practical case in which there is a finite number of classes.

**High-volume analysis.** When dealing with multiple (but finite number of) classes, the social and revenue maximizer problems are rather intractable. Fortunately, when the volume is high we can characterize the decisions of the providers and compare them. We build on the approach of [Nazerzadeh and Randhawa 2015] and study the asymptotic performance, as the arrival rate and service rate are scaled up by a multiplier $n$: the nominal arrival rate is $\Lambda n$ and the service rate is $n$. We superscript all relevant notation with $n$ to capture the dependence on this multiplier. Thus, for example, $v^{n*}_{i,R} \ [v^{n*}_{i,S}]$ is the optimal class-$i$ cutoff of the revenue maximizer [social planner]. Since both the nominal arrival rate and the service rate are multiplied by $n$, $\bar{v}$ does not depend on $n$. 

For large values of \( n \), it turns out, the differences between subsequent optimal cutoff values, \( v_{i+1}^n - v_i^n \) becomes small. Roughly speaking we can then replace \( V(v_{i+1}^n) - V(v_i^n) \) in the objective function (4) of the social planner with a Taylor expansion around \( v_i^n \). It is such Taylor expansions that enable the analysis of an otherwise intractable problem. Our analysis allows us to nail down the optimal arrival-rate vectors (and revenue/welfare outcomes) up to an error that is negligible relative to the square root of the multiplier \( n \). Accordingly, two coverage/classification decisions are distinguishable if they are \( \sqrt{n} \) apart.

We first re-visit the single-class FIFO queue. From Afche and Mendelson (2004) it follows that, in our setting, the revenue maximizer admits fewer customers than socially optimal. We can now characterize more fully the difference in coverage.

Following standard notation, we say that \( \xi(n) = o(\sqrt{n}) \) if \( \xi(n)/\sqrt{n} \to 0 \) as \( n \to \infty \).

**Lemma 4 (coverage difference with FIFO).** With a single (FIFO) class the cut-off entry valuations of the social planner and the revenue maximizer satisfy

\[
v_{1,S}^n = \bar{v} + \sqrt{\alpha} \sqrt{\frac{F(\bar{v}) V(\bar{v})}{f(\bar{v}) V'(\bar{v})}} n^{-\frac{1}{2}} + o(n^{-\frac{1}{2}}), \quad v_{1,R}^n = \bar{v} + \sqrt{\alpha} \sqrt{\frac{F(\bar{v}) \rho(\bar{v})}{f(\bar{v}) |\rho'(\bar{v})|}} n^{-\frac{1}{2}} + o(n^{-\frac{1}{2}}).
\]

Optimal admission rates consequently satisfy,

\[
\lambda_{1,S}^n = n - \lambda f(\bar{v}) \sqrt{\alpha} \sqrt{\frac{F(\bar{v}) V(\bar{v})}{f(\bar{v}) |V'(\bar{v})|}} \sqrt{n} + o(\sqrt{n}), \quad \lambda_{1,R}^n = n - \lambda f(\bar{v}) \sqrt{\alpha} \sqrt{\frac{F(\bar{v}) \rho(\bar{v})}{f(\bar{v}) |\rho'(\bar{v})|}} \sqrt{n} + o(\sqrt{n})
\]

If \( F \) is IGFR then \( V(\bar{v})/V'(\bar{v}) > \rho(\bar{v})/\rho'(\bar{v}) \) and, consequently, the social planner has a larger coverage. Furthermore, the difference in coverage is non-negligible, \( \lambda_{1,S}^n - \lambda_{1,R}^n \neq o(\sqrt{n}) \) if the inequality is strict.

We next show that with multiple customer classes the coverage gap disappears. Classification, however, differs depending on the mean residual life of the distribution. Theorem 1 below is the main result of this paper and focuses on the comparison of the actions of the social planner and the revenue maximizer. The full characterization of their decisions appears in Theorems EC.1 and EC.2 in the appendix.

**Theorem 1. (coverage, coarseness and classification: SP Vs. RM)** With \( K > 1 \) levels of service, the coverage of the social planner and the revenue maximizer are asymptotically identical in the sense

\[
\lambda_{1,R}^n - \lambda_{1,S}^n = o(\sqrt{n}). \quad \text{(Coverage)}
\]

For both, two classes are sufficient:

\[
S_K^* = S_2^* + o(\sqrt{n}), \quad \text{and} \quad R_K^* = R_2^* + o(\sqrt{n}). \quad \text{(Coarseness)}
\]
Classification is asymptotically different except for linear MRL:
\[
\bar{\lambda}_{n}^{*2,R} - \bar{\lambda}_{n}^{*2,S} = \gamma n^{3/4} + o(n^{3/4}),
\]
(Classification)

where \(\gamma \geq 0\) (resp. \(\gamma \leq 0\)) if the valuation distribution has a convex (resp. concave) MRL and \(\gamma = 0\) if the MRL is linear. In particular, the revenue maximizer directs more volume to the high priority when \(F\) has a convex MRL. Further, if \(\gamma \neq 0\) (classification is different), the social cost of revenue maximization grows as \(\sqrt{n}\),
\[
\liminf_{n \to \infty} \frac{S_{n}^{*2,R} - S_{n}^{*2,S}(v_{n}^{*2,R})}{\sqrt{n}} > 0.
\]
(Social welfare gap)

We find then that both providers choose coarse priority schemes and offer identical coverage (up to a small difference) but that their admission to the high priority class differs. Table 1 is a numerical illustration of this result. The example serves to illustrate three additional points: (i) that the asymptotics-based comparison holds also for small values of \(n\), (ii) that the difference in classification can be rather large: for Weibull(1,0.3) which has a concave MRL \(\gamma = -140 < 0\). It can also be rather small: for the Weibull(1,2) which has convex MRL, \(\gamma = 0.011\), and (iii) that both the social planner and the revenue maximizer have a high priority class that is much larger than the low priority class. This latter observation is supported in Theorems EC.1 and EC.2 in the appendix.

<table>
<thead>
<tr>
<th>(n)</th>
<th>Coverage</th>
<th>High Priority</th>
<th>(n)</th>
<th>Coverage</th>
<th>High Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.573</td>
<td>8.568</td>
<td>10</td>
<td>8.44</td>
<td>7.9</td>
</tr>
<tr>
<td>100</td>
<td>95.495</td>
<td>95.485</td>
<td>100</td>
<td>95.177</td>
<td>92.365</td>
</tr>
<tr>
<td>1000</td>
<td>985.789</td>
<td>985.767</td>
<td>1000</td>
<td>985.144</td>
<td>975.720</td>
</tr>
</tbody>
</table>

Table 1 \(\Lambda = 30\), \(\alpha = 0.2\): (LHS) Weibull(1,2) for convex MRL (RHS) Weibull(1,0.3) for concave MRL.

The differences in classification matter. They suffice to make the revenue maximizer’s actions socially inefficient. Figure 2 (LHS) displays \((S_{n}^{*2,R} - S_{n}^{*2,S}(v_{n}^{*2,R})/\sqrt{n}\) as a function of \(n\) for a Weibull(1,0.3) distribution (concave MRL). Thus, while a continuum of classes means that maximizing revenue is equivalent to maximizing welfare, for each finite \(K\) revenue maximization leads to non-negligible social inefficiency.

Further, any number of classes larger than two classes has little impact on either revenue or social welfare. This was proved already for the revenue maximizer by Nazerzadeh and Randhawa (2015) and we extend this result to the social planner. Figure 2 (RHS) provides numerical evidence for this result.
Remark 1 (The role of Mean Residual Life). The mean residual life is often studied in the reliability literature. It is argued that it can be more informative than the failure rate; the latter captures only local information – i.e., the chance of imminent failure – while the MRL reflects the entire distribution of the component’s remaining life (Lai and Xie (2006)). A similar argument applies here. Elasticity conditions as in Afeche and Mendelson (2004) are local but, as our analysis reveals, one must capture the entire distribution of value in making classification decisions.

The MRL is also intimately linked with consumer surplus and, in turn, to the observations made in §4. Specifically, \( MRL(v) = CS(v) / \bar{F}(v) \) so that \( MRL(\bar{v}) / \Lambda \) is the consumer surplus per admitted customer. In §4 we proved that maximizing welfare requires putting those with high valuations at the front of the line when the valuation distribution is DFR. An IFR valuation distribution reverses the order giving those with low valuations high priority. Standard results (Lai and Xie (2006)) give that the MRL function of an IFR [DFR] distribution is decreasing [increasing]. Going the other way requires a bit more structure: if \( MRL(v) \) is decreasing and convex, the underlying distribution is IFR. If \( MRL(v) \) is increasing and concave, \( F \) must be DFR.

Now suppose that the MRL is increasing and concave (implying a DFR valuation distribution). From the social planner’s perspective, giving high value customers high priority boosts both firm revenue and consumer welfare. She then has an incentive to classify a large number of customers as high priority. The revenue maximizer, of course, does not see any benefit to increasing consumer surplus and therefore is more conservative than the social planner in expanding the higher priority class.

Things are reversed when \( MRL(v) \) is decreasing and convex (so \( F \) is IFR). Now, the social planner faces a trade off: Maximizing revenue calls for putting those with high valuations at the
front of the line but maximizing social welfare requires putting how with low valuations first. In aggregate, the former effect dominates but limiting the size of the higher class minimizes the impact on low valuation consumers. The revenue maximizer has no such compunctions and thus opts for a larger high priority class.

**Remark 2 (Who pays for social inefficiency).** When the social planner and the revenue maximizer’s classification decisions diverge, consumer surplus is smaller than socially optimal. While the average customer stands to lose from revenue maximization, some customers may gain. In fact, with a concave MRL, the revenue maximizer has a smaller high priority class and, consequently, offers a shorter delay to both classes and a higher customer utility (per class); see Table 2 for a numerical example. The total consumer surplus is higher under social planning. Who gains depends on the convexity/concavity of the MRL.

<table>
<thead>
<tr>
<th></th>
<th>((p_i, W_i))</th>
<th>Avg. Utility</th>
<th></th>
<th>((p_i, W_i))</th>
<th>Avg. Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SP - Low</strong></td>
<td>(53.396, 1.163)</td>
<td>16.046</td>
<td><strong>RM - Low</strong></td>
<td>(63.071, 0.737)</td>
<td>23.499</td>
</tr>
<tr>
<td><strong>SP - High</strong></td>
<td>(76.921, 0.181)</td>
<td>198.197</td>
<td><strong>RM - High</strong></td>
<td>(79.804, 0.155)</td>
<td>236.390</td>
</tr>
</tbody>
</table>

*Table 2* Waiting and pricing menus for \(F = \text{Weibull}(1,0.3)\) (concave MRL), \(\alpha = 0.2\), and nominal arrival rate \(A_n = 300\): (LHS) Social planner (RHS) Revenue Maximizer

Figure 3 provides a schematic view of how the utility \(U(v; i)\) changes with the customer valuation \(v\). With a convex MRL (LHS), the revenue maximizer has a smaller high priority cutoff (represented by a square) than that of the social planner (represented by a circle) and, in turn, a larger high priority class. The story is reversed with a concave MRL (RHS).
More in detail: let \((P^L_S, W^L_S)\) and \((P^H_S, W^H_S)\) be the two menu items offered by the social planner (where \(L\) stands for low priority, i.e., \(W^L_S \geq W^H_S\)). Define similar notation for the revenue maximizer (with the subscript \(R\)). Since the providers offer identical coverage in both cases, 
\[ v_1 - p^L_R - \alpha v_1 W^R_L = v_1 - p^L_S - \alpha v_1 W^L_S = 0 \] 
so that \((p^L_S - p^L_R) = \alpha v_1 (W^L_S - W^L_R)\).

For a convex MRL, we have \(W^L_R \geq W^L_S\) so that the revenue maximizer charges a lower price than the social planner. However, that price reduction is insufficient to make any other low priority customer whole; any customer with valuation greater than \(v_1\) who would be classified as low priority by either decision maker is worse off under the revenue maximizer. One gets similar results at the other extreme. High priority customers also get a price break from the revenue maximizer but for those with very high valuations it does not adequately compensate for increased waits. Thus, with a convex MRL, those with very high or very low valuations are certain to lose when we move from the social planner to the revenue maximizer.

With a concave MRL, the story is reversed. Customers with valuation greater than than \((p^H_R - p^H_S)/(\alpha (W^H_S - W^H_R))\) or smaller than \((p^H_S - p^L_R)/(\alpha (W^H_S - W^L_R))\) are better off under revenue maximization. These are the customers that absorb all the gain in the class utilities seen in Table 2. The burden of social inefficiency is all carried by the “middle class”: the customers with intermediate valuations. With convex MRL, in contrast, the high and low valuation customers pay for the welfare loss under revenue maximization.

6. Non-Preemptive policies

Thus far we allowed the provider to use preemptive policies and, indeed, in their optimal solutions both providers use preemptive static priorities. In practice it is often unrealistic to preempt customers in the middle of their service. If a provider is restricted to non-preemptive policies, the menus offered to customers must change. We prove, however, that a restriction to non-preemption does not change our results: while there are differences between preemptive and non-preemptive, these are relatively negligible.

The objective functions \(S_K(v, W)\) and \(R_K(v, W)\) are the same as in §3. The difference is that, in optimizing, \(W\) must be taken from the non-preemptive achievable region \(W_{NP}(v)\). Let \(S^*_{2,NP}\) be the optimal social welfare (optimized over all valuation cutoffs and non-preemptive disciplines). Since the providers are now constrained now to use non-preemptive policies it holds that \(S^*_{2,NP} \leq S^*_{2}\) and \(R^*_{2,NP} \leq R^*_{2}\). Our last theorem shows that the loss is minimal.

Let \(S^*_{2,NP}(v^n)\) be the social planner’s objective function value when two-class are used, the cutoff vector \(v^n\) is used for classification and static non-preemptive priority is used with the highest
priority provided to the highest valuations customers. Similarly define \( R_{2,NP}^n(v^n) \) for the revenue maximizer.

**Theorem 2.** (Optimality of preemptive cut-offs with non-preemptive service) Using two non-preemptive priority classes with the optimal preemptive cut-off valuations is nearly optimal in the \( \sqrt{n} \) scale for both the social planner and the revenue maximizer. That is,

\[
S_{2,NP}^n(v^n) - S_{2} = o(\sqrt{n}) \quad \text{and} \quad R_{2,NP}^n(v^n) - R_{2} = o(\sqrt{n}).
\]

respectively.

The non-preemptive provider can use the same coarseness, coverage and classification as the preemptive provider with negligible compromise to optimality. In turn, the comparisons that apply to the preemptive case, apply to the non-preemptive restriction.

7. Conclusion

Managing a service system as a priority queue is a challenging endeavor. To map a continuum of customers into a finite number of priority classes requires multiple decisions. One must determine how much of the market to cover, how coarse a set of priorities to offer, and how to classify customers into specific grades of service. Given such complexity, it is remarkable that decision makers pursuing distinct goals – maximizing social welfare and maximizing revenue – can agree on two out of three of these dimensions. We show in a limiting regime that a social planner and a revenue maximizer choose essentially identical levels of coverage while being content to offer very coarse priority schemes.

Meaningful differences exist, however, in how customers are classified. The revenue maximizer may pursue an ultra-luxury strategy and admit too few customers to the high priority class (in comparison to the social optimal) or a mass-luxury strategy and admit too many to the high priority class. These differences in classification are driven by differences in a measure of consumer surplus specifically, consumer surplus per admitted customer, which is captured by the mean residual life of the customer valuation distribution. A concave MRL results in the revenue maximizer following an ultra-luxury strategy while a convex MRL results in the revenue maximizer following a mass-luxury strategy.

We also address the more general question of how priority schemes affect consumer surplus. We show that if the valuation distribution has a decreasing failure rate, consumer surplus is maximized by the most natural priority scheme, the one that puts those with high costs at the front of the line (which coincidently is the scheme that both the social planner and revenue maximizer would use).
However, when the valuation distribution has an increasing failure rate, that natural priority scheme results in a lower consumer surplus than simple first-in, first-out service. Since a decreasing, convex MRL implies an increasing failure rate, this suggests that a convex MRL should cause the social planner to limit deviations from FIFO waits to protect consumer surplus; the revenue maximizer has no such concerns.

There are several ways this work can be extended. One could consider alternative utility structures (see, for example, Af` eche and Pavlin (2016)) or alternative information structure (for example, visible queues). Introducing competition could also be fruitful. Competition generally results in consumers capturing more of the value the system creates. Here value can be shifted to customers through lower prices or more efficient classification. Which is the better means for rewarding customers is an interesting question.

References


Proofs

EC.1. Proofs of Lemmas

Proof of Lemma 1: The customer with valuation $\tilde{v}$ prefers to join class $i$ over class $j$, meaning that
\[(1 - \alpha W_i) \tilde{v} - p_i \geq (1 - \alpha W_j) \tilde{v} - p_j.\]
The customer with valuation $\hat{v}$ prefers to join class $j$ over class $i$, meaning that
\[(1 - \alpha W_j) \hat{v} - p_j \geq (1 - \alpha W_i) \hat{v} - p_i.\]
Adding up these two inequalities gives
\[(\tilde{v} - \hat{v}) \alpha (W_j - W_i) \geq 0.\]
Since $W_i > W_j$, we have $\tilde{v} \leq \hat{v}$ concluding that the customers who choose the shorter waiting time $W_j$ have higher valuation.

Proof of Lemma 2: Part 1 of the lemma follows from Theorem 2 of Lariviere (2006). Parts 2 and 3 follow from Proposition 5.1 in Ziya et al. (2004). Part 4 follows noting that $\varepsilon(\lambda) = \frac{\lambda \hat{F}^{-1}(\lambda/\Lambda)}{\chi} f\left(\hat{F}^{-1}(\lambda/\Lambda)\right)$ and substituting $\hat{F}(v(\lambda))$ for $\lambda/\Lambda$.

Proof of Lemma 3: Recall that $\bar{\lambda}_i(v) = \Lambda \hat{F}(v_i)$. In what follows, we fix $v$ and suppress the dependence of $\bar{\lambda}_i$ on $v$ in the notation. Let $\tilde{CS}(x) = CS\left(\hat{F}^{-1}(x/\Lambda)\right)$ and notice that
\[\tilde{CS}'(x) = \frac{1}{\Lambda h\left(\hat{F}^{-1}(x/\Lambda)\right)}.\] (EC.1)
$\tilde{CS}(x)$ is then increasing and concave in its argument [convex] if $h(v)$ is decreasing [increasing]. If $h(v)$ is constant, $\tilde{CS}(x)$ is linear. In turn, since $\bar{\lambda}_i$ is decreasing in $i$, $\tilde{CS}(\bar{\lambda}_i)$ is decreasing and convex [concave] if $F$ is DFR [IFR]. The result then follows from noting that $c_i^{CS}$ can be written as
\[c_i^{CS} = \frac{\tilde{CS}(\bar{\lambda}_i) - \tilde{CS}(\bar{\lambda}_{i+1})}{\bar{\lambda}_i - \bar{\lambda}_{i+1}}.\]
By the definition of convexity/concavity $c_i^{CS}$ is then increasing in $i$ if $F$ is DFR and decreasing if $F$ is IFR. It is constant if the hazard rate is constant.

Proof of Lemma 4: The expressions for $\bar{\lambda}_{n,S}^{i*}$ and $\bar{\lambda}_{n,R}^{i*}$ follow from Theorems EC.1 and EC.2 stated further below.
Here we only need to prove the comparison. The social has a strictly larger coverage (for all sufficiently large n) if \( V(\bar{v})/|V'(\bar{v})| < \rho(\bar{v})/|\rho'(\bar{v})| \). Note that, since \( F \) is IGFR, \( \rho'(\bar{v}) < 0 \), we also have \( V'(\bar{v}) = -\bar{v}f(\bar{v}) < 0 \). Hence, it remains to show that

\[
V(\bar{v})/V'(\bar{v}) > \rho(\bar{v})/\rho'(\bar{v}). \tag{EC.2}
\]

Because we assume that \( F \) has a strictly positive density there exits, for any valuation \( v \), a unique \( \lambda \) such that \( \lambda = \Lambda F(v) \) or \( v(\lambda) = F^{-1}(\lambda/\Lambda) \). Let \( \hat{V}(\lambda) = V(v(\lambda)) \) and \( \hat{\rho}(\lambda) = \rho(v(\lambda)) \). Then, \( \hat{V}'(\lambda) = V'(v(\lambda))v'(\lambda) \). Since \( v'(\lambda) \leq 0 \), [EC.2] will be established if we show that

\[
\frac{\hat{V}'(\lambda)}{\hat{V}(\lambda)} > \frac{\hat{\rho}'(\lambda)}{\hat{\rho}(\lambda)}. \tag{EC.3}
\]

Notice, to that end, that

\[
\frac{\hat{V}'(\lambda)}{\hat{V}(\lambda)} = \Lambda \frac{\hat{F}^{-1}(\frac{\lambda}{\Lambda}) f(\hat{F}^{-1}(\frac{\lambda}{\Lambda}))}{\lambda} = GFR(v(\lambda)).
\]

Since \( F \) is IGFR and \( v'(\lambda) \leq 0 \) we have that \( \frac{\hat{V}'(\lambda)}{\hat{V}(\lambda)} \) is decreasing in \( \lambda \) and the reciprocal \( \frac{\hat{V}'(\lambda)}{\hat{V}(\lambda)} \) is increasing in \( \lambda \). Notice that \( \hat{V}'(\lambda) + \hat{V}''(\lambda) = \hat{\rho}'(\lambda) \). Thus, we have that \( 1 + \frac{\hat{V}'(\lambda)}{\hat{V}(\lambda)} = \frac{\hat{\rho}'(\lambda)}{\hat{V}(\lambda)} \) decreasing in \( \lambda \) so that, fixing \( \lambda \), \( \frac{\hat{\rho}'(\lambda)}{\hat{V}(\lambda)} > \frac{\hat{\rho}'(\lambda)}{\hat{V}(\lambda)} \) for any \( \lambda \leq \lambda \). Hence we have \( \hat{V}'(\lambda) > \hat{\rho}'(\lambda) \hat{V}'(\lambda) \) and, consequently,

\[
\hat{V}'(\lambda) \hat{\rho}(\lambda) = \int_0^\lambda \hat{V}'(\lambda) \hat{\rho}'(\lambda) d\lambda > \int_0^\lambda \hat{\rho}'(\lambda) \hat{V}'(\lambda) d\lambda = \hat{\rho}'(\lambda) \hat{V}(\lambda)
\]

which gives [EC.3]. All inequalities are replaced with equalities if the generalize failure rate \( g \) is constant.

\[\blacksquare\]

**EC.2. Proof of Theorem 1**

We use standard sequence notation collected in the following definition.

**Definition EC.1 (Scaling Comparisons).** Given a non-negative function \( g \) with \( g(n) \to \infty \) as \( n \to \infty \) we write, \( \xi(n) = o(g(n)) \) if

\[
\lim_{n \to \infty} \frac{\xi(n)}{g(n)} = 0.
\]

We write \( f(n) = O(g(n)) \) if

\[
\limsup_{n \to \infty} \frac{\xi(n)}{g(n)} < \infty.
\]

Finally, \( \xi(n) = \Omega(g(n)) \) is the negation of \( \xi(n) = o(g(n)) \), i.e.,

\[
\xi(n) = \Omega(g(n)) \iff \liminf_{n \to \infty} \frac{\xi(n)}{g(n)} > 0.
\]
The characterization of the optimal actions of the social planner and the revenue maximizer in Theorem [EC.1] and [EC.2] is used to prove Theorem [1]. The proofs of these theorems appear then in [EC.3]

**Theorem EC.1. (optimal decisions of the SP)** The cutoffs and admission rates

\[
\hat{v}^n_{i,s} = \bar{v} + \varphi \frac{i-1}{K} \theta \frac{K-i+1}{n} - \frac{K-i+1}{2K}, \quad \text{and} \quad \hat{\lambda}^n_{i,S} = n - \Lambda f(\bar{v}) \varphi \theta \frac{i-1}{K} \frac{K-i+1}{n} - \frac{K-i+1}{2K},
\]

where

\[
\theta = \frac{\bar{F}(\bar{v})}{f(\bar{v})} \sqrt{\alpha}, \quad \varphi = 2 \frac{V(\bar{v}) - \rho(\bar{v})}{F(\bar{v})} = 2MRL(\bar{v}),
\]

are nearly optimal in the sense that \(S^n_K - S^n_K(\hat{\lambda}^n_S) = o(\sqrt{n})\).

The welfare maximizing decisions must be at most small perturbation of \(\hat{v}^n_S\) and \(\hat{\lambda}^n_S\). That is,

\[
v^n_{i,S} = \hat{v}^n_{i,S} + o(n^{-\frac{K-i+1}{2K}}), \quad i = 2, \ldots, K, \quad \text{and} \quad \hat{\lambda}^n_{i,S} = \hat{\lambda}^n_{i,S} + o(n^{-\frac{K-i+1}{2K}}), \tag{EC.4}
\]

Finally, increasing the number of classes beyond 2 can increase social welfare by at most \(o(\sqrt{n})\):

\[
\Lambda nV(\bar{v}) - 2\bar{v} \sqrt{\alpha n} + o(\sqrt{n}), \quad \text{for any} \ K \geq 2.
\]

The following is a strengthened version of Theorem 3 in [Nazerzadeh and Randhawa (2015)] where we add, to their result, that the optimal actions are asymptotically unique. This is crucial for the comparison of the social-planner and revenue-maximizer actions.

**Theorem EC.2. (optimal decisions of the RM)** The cutoffs and arrival rates

\[
\hat{v}^n_{i,R} = \bar{v} + \Phi \frac{i-1}{K} \theta \frac{K-i+1}{n} - \frac{K-i+1}{2K} + o(n^{-\frac{K-i+1}{2K}}), \quad i = 2, \ldots, K, \quad \text{and} \quad \hat{\lambda}^n_{i,R} = n - \Lambda f(\bar{v}) \Phi \frac{i-1}{K} \frac{K-i+1}{n} - \frac{K-i+1}{2K} \tag{EC.5}
\]

where

\[
\theta = \frac{\bar{F}(\bar{v})}{f(\bar{v})} \sqrt{\alpha}, \quad \Phi = 2 \left( \frac{f(\bar{v}) \bar{F}(\bar{v})}{f'(\bar{v}) \bar{F}(\bar{v}) + 2(f(\bar{v}))^2} \right) = \frac{2}{\Lambda \bar{v}'(1) f(\bar{v})^2},
\]

are nearly optimal in the sense that \(R^n_K - R^n_K(\hat{\lambda}^n_R) = o(\sqrt{n})\).

The welfare maximizing decisions must be at most small perturbation of \(\hat{v}^n_R\) and \(\hat{\lambda}^n_R\). That is,

\[
v^n_{i,R} = \hat{v}^n_{i,R} + o(n^{-\frac{K-i+1}{2K}}), \quad i = 2, \ldots, K, \quad \text{and} \quad \hat{\lambda}^n_{i,R} = \hat{\lambda}^n_{i,R} + o(n^{-\frac{K-i+1}{2K}}). \tag{EC.6}
\]

Finally, increasing the number of classes beyond 2 can increase revenue by at most \(o(\sqrt{n})\):

\[
R^n_K = \Lambda n\rho(\bar{v}) - 2\bar{v} \sqrt{\alpha n} + o(\sqrt{n}), \quad \text{for any} \ K \geq 2.
\]
In the statement of this theorem $\theta \geq 0$. Also, since $\tilde{\rho}(\lambda)$ is concave for $\lambda < \lambda_{0}^{*}$ (see Lemma 2) and, in particular, for $\lambda = 1 = \Lambda F(\bar{v}) < \Lambda F(\bar{v}^*_{o})$, we have that $\tilde{\rho}'(1) \geq 0$ so that $\Phi \geq 0$.

Theorems EC.1 and EC.2 provide the basis for Theorem 1: the constant $\gamma$ in Theorem 1 stands for $\varphi - \Phi$. With $K = 2$, if $\gamma = \varphi - \Phi > 0$, $\bar{\lambda}_{2,R}^{*} > \bar{\lambda}_{2,S}^{*}$ so that the revenue maximizer has a larger high priority class. The following lemma studies, then, $\gamma = \varphi - \Phi$.

**Lemma EC.1.** Suppose that $F$ has strictly positive density on its support and that $m(\cdot)$ is a convex (respectively concave, linear respectively) MRL. Then,

$$MRL(x) \left( \frac{h'(x)}{h(x)} + h(x) \right) \geq (\leq, = \text{respectively}) 1$$

for any $x$ in the support of $F$.

Recalling the definition of $\varphi$ and $\Phi$, we must show is that

$$\varphi = \frac{(V(\bar{v}) - \bar{v}F(\bar{v}))}{F(\bar{v})} \geq \frac{f(\bar{v})F(\bar{v})}{f'(\bar{v})F(\bar{v})} = \Phi.$$ 

and that the opposite holds for concave MRL. The left hand side of the inequality is precisely the MRL of $F$ at the point $\bar{v}$ so that this inequality is equivalent to

$$MRL(\bar{v}) \left( \frac{h'(\bar{v})}{h(\bar{v})} + h(\bar{v}) \right) \geq 1,$$

which follows, with convex MRL, from Lemma EC.1.

**Proof of Theorem 2** By definition, the optimal objective function value when policies are restricted to non-preemption is smaller than the optimal value under the larger family of pre-emptive policies. That is, $S_{K,NP}^{n} \leq S_{K}^{n}$. Then, we have

$$S_{K,NP}^{n}(\bar{v}^*_{S}) \leq S_{K,NP}^{n} \leq S_{K}^{n} = S_{2}^{n} + o(\sqrt{n}) = S_{2}^{n}(\bar{v}^*_{S}) + o(\sqrt{n}),$$

where $\bar{v}^*_{S,i} = \bar{v} + \varphi \frac{i-1}{2} \theta^{\frac{3i+1}{2}} - \frac{3i}{4}$ for $i = 1, 2$. The last two equalities follow from Theorem EC.1. This holds for all $K$, corresponding arrival rates for $\bar{v}^*_{S,i}$ are

$$\hat{\lambda}_{S,i}^{n} = n - C_{S,i} n^{\frac{K+\gamma+1}{2k}},$$

where $C_{S,i} = f(\bar{v}) \varphi \frac{i-1}{2} \theta^{\frac{3i+1}{2}}$. Similarly for revenue maximization

$$R_{K,NP}^{n}(\bar{v}^*_{R}) \leq R_{K,NP}^{n} \leq R_{K}^{n} = R_{2}^{n} + o(\sqrt{n}) = R_{2}^{n}(\bar{v}^*_{R}) + o(\sqrt{n}),$$

where $\bar{v}^*_{R,i} = \bar{v} + E_{i} n^{\frac{3i}{4}}$ with $E_{i} = \Phi \frac{i-1}{2} \theta^{\frac{3i+1}{2}}$ for $i = 1, 2$. Corresponding cumulative arrival rates for $\bar{v}^*_{R,i}$ are

$$\hat{\lambda}_{R,1}^{n} = n - C_{R,i} n^{\frac{K+\gamma+1}{2k}}.$$
where \( C_{R,i} = f(\hat{v}) \Phi^{i-1} \theta^3 \). It then suffices to prove that

\[
R^n_R(\hat{v}_R^n) - R^n_{2,NP}(\hat{v}_R^n) = o(\sqrt{n}), \quad \text{and} \quad S^n_s(\hat{v}_S^n) - S^n_{2,NP}(\hat{v}_S^n) = o(\sqrt{n}).
\]

Let us start with the revenue maximizer. Notice that

\[
R^n_{2,NP}(v^n_R) - R^n_2(v^n_R) = \Lambda n \Delta R_2(W^n_{R,H} - W^n_{R,L}) + \Lambda n \Delta R_1(W^n_{R,L} - W^n_{R,R}),
\]

where \( \Delta R_2 = \hat{\lambda}_{R,2} \hat{v}_{R,2} \) and \( \Delta R_1 = \hat{\lambda}_{R,1} \hat{v}_{R,1} - \hat{\lambda}_{R,2} \hat{v}_{R,2} \). With preemption, the steady-state sojourn times satisfy

\[
W^n_{R,L} = \frac{n}{(n - \hat{\lambda}_{R,1})(n - \hat{\lambda}_{R,2})} + \frac{n}{C_{R,1} n_1^{1/2} C_{R,2} n_3^{1/4}}, \quad W^n_{R,H} = \frac{\hat{\lambda}_{R,1}}{n(n - \hat{\lambda}_{R,2})} + \frac{n}{n C_{R,2} n_3^{3/4}} + \frac{1}{n}.
\]

Hence

\[
W^n_{R,L} - W^n_{R,H} = n^{-1} - C_{R,2} n^{-3/4}, \quad \text{and} \quad W^n_{R,R} - W^n_{R,H} = n^{-1} - C_{R,1} C_{R,2} n^{-5/4}.
\]

Then,

\[
\Lambda n \Delta R_2(W^n_{R,H} - W^n_{R,R}) = \hat{\lambda}_{R,2} \hat{v}_{R,2} (n^{-1} - C_{R,1} C_{R,2}^{-1} n^{-5/4}) = \mathcal{O}(1) = o(\sqrt{n}),
\]

and

\[
\Lambda n \Delta R_1(W^n_{R,L} - W^n_{R,R}) = \hat{\lambda}_{R,1} \hat{v}_{R,1} - \hat{\lambda}_{R,2} \hat{v}_{R,2} (n^{-1} - C_{R,1} C_{R,2}^{-1} n^{-5/4}) = \mathcal{O}(1) = o(\sqrt{n}),
\]

so that

\[
R^n_{2,NP}(v^n_R) - R^n_2(v^n_R) = \Lambda n \Delta R_2(W^n_{R,H} - W^n_{R,R}) + \Lambda n \Delta R_1(W^n_{R,L} - W^n_{R,R}) = \mathcal{O}(1) = o(\sqrt{n}).
\]

For the social planner,

\[
S^n_s(\hat{v}_S^n) - S^n_{2,NP}(\hat{v}_S^n) = \Lambda n V(\hat{v}_S^n)(W^n_{S,H} - W^n_{S,L}) + \Lambda n (V(\hat{v}_S^n) - V(\hat{v}_S^n))(W^n_{S,L} - W^n_{S,L}),
\]

where the expressions for the waiting times are the same as for the revenue maximizer with the obvious replacements of \( R \) with \( S \) everywhere. Using Taylor expansion on \( V(\cdot) \) at \( \hat{v} \) and that \( \hat{v}_{S,i} = \hat{v} + D_i n^{-3/4} \), we have

\[
\Lambda n V(\hat{v}_{S,2})(W^n_{S,H} - W^n_{S,L}) = \Lambda n (V(\hat{v}) - \hat{v} f(\hat{v}) D_2 n^{-1/4} + O(n^{-1/2}))(n^{-1} - C_{S,1} C_{S,2}^{-1} n^{-5/4}) = \mathcal{O}(1) = o(\sqrt{n}),
\]

and,

\[
\Lambda n V(\hat{v}_{S,1})(W^n_{S,H} - W^n_{S,L}) = \mathcal{O}(1) = o(\sqrt{n}).
\]
and
\[
\Lambda n (V(\tilde{v}_{S,1}^n) - V(\tilde{v}_{S,2}^n)) (W_{S,L}^n - W_{S,L}^p) = \Lambda n (\tilde{v} f(\tilde{v}) D_2 n^{-1/4} + \mathcal{O}(n^{-1/2})) (n^{-1} - C_{S,1} C_{S,2} n^{-5/4})
\]
\[
= \mathcal{O}(n^{-1/4}) = o(\sqrt{n}),
\]
so that \(S_{2}^n(\tilde{v}_S^n) - S_{2,SP}^n(\tilde{v}_S^n) = \mathcal{O}(1) = o(\sqrt{n})\), as stated.

In passing, it is worthwhile noticing the subtlety in the argument above. It builds on the fact that, under the optimal preemptive actions, the high-priority volume is order-of-magnitude larger than that of the low priority. The latter’s is of the order of \(n^{3/4}\); see equation (EC.9).

\[\Box\]

**EC.3. Proofs of Theorems [EC.1 and EC.2]**

We consider a sequence of queues indexed by the service rate \(n\). The nominal arrival rate in the \(n^{th}\) queue is \(\Lambda n\).

**A perturbation formulation:**

We express the cutoffs as deviations from \(\tilde{v} \equiv \bar{F}^{-1}(1/\Lambda)\): \(v_i = \tilde{v} + u_i\) or, in vector notation, \(v = \tilde{v} e + u\). As no customer with valuation smaller than \(\tilde{v} \equiv \bar{F}^{-1}(1/\Lambda)\) joins the queue \(u\) is a non-negative vector. Let \(W_{i}^n(u)\) be the expected waiting time of class \(i\) under preemptive static priority under the cutoff vector \(\tilde{v} e + u\). The social planner’s problem [1] with nominal arrival rate \(\Lambda n\) and service rate \(n\) is re-written as

\[
S_{K}^n = \max_{u_{i}^+} S_{K}^n(u) := \max_{u_{i}^+} \Lambda n \left[ V(\tilde{v} + u_1) - \alpha \sum_{i=1}^{K} (V(\tilde{v} + u_i) - V(\tilde{v} + u_{i+1})) W_{i}^n(\tilde{v} e + u) \right], \quad (SP_{n})
\]

and that for the revenue maximizer as

\[
R_{K}^n = \max_{u_{i}^+} R_{K}^n(u) := \max_{u_{i}^+} \Lambda n \left[ \rho(\tilde{v} + u_1) - \alpha \sum_{i=1}^{K} (\rho(\tilde{v} + u_i) - \rho(\tilde{v} + u_{i+1})) W_{i}^n(\tilde{v} e + u) \right]. \quad (RM_{n})
\]

Given optimal solutions \(u_{i,S}^{n^*}\) and \(u_{i,R}^{n^*}\) for \((SP_{n})\) (respectively \((RM_{n})\)), the optimal cutoffs are given by \(v_{i,S}^{n^*} = \tilde{v} + u_{i,S}^{n^*}\) (respectively \(v_{i,R}^{n^*} = \tilde{v} e + u_{i,R}^{n^*}\)).

We first state several auxiliary lemmas, the proofs of which appear at the end of this companion. The first of these, analogous to Lemma 2 in [Nazerzadeh and Randhawa (2015)], shows that the optimal cut-offs \(v_{i,R}^{n^*}\) and \(v_{i,S}^{n^*}\) are clustered around \(\tilde{v}\) when the volume is high.

**Lemma EC.2.** For each \(n\), there exist optimal solutions \(v_{i,S}^{n^*}\) and \(v_{i,R}^{n^*}\) for \((SP_{n})\) and \((RM_{n})\) respectively. Let \(\{(v_{i,R}^{n^*}, v_{i,S}^{n^*}); n = 1, 2, \ldots\}\) be a sequence of optimal solutions. Then, \(v_{i,R}^{n^*} \to \tilde{v}\) and \(v_{i,S}^{n^*} \to \tilde{v}\) as \(n \to \infty\).

That all “good” decisions must be small perturbations around \(\tilde{v}\) means that Taylor expansion should be useful in uncovering these perturbations.
Lemma EC.3. Fix a sequence of \( u^n = o(1) \) of cutoff values. Then,

\[
\sum_{i=1}^{K} u_i^n \beta (\bar{v}) + \frac{\alpha V' (\bar{v}) + \frac{1}{\Lambda (\bar{v})^2}}{u_i^n} - \left( \frac{1}{u_i^n} \beta (\bar{v}) + \gamma (\bar{v}) \sum_{i=1}^{K} \frac{u_i^n}{u_i^n} \right) + \epsilon^n, \tag{EC.10}
\]

where

\[
\beta (\bar{v}) := \frac{\alpha V (\bar{v})}{f (\bar{v})} - \frac{\alpha \bar{v}}{\Lambda (f (\bar{v}))}, \quad \gamma (\bar{v}) := \frac{\alpha \tilde{F} (\bar{v})}{2 f (\bar{v})},
\]

and

\[
\epsilon^n := O \left( \sum_{i=1}^{K} \frac{u_i^n}{u_i^n} \right) + O(n (u_i^n)^2) + O(1).
\]

Let \( u_i^n = \hat{u}_i^n + o(n^{-\frac{K+1}{2K}}) \) for \( 1 \leq i \leq K \) where \( \hat{u}_i^n = \varphi_i \frac{1}{n} \theta_i \frac{1}{n^{-\frac{K+1}{2K}}} \).

(Step 1) \( S_K^n - S_2^n = o(\sqrt{n}) \).

(Proof of Step 1) Re-write

\[
S_K^n (u^n) = n \Lambda V (\bar{v}) + M (\hat{u}_i^n) + B (u^n) + E (u^n), \tag{EC.11}
\]

where,

\[
B (u^n) = - \left( \frac{1}{u_i^n} \beta (\bar{v}) + \gamma (\bar{v}) \sum_{i=1}^{K} \frac{u_i^n}{u_i^n} \right), \quad M (u_i^n) = n \Lambda V' (\bar{v}) u_i^n + \frac{\alpha V' (\bar{v})}{\Lambda (f (\bar{v}))^2} u_i^n, \tag{EC.12}
\]

and, using Lemma [EC.3]

\[
E (u^n) = \sum_{i=1}^{K} O \left( \frac{(u_i^n)^2}{u_i^n} \right) + O(n (u_i^n)^2) + O(1). \tag{EC.13}
\]

Suppose that \( u^n \) is a sequence that has \( \epsilon_{i,n} := |u_i^n - \hat{u}_i^n| = \Omega \left( n^{-\frac{K+1}{2K}} \right) \) for some \( i \). Then, we will show that it must be sub-optimal.

Step 1.1. First, consider the case of \( i = 1 \), namely, that \( \epsilon_{1,n} := |u_1^n - \hat{u}_1^n| = \Omega(n^{-1/2}) \). We will show that, in this case

\[
M (\hat{u}_1^n) - M (u_1^n) = \Omega (n \epsilon_{1,n}) = \Omega (\sqrt{n}), \tag{EC.14}
\]

but

\[
B (\hat{u}^n) - B (u^n) = o (\sqrt{n}), \quad E (\hat{u}^n) - E (u^n) = o (n \epsilon_{1,n}), \tag{EC.15}
\]

so that, overall

\[
S_K (\hat{u}^n) - S_K^n (u^n) = \Omega (n \epsilon_{1,n}) = \Omega (\sqrt{n}). \tag{EC.16}
\]

In particular, \( u^n \) is sub-optimal for all \( n \) sufficiently large.
To prove (EC.14), notice that the function $M(\cdot)$ is maximized by

$$
\hat{u}^n_i = \frac{1}{\Lambda f(\bar{v})} n^{-\frac{1}{2}} \sqrt{\alpha} = \frac{\tilde{F}(\bar{v})}{f(\bar{v})} n^{-\frac{1}{2}} \sqrt{\alpha} n^{-\frac{1}{2}},
$$

where we used the fact that $\Lambda \tilde{F}(\bar{v}) = 1$ and, recall, $\theta = \frac{\tilde{F}(\bar{v})}{f(\bar{v})} \sqrt{\alpha}$. By Lemma EC.2, we can assume, without loss of generality, that $\epsilon_{i,n} = o(1)$. By definition

$$
\frac{M(\hat{u}^n_i) - M(\hat{u}^n_i + \epsilon_n)}{n\epsilon_{i,n}} = -\Lambda V'(\bar{v}) + \frac{\alpha V'(\bar{v})}{\Lambda (f(\bar{v}))^2 n\epsilon_{i,n}} (\hat{u}^n_i + \epsilon_n)
$$

$$
= -\Lambda V'(\bar{v}) + \frac{\alpha V'(\bar{v})}{\Lambda (f(\bar{v}))^2 n\theta} \left( \frac{\sqrt{n}}{\alpha \theta} + \epsilon_n \right)
$$

$$
= V'(\bar{v}) \left( -1 + \frac{1}{\bar{F}(\bar{v})} + \frac{\tilde{F}(\bar{v})}{f(\bar{v})} \frac{\alpha}{(\tilde{F}(\bar{v})/f(\bar{v}))^2} \frac{1}{\alpha + \theta \sqrt{n} \epsilon_{i,n}} \right),
$$

where we use again the definition of $\theta$ and the fact that $\Lambda = \frac{1}{\bar{F}(\bar{v})}$. Further simplification gives

$$
\frac{M(\hat{u}^{1,n}_i) - M(\hat{u}^{1,n}_i + \epsilon_{1,n})}{n\epsilon_{1,n}} = \frac{V'(\bar{v})}{\bar{F}(\bar{v})} \left( -1 + \left( \frac{\tilde{F}(\bar{v})}{f(\bar{v})} \right)^2 \frac{\alpha}{(\tilde{F}(\bar{v})/f(\bar{v}))^2} \frac{1}{\alpha + \theta \sqrt{n} \epsilon_{1,n}} \right)
$$

$$
= V'(\bar{v}) \left( -1 + \frac{1}{1 + \frac{\sqrt{n} \epsilon_{1,n}}{\theta}} \right), \quad \text{(EC.17)}
$$

Since $V'(\bar{v}) < 0$, $\bar{F}(\bar{v}) = \frac{1}{\Lambda} > 0$ and $\theta > 0$, we have the relation

$$
\liminf_{n \to \infty} \frac{M(\hat{u}^{1,n}_i) - M(\hat{u}^{1,n}_i + \epsilon_{1,n})}{n\epsilon_{1,n}} > 0 \iff \liminf_{n \to \infty} \sqrt{n} \epsilon_{1,n} > 0
$$

Since, $\epsilon_{1,n} = \Omega \left( n^{-\frac{1}{2}} \right)$, we have $\liminf_{n \to \infty} \sqrt{n} \epsilon_{1,n} > 0$ and, in particular, that

$$
\liminf_{n \to \infty} \frac{M(\hat{u}^{1,n}_i) - M(\hat{u}^{1,n}_i + \epsilon_{1,n})}{n\epsilon_{1,n}} > 0,
$$

equivalently,

$$
M(\hat{u}^{1,n}_i) - M(\hat{u}^{1,n}_i + \epsilon_{1,n}) = \Omega(n\epsilon_{1,n}).
$$

Since $\hat{u}^{1,n}_i$ is the maximizer of $M(\cdot)$ we have \text{(EC.14)}.

We next prove (EC.15) starting with $B(\cdot)$. Since $|B(\hat{u}^n) - B(\hat{u}^n)| \leq B(\hat{u}) + B(\hat{u}^n)$, it suffices to prove that $B(\hat{u}^n) = o(\sqrt{n})$ and $B(\hat{u}^n) = o(\sqrt{n})$. Since $\hat{u}_{i+1}^n = o(1)$ for all $i$ and since cut-offs satisfy $\hat{u}_{i+1}^n \geq \hat{u}_i$, it suffices to have $\frac{1}{\hat{u}_{i+1}^n} = O(\sqrt{n})$ to conclude that $\hat{u}_{i+1}^n = o(\sqrt{n})$ for all $i$ and, in turn, that $B(\hat{u}^n) = o(\sqrt{n})$. Similarly, it suffices to prove that $\frac{1}{\hat{u}_{i+1}^n} = O(\sqrt{n})$ to have $B(\hat{u}^n) = o(\sqrt{n})$. 


First, by definition, \( \hat{u}_i^n = \theta n^{-1/2} \) so that \( 1/\hat{u}_i^n = O(\sqrt{n}) \). Because \( \theta \sqrt{n} \epsilon_1 = \Omega(1) \) we have that

\[
\limsup_{n \to \infty} \frac{1}{\sqrt{n} \hat{u}_i^n} = \limsup_{n \to \infty} \frac{1}{\theta \sqrt{n} \epsilon_1} < \infty.
\]

Since \( \sqrt{n} \epsilon_1 = \Omega(1) \) by assumption, we also have that \( \frac{1}{\sqrt{n} \hat{u}_i^n + \sqrt{n} \epsilon_1} = O(1) \) and, in turn, that \( \frac{1}{\hat{u}_i^n + \epsilon_1} = O(\sqrt{n}) \).

We turn to \( E(\cdot) \). Since \( |E(\hat{u}^n) - E(u^n)| \leq E(\hat{u}^n) + E(u^n) \), it again suffices to show that \( E(\hat{u}) = o(\sqrt{n}) \) and \( E(u^n) = o(\sqrt{n}) \). Notice that

\[
E(\hat{u}^n) = \sum_{i=1}^K O \left( \left( \frac{\hat{u}_{i+1}^n}{\hat{u}_i^n} \right)^2 + O(n(\hat{u}_i^n)^2) \right) \tag{EC.18}
\]

\[
E(u^n) = \sum_{i=1}^K O \left( \left( \frac{\hat{u}_{i+1}^n + \epsilon_{i+1,n}}{\hat{u}_i^n + \epsilon_{i,n}} \right)^2 + O(n(\hat{u}_i^n + \epsilon_{i,n})^2) \right)\]

Since \( n(\hat{u}_i^n)^2 = n\theta^2 n^{-1} = \Theta(1) \) and \( n\hat{u}_i^n \epsilon_n = \theta \sqrt{n} \epsilon_n \), we have that \( O((n\hat{u}_i^n)^2) = O(1) = o(\sqrt{n}) \) and

\[
O(n(\hat{u}_i^n + \epsilon_{i,n})^2) = O(n(\hat{u}_i^n)^2 + 2n\hat{u}_i^n \epsilon_{i,n} + n\epsilon_{i,n}^2) = O(1 + \sqrt{n} \epsilon_{i,n} + n\epsilon_{i,n}^2) = O(n\epsilon_{i,n}^2) = o(n\epsilon_{i,n}).
\]

The second to last equality follows since \( \epsilon_{i,n} = \Omega \left( n^{-\frac{3}{2}} \right) \), \( \sqrt{n} \epsilon_{i,n} = O(n\epsilon_{i,n}^2) \) and the last equality follows since \( \epsilon_{i,n} = o(1) \). To take care of the other terms of \( E(\hat{u}^n) \) and \( E(u^n) \) notice that, since \( u^n = o(1) \) and \( \epsilon_{i,n} = o(1) \) for all \( i \),

\[
O \left( \frac{\hat{u}_{i+1}^n + \epsilon_{i+1,n}}{\hat{u}_i^n + \epsilon_{i,n}} \right)^2 = O \left( \frac{\hat{u}_{i+1}^n + \epsilon_{i+1,n}}{\hat{u}_i^n + \epsilon_{i,n}} \right) \quad \text{and} \quad O \left( \frac{\hat{u}_{i+1}^n}{\hat{u}_i^n} \right)^2 = o \left( \frac{\hat{u}_{i+1}^n}{\hat{u}_i^n} \right).
\]

It is therefore sufficient to show that

\[
o \left( \frac{\hat{u}_{i+1}^n + \epsilon_{i+1,n}}{\hat{u}_i^n + \epsilon_{i,n}} \right) = o(\sqrt{n}), \quad \text{and} \quad o \left( \frac{\hat{u}_{i+1}^n}{\hat{u}_i^n} \right) = o(\sqrt{n}).
\]

Moreover, since \( \hat{u}_{i+1}^n + \epsilon_{i+1,n} = o(1) \) and \( \hat{u}_i^n + \epsilon_{i,n} = o(1) \), it is sufficient to show that \( \frac{1}{\hat{u}_{i+1}^n + \epsilon_{i+1,n}} = O(\sqrt{n}) \) and \( \frac{1}{\hat{u}_i^n + \epsilon_{i,n}} = O(\sqrt{n}) \). We also know that \( \frac{1}{\sqrt{n}} < \frac{1}{\hat{u}_i^n} \) and \( \frac{1}{\sqrt{n} + \epsilon_{i,n}} < \frac{1}{\hat{u}_i^n + \epsilon_{i,n}} \). We already showed that \( \frac{1}{\sqrt{n}} = O(\sqrt{n}) \) and \( \frac{1}{\hat{u}_i^n + \epsilon_{i,n}} = O(\sqrt{n}) \).

This completes the proof of \( \text{EC.14} \) and \( \text{EC.15} \) and hence of \( \text{EC.16} \). We reached a contradiction to \( \|u^n - \hat{u}_i^n\| = \Omega(1/\sqrt{n}) \). Notice that we can repeat the above for any subsequence. We may thus conclude that \( |u^n - \hat{u}_i^n| = o(1/\sqrt{n}) \).

**Step 1.2.** We proved that any optimal sequence \( u^n \) must satisfy that \( u_i^n = \hat{u}_i^n + o(n^{-1/2}) \). We turn to prove that this, in turn, implies that any such sequence must have \( u_i^n = \hat{u}_i^n + o(n^{-K+1/2}) \) for \( i = 2, \ldots, K \). We will use the following lemma where, given a vector \( u \in \mathbb{R}_n^K \) and \( j < K \), we write \( u_{[j]} = (u_1, \ldots, u_j) \) and \( u_{-[j]} = (u_{j+1}, \ldots, u_K) \).
Lemma EC.4. Given \( u_1, \ldots, u_j \) for some \( j < K \),
\[
u_i = f_{i,j}(u_j) := \varphi_{K-j+1}^{i-j+1}(u_j) \frac{K-i}{K-j+1}, \quad i = j+1, \ldots, K,
\]
is the unique solution to
\[
\max_{0 \leq u_{j+1} \leq \ldots \leq u_K} B(u_{-j}; u_{[j]}).
\]

We define \( f_j(u_j) \) to be the vector \( f_{i,j}(u_j) \) for \( i = j+1, \ldots, K \). Notice that in the special case that \( j = 1 \), we have
\[
u_i = \varphi_{1-K}^{i-1}(u_1) \frac{K-i}{K-1}, \quad i = 2, \ldots, K.
\]

Fix a sequence of cutoffs \( \tilde{u}^n \) such that \( \tilde{u}_1^n = \hat{u}_1^n + o(1/\sqrt{n}) \) and such that \( \tilde{u}_k^n \) for \( k = 1, \ldots, K \) are determined by Lemma [EC.4] with \( j = 2 \) there. In particular, notice, \( \tilde{u}_i^n = \varphi^{1/K}(\tilde{u}_i^n) \frac{K-i+1}{K} = \hat{u}_i^n + o\left(n^{-\frac{K-i+1}{2K}}\right) \). Let \( u^n = o(1) \) be a sequence of cutoffs where \( u_1^n = \hat{u}_1^n + o(n^{-1/2}) \). We will show first that if \( |u_2^n - \tilde{u}_2^n| = \left(n^{-\frac{K-i+1}{2K}}\right) \), then \( u_2^n \) must be sub-optimal; specifically that
\[
S_K^n(\tilde{u}^n) - S_K^n(u^n) = \Omega(n^{-\frac{1}{2K}}) > 0.
\]

Recall that \( S_K^n(u) = M(u^n) + B(u^n) + E(u^n) \) where \( M, B \) and \( E \) are as defined in [EC.12] and [EC.13]. First, because both \( \tilde{u}_1^n = \hat{u}_1^n + o(n^{-1/2}) \) and \( u_1^n = \hat{u}_1^n + o(n^{-1/2}) \) and, because \( u^n = o(1) \), we have that \( M(\tilde{u}_1^n) - M(u_1^n) = \mathcal{O}(1) \), \( B(u_{-1}^n, u_1^n) - B(u_{-1}^n, \tilde{u}_1^n) = \mathcal{O}(1) \) and \( E(u_{-1}^n, u_1^n) - E(u_{-1}^n, \tilde{u}_1^n) = \mathcal{O}(1) \). It suffices, then to consider sequence \( u^n \) with \( u_1^n = \tilde{u}_1^n \). In that case,
\[
S_K^n(\tilde{u}^n) - S_K^n(u_{-1}^n, \tilde{u}_1^n) = B(\tilde{u}_{-1}^n; \tilde{u}_1^n) - B(u_{-1}^n, \tilde{u}_1^n) + E(\tilde{u}_{-1}^n; \tilde{u}_1^n) - E(\tilde{u}_{-1}^n, \tilde{u}_1^n).
\]

Since cutoff vectors are increasing \( (u_{i+1}^n > u_i^n; i = 1, \ldots, K-1) \), we have that \( E(u^n_{-1}) = \mathcal{O}(1) \). Since we are considering in this step the case that \( u_1^n = \tilde{u}_1^n + o(n^{-1/2}) \), we also have that \( n(u_1^n)^2 = \mathcal{O}(1) \). Thus, for any sequence of cutoff \( u^n \) that has \( u_1^n = \hat{u}_1^n + o(n^{-1/2}) \):
\[
E(u^n) = E(u^n) = \sum_{i=1}^{K} \mathcal{O}\left(\frac{(u_{i+1}^n)^2}{u_i^n}\right) + \mathcal{O}(n(u_1^n)^2) + \mathcal{O}(1) = o(-B(u^n)) + \mathcal{O}(1). \tag{EC.19}
\]

By definition of \( \tilde{u}^n \) we have \( 0 \geq B(\tilde{u}_{-1}^n; \tilde{u}_1^n) \geq B(u_{-1}^n; \tilde{u}_1^n) \) so that we further have
\[
S_K^n(\tilde{u}^n) - S_K^n(u_{-1}^n, \tilde{u}_1^n) = B(f_1^n; \tilde{u}_1^n) - B(u_{-1}^n, \tilde{u}_1^n) + E(\tilde{u}_{-1}^n; \tilde{u}_1^n) - E(u_{-1}^n, \tilde{u}_1^n) > B(\tilde{u}_{-1}^n; \tilde{u}_1^n) - B(u_{-1}^n, \tilde{u}_1^n) + \epsilon B(u_{-1}^n; \tilde{u}_1^n) + \mathcal{O}(1) \geq B(\tilde{u}_{-1}^n; \tilde{u}_1^n) - (1 - \epsilon) B(u_{-1}^n, \tilde{u}_1^n) + \mathcal{O}(1),
\]
where can be taken to be an arbitrarily small strictly positive constant.

By definition \( 0 \geq B(f^n_2(u^n_2); u^n_{[2]}) \geq B(u^n_1; \tilde{u}^n_1) \), so that, further

\[
S^n_\delta(\tilde{u}^n) - S^n_\delta(\tilde{u}^n_1, u^n_{[1]}) \geq B(\tilde{u}^n_1; \tilde{u}^n_1) - \beta \frac{(v)(K-2)(u^n_2)}{\varphi(u^n_2)^{\frac{1}{\gamma}}} \geq B(\tilde{u}^n_1; \tilde{u}^n_1) - (1-\epsilon)B(f^n_2(u^n_2); u^n_{[2]}) + O(1), \tag{EC.20}
\]

By lemma [EC.4] with \( j = 2 \), we have

\[
B(f_2(u^n_2), u^n_{[2]}) = -\gamma(v) \frac{u^n_2}{\bar{u}^n_1} + \varphi \frac{1}{\gamma} \gamma(v)(K-2)(u^n_2)\varphi^{-\frac{1}{\gamma}} + \varphi \frac{1}{\gamma} \beta(v) (u^n_2) \varphi^{-\frac{1}{\gamma}} = -\gamma(v) \frac{u^n_2}{\bar{u}^n_1} + (K-1)\varphi \frac{1}{\gamma} (u^n_2) \varphi^{-\frac{1}{\gamma}}
\]

\[
= -\gamma(v) \left( \frac{1}{\bar{u}^n_1} \right)^{\frac{1}{\gamma}} \left( \zeta^n + (K-1)\varphi \frac{1}{\gamma} \right)
\]

where \( \zeta^n = \frac{u^n_2}{\bar{u}^n_1} = (1/\varphi)^{\frac{1}{\gamma}} = u^n_2(\bar{u}^n_1)^{-\frac{K-1}{\gamma}} \) and we used the fact that \( \frac{\beta(v)}{\gamma(v)} = \varphi \). Notice that \( g(x) = x + (K - 1)\varphi \frac{1}{\gamma} (\frac{1}{x})^{\frac{1}{\gamma}} \) is convex in \( x \) and minimized at \( x^* = \varphi \frac{1}{k} \) with \( g(x^*) = \varphi \frac{1}{k} K \). Notice that

\[
B(\tilde{u}^n) = B(f_1(\tilde{u}^n_1), \tilde{u}^n_1) = -\varphi \frac{1}{\gamma} \gamma(v) K(\tilde{u}^n_1)^{-\frac{1}{\gamma}} = -\gamma(v) \left( \frac{1}{\bar{u}^n_1} \right)^{\frac{1}{\gamma}} g(x^*). \tag{EC.21}
\]

Suppose that \( \eta := \liminf_{\eta \to \infty} \zeta^n > \varphi \frac{1}{k} \) we have a constant \( c_{\eta} > 1 \) such that

\[
B(f_2(u^n_{[2]}), u^n_2) \leq c_{\eta} B(\tilde{u}^n),
\]

for all sufficiently large \( n \) in which case by [EC.23] (choosing \( \epsilon \) so that \( \xi := c_{\eta}(1-\epsilon) - 1 > 0 \); recall \( \epsilon \) was arbitrary) we have that

\[
S^n_\delta(\tilde{u}^n) - S^n_\delta(\tilde{u}^n_1, u^n_{[1]}) \geq B(\tilde{u}^n_1; \tilde{u}^n_1) - (1-\epsilon)B(f^n_2(u^n_2); u^n_{[2]}) + O(1)
\]

\[
\geq -\zeta B(\tilde{u}^n) + O(1) = \Omega(n^{1/k}) > 0. \tag{EC.22}
\]

The last equality follows since \( \tilde{u}^n_1 = \hat{u}^n_1 + o(n^{-1/2}) = \theta n^{-1/2} + o(n^{-1/2}) \) so that \( -B(\tilde{u}^n) = \gamma(v) \left( \frac{1}{\bar{u}^n_1} \right)^{\frac{1}{\gamma}} g(x^*) = \Omega(n^{1/2}) \) and we would conclude that, for all \( n \) sufficiently large \( S^n_\delta(\tilde{u}^n) > S^n_\delta(\tilde{u}^n_1, u^n_{[1]}) \) meaning that \( u^n \) is sub-optimal. The same argument applies if, instead of \( \liminf_{\eta \to \infty} \zeta^n > \varphi \frac{1}{k} \), we have \( \limsup_{\eta \to \infty} \zeta^n < \varphi \frac{1}{k} \).

It remains to prove that one of these hold if \( |u^n_2 - \tilde{u}^n_2| = \Omega\left(n^{-\frac{K-1}{2K}}\right) \). Indeed, if \( u^n_2 = \hat{u}^n_2 + \Omega(n^{-\frac{K-1}{2K}}) \), then there exist \( \delta > 0 \) such that, for all sufficiently large \( n \), either

\[
u^n_2 > \varphi \frac{1}{k} \theta n^{-\frac{K-1}{2k}} + \delta n^{-\frac{K-1}{2k}}, \text{ or } u^n_2 < \varphi \frac{1}{k} \theta n^{-\frac{K-1}{2k}} - \delta n^{-\frac{K-1}{2k}}.
\]
Thus, there exists \( \tilde{\delta} > 0 \) such that \( \zeta^n = u^n_2(\tilde{u}_1^n) = u^n_2(\tilde{u}_1^n) - \frac{K-1}{K} = u^n_2(\theta n^{-1/2} + o(\sqrt{n})) - \frac{K-1}{K} > \varphi^{1/K} + \tilde{\delta} \) or \( \varsigma^n \leq \varphi^{1/K} - \tilde{\delta} \) for all sufficiently large \( n \) so that (EC.22) holds. We conclude that any optimal sequence \( u^n \) must have that \( |u^n_2 - \tilde{u}_2^n| = o(n^{-\frac{K-1}{2K}}) \).

Now, one proceeds sequentially. Fixing a sequence such that \( \hat{u}_n^{v_1} = \tilde{u}_1^n \) and \( u_2^n = \hat{u}_2^n \) but \( |u_3^n - \hat{u}_3^n| = \Omega(n^{-\frac{K-1}{2K}}) \) we have

\[
S^n_2(\tilde{u}^n) - S^n_2(\hat{u}^n_2, u_{-2}^n) \geq B(\tilde{u}^n_2; \hat{u}^n_2) - (1 - \epsilon)B(f_n^{\ast}_1(u^n_1); u^n_1) + O(1), \tag{EC.23}
\]

and one proceeds similarly to our argument above to show that \( u^n \) is sub-optimal if \( |u^n_3 - \hat{u}_3^n| = \Omega(n^{-\frac{K-1}{2K}}) \). One then proceeds to \( u_4 \) and so on.

(Proof of Step 2) In this step, we first calculate the optimal objective function value for the social planner by using the optimal decisions we find in step 2. To do so, we consider \( S^n_K(\cdot) \) as defined in (EC.11)

\[
S^n_K(u^{v_1}) = M(u^{v_1}) - M(\tilde{u}^{v_1}) + M(\hat{u}^{v_1}) + B(u^{v_1}) - B(\tilde{u}^{v_1}) + Y(\hat{u}^{v_1}) + E(u^{v_1})
\]

where \( \hat{u}^n \equiv (\hat{u}_1^n, \cdots, \hat{u}_K^n) \) and \( u^{v_1} \equiv (u_1^{v_1}, \cdots, u_K^{v_1}) \). Note that \( E(u^{v_1}) = O(1) \) and \( M(u^{v_1}) - M(\hat{u}^{v_1}) = o(\sqrt{n}) \) by using (EC.17). Similarly, we have \( B(u^{v_1}) - B(\hat{u}^{v_1}) = o(\sqrt{n}) \). Hence we have

\[
S^n_K = n\Lambda V(\tilde{v}) - 2\tilde{v}\sqrt{n}\sqrt{\alpha} - n^{\frac{1}{2K}} \left( \beta(\tilde{v}) \varphi^{\frac{K-1}{K}} \theta^{\frac{1}{K}} - \gamma(\tilde{v})(K-1)\varphi^{\frac{1}{K}} \theta^{-\frac{1}{K}} \right) + o(\sqrt{n}) + O(1)
\]

for \( K \geq 2 \). Hence, two class policy is asymptotic optimal on \( o(\sqrt{n}) \) scale, i.e.

\[
\lim_{n \to \infty} \frac{S^n_K - S^n_2}{\sqrt{n}} = 0.
\]

Now, we check if it’s worth to offer whether two classes or only single class. Therefore, we calculate the optimal objective function value for the case of single class. The social planner’s problem reduces to the following when \( K = 1 \)

\[
\max_{v_1^n \in \mathbb{R}_+} S^n_1 = n\Lambda [V(v_1^n) - \alpha d_K V(v_1^n)]
\]

\[
= \max_{u_1^n \in \mathbb{R}_+} n\Lambda V(u_1^n) \left[ 1 - \frac{1}{n\Lambda f(\bar{v})} - O\left(\frac{1}{n}\right) \right]
\]

\[
= \max_{u_1^n \in \mathbb{R}_+} n\Lambda \left[ V(\tilde{v}) + V'(\tilde{v})u_1^n + O\left( (v_1^n)^2 \right) \right] \left[ 1 - \frac{1}{n\Lambda f(\bar{v})} - O\left(\frac{1}{n}\right) \right]
\]

\[
= n\Lambda V(\tilde{v}) - \frac{\alpha V'(\tilde{v})}{f(\bar{v})} + \max_{u_1^n \in \mathbb{R}_+} -\alpha V(\tilde{v}) + n\Lambda V'(\tilde{v})u_1^n + O\left(n(v_1^n)^2 + 1\right)
\]
which gives the following optimal $u^*_i$ and $S^*_i$

$$u^*_i = \sqrt{\alpha n^{-\frac{1}{2}}} \sqrt{\frac{\bar{F}(\bar{v})}{f(\bar{v})}} \sqrt{\frac{V(\bar{v})}{\bar{v}f(\bar{v})}} + o\left(n^{-\frac{1}{2}}\right)$$

$$S^*_1 = n\Lambda V(\bar{v}) + \alpha \bar{v} - 2\sqrt{\alpha \sqrt{n}} \sqrt{\bar{v}} \sqrt{\frac{V(\bar{v})}{F(\bar{v})}} + O(1)$$

Then we have

$$\lim_{n \to \infty} \frac{S^*_1 - S^*_K}{\sqrt{n}} = 2\bar{v} \sqrt{\alpha} - 2\sqrt{\alpha \sqrt{n}} \sqrt{\bar{v}} \sqrt{\frac{V(\bar{v})}{F(\bar{v})}}$$

where $K \geq 2$. Therefore, there is a significant benefit of offering more than 1 class and we can conclude that offering 2 classes is asymptotically optimal on $o(\sqrt{n})$ scale.

**Proof of Theorem EC.2.** Similar to the proof of Theorem EC.1. We omit the details.

**EC.4. Proofs of Auxiliary lemmas**

**Proof of Lemma EC.2.** We prove this result for the social planner. The argument for the revenue maximizer requires only minor changes. We first prove that $v^*_i \to \bar{v}$ and then proceed to show that $v^*_i \to \bar{v}$ for all $i = 2, \ldots, K$.

Suppose $v^*_i \to v^0$ for some vector $v^0 \in \mathbb{R}^K_+$ (if the sequence does not converge we can apply to argument below to any convergent subsequence). Suppose, further, that $v^0_i > \bar{v}$. We will prove that this leads to a contradiction to the optimality of $v^*_i$.

Since the cutoffs $u^*_i$ increase in $i$, we then have that $u^*_i = v^*_i - \bar{v} = \Omega(1)$. This, in turn, implies (recall that $\Lambda \bar{F}(\bar{v}) = 0$) the existence of $\delta < 1$ such that $\Lambda \bar{F}(v^*_i) \leq \delta$ for all $i = 1, \ldots, K$, and all $n$ sufficiently large. Consequently, for all such $n$,

$$W^*_i(\bar{v} + u^*) = \frac{1}{n(1 - \Lambda \bar{F}(v^*_i))(1 - \Lambda \bar{F}(v^*_i))} \leq \frac{1}{n(1 - \delta)^2} = O(1), \ i = 1, \ldots, K,$$

so that $AnW^*_i(\bar{v} + u^*) = O(1)$. Since $0 < V(x) \leq V(\bar{v})$ for all $x \geq \bar{v}$, we then have

$$S^*_K(u^*) = An \left[ V(\bar{v} + u^*_i) - \alpha \sum_{i=1}^{K} (V(\bar{v} + u^*_i) - V(\bar{v} + u^*_i)) W_i^*(\bar{v} + u^*) \right]$$

$$= An \left[ V(\bar{v} + u^*_i) \right] + O(1).$$

Take $u = (u^*_i/2, u^*_2, \ldots, u^*_K)$ to be the vector obtained from $u^*$ by replacing $u^*_i$ with $u^*_i/2$ and keeping all other entries the same. Let $u^0 = \bar{v} + u^*$. Notice that

$$u^0 \to u^0 = \left( \frac{v^0_1 - \bar{v}}{2}, v^0_2, \ldots, v^0_K \right).$$
Since $v_i^0 - \frac{v_i^0 - \bar{v}}{2} > \bar{v}$ we have, as before, that $W_i^n = O(1/n)$ for all $i$ and all sufficiently so that

$$S_K^n(u^n) = \Lambda n \left[ V\left(\bar{v} + \frac{u_i^{n*}}{2}\right)\right] + O(1)$$

Therefore,

$$S_K^n(u^n) - S_K^n(v^{n*}) = \Lambda n \left[ V\left(\bar{v} + \frac{u_i^{n*}}{2}\right) - V(\bar{v} + u_1^{n*})\right] + O(1)$$

It follows from the strictly positive density of $F$, and from $u_1^{n*} \rightarrow v_1^0 - \bar{v}$ and $u_1^{n*}/2 \rightarrow (v_1^0 - \bar{v})/2$, that

$$V\left(\bar{v} + \frac{u_i^{n*}}{2}\right) - V(\bar{v} + u_1^{n*}) = \int_{u_1^{n*}}^{u_i^{n*}} x f(x) dx \geq \frac{u_i^{n*}}{2} \left(\int_{u_1^{n*}}^{u_i^{n*}} f(x) dx - \int_{u_1^{n*}}^{u_i^{n*}} f(x) dx\right) = \Omega(1), \quad (EC.24)$$

and, in particular, that

$$S_K^n(u^n) - S_K^n(v^{n*}) \geq \Lambda n \frac{u_i^{n*}}{2} \left(\int_{u_1^{n*}}^{u_i^{n*}} f(x) dx - \int_{u_1^{n*}}^{u_i^{n*}} f(x) dx\right) + O(1) \geq Ln,$$

for some $L > 0$ contradicting the optimality of $v^{n*}$. We conclude that optimal cutoffs must satisfy $v_1^{n*} \rightarrow \bar{v}$.

From this first step it follows in particular that a limit $v^0$ of $v^{n*}$ must have $v_i^0 = \bar{v}$. Suppose that there exists an index $i \leq K$ such that $v_i^0 > \bar{v}$. Let $i_0$ be the smallest such index. Then, $W_i^n(\bar{v}e + u^{n*}) = O(1/n)$ for all $i \geq i_0$ and, in turn,

$$S_K^n(u^{n*}) = \Lambda n \left[ V(\bar{v} + u_1^{n*}) - \alpha \sum_{i=1}^{i_0-1} (V(\bar{v} + u_i^{n*}) - V(\bar{v} + u_{i+1}^{n*})) W_i^n(\bar{v}e + u^{n*})\right] + O(1).$$

Replicating our arguments above, take

$$u^n = \left(u_1^{n*}, ..., u_{i_0-1}^{n*}, \frac{u_{i_0}^{n*}}{2}, u_{i_0+1}^{n*}, ..., u_K^{n*}\right) \rightarrow \left(0, ..., \frac{v_{i_0}^0 - \bar{v}}{2}, ..., \frac{v_K^0}{2}\right).$$

Notice that, by the definition of $i_0$ we have $u_{i_0}^{n*}/2 > u_{i_0-1}^{n*}$ for all sufficiently large $n$ so the monotonicity of cutoffs (in $i$) is maintained. Let $v^n = \bar{v}e + u^n$. Then, also for $u^n$,

$$S_K^n(u^n) = \Lambda n \left[ V(\bar{v} + u_1^n) - \alpha \sum_{i=1}^{i_0-1} (V(\bar{v} + u_i^n) - V(\bar{v} + u_{i+1}^n)) W_i^n(\bar{v}e + u^n)\right] + O(1),$$

so that

$$S_K^n(u^n) - S_K^n(u^{n*}) = -\alpha \Lambda n (V(\bar{v} + u_{i_0-1}^{n*}) - V(\bar{v} + u_{i_0}^{n*}/2)) W_{i_0-1}^n(\bar{v} + u^n)$$

$$+ \alpha \Lambda n (V(\bar{v} + u_{i_0-1}^{n*}) - V(\bar{v} + u_{i_0}^{n*})) W_{i_0-1}^n(\bar{v} + u^{n*}) + O(1)$$
Using
\[ W_{i_0-1}^n (\bar{v} + u_{i_0}^{n*}) = \frac{1}{n \left( 1 - \Lambda F(\bar{v} + u_{i_0}^{n*}) \right) \left( 1 - \Lambda F(\bar{v} + u_{i_0}^{n*}) \right)}, \]
\[ W_{i_0-1}^n (\bar{v} + u^n) = \frac{1}{n \left( 1 - \Lambda F(\bar{v} + u_{i_0}^{n*}) \right) \left( 1 - \Lambda F(\bar{v} + u_{i_0}^{n*}) \right)}, \]
we get
\[ S_K^n (\underline{u}) - S_K^n (u_{i_0}^{n*}) = \frac{\alpha \Lambda}{1 - \Lambda F(\bar{v} + u_{i_0}^{n*})} (g^n (u_{i_0}^{n*}) - g^n (u_{i_0}^{n*}/2)) + o(1), \]
where, for \( x \geq u_{i_0}^{n*} \),
\[ g^n (x) := \frac{V (\bar{v} + u_{i_0}^{n*}) - V (\bar{v} + x)}{1 - \Lambda F(\bar{v} + x)} = \frac{V (\bar{v}) - V(\bar{v} + x)}{1 - \Lambda F(\bar{v} + x)} + \frac{V (\bar{v} + u_{i_0}^{n*}) - V(\bar{v})}{1 - \Lambda F(\bar{v} + x)}. \]

Notice that since \( u_{i_0}^{n*} \to 0 \), we have that \((1 - \Lambda F(\bar{v} + u_{i_0}^{n*}))^{-1} \to \infty \) as \( n \to \infty \) so that, to prove that \( u_{i_0}^{n*} \) is sub-optimal, it suffices to show that \((g^n (u_{i_0}^{n*}) - g^n (u_{i_0}^{n*}/2)) = \Omega(1) \) as this will imply that \( S_K^n (\underline{u}) - S_K^n (u_{i_0}^{n*}) \to \infty \) as \( n \to \infty \).

To that end, let
\[ \bar{g}(x) := \frac{V (\bar{v}) - V(\bar{v} + x)}{1 - \Lambda F(\bar{v} + x)} = \frac{V (\bar{v}) - V(\bar{v} + x)}{\Lambda F(\bar{v}) - \Lambda F(\bar{v} + x)}. \]
\( (\Lambda \bar{g}(x) \) is the expected valuation conditional on it being between \( \bar{v} \) and \( \bar{v} + x \). Since \( F \) is assumed \( F \) to have a strictly positive density, \( \bar{g}(x) \) is strictly increasing in \( x \) so that, since \( u_{i_0}^{n*}/2 = \Omega(1) \) and \( u_{i_0}^{n*} - u_{i_0}^{n*}/2 = \Omega(1) \),
\[ \bar{g}(u_{i_0}^{n*}) - \bar{g}(u_{i_0}^{n*}/2) = \Omega(1). \]

Also, since \( u_{i_0}^{n*} = \Omega(1) \) but \( u_{i_0}^{n*} \to 0 \), we have that
\[ \frac{V (\bar{v} + u_{i_0}^{n*}) - V(\bar{v})}{1 - \Lambda F(\bar{v} + u_{i_0}^{n*})} = o(1), \]
and the same holds with \( u_{i_0}^{n*} \) replaced with \( u_{i_0}^{n*}/2 \). Combined, we have that
\[ g^n (u_{i_0}^{n*}) - g^n (u_{i_0}^{n*}/2) = \bar{g}(u_{i_0}^{n*}) - \bar{g}(u_{i_0}^{n*}/2) + o(1) = \Omega(1), \]
contradicting the optimality of \( v_{i_0}^{n*} \).

Finally, in repeating the proof for the revenue maximizer, \( \bar{g}(x) \) will be replaced by \((\rho(\bar{v}) - \rho(\bar{v} + x))/\Lambda F(\bar{v}) - \Lambda F(\bar{v} + x) \) which is increasing in \( x \) by Lemma 2. \( \blacksquare \)
Proof of Lemma EC.3: The sequence $u^n$ is fixed throughout the proof. For simplicity of notation, we write $W^n_i$ for $W^n_i(\bar{v} + u^n)$. Recall that

$$ S^n_K(v^n) = n\Lambda \left[ V(v^n_i) - \alpha \sum_{i=1}^{K-1} W^n_i \left( V(v^n_i) - V(v^n_{i+1}) \right) - \alpha W^n_K V(v^n_K) \right] $$

(EC.25)

Taking a Taylor expansion of $V(v^n_i)$ for $i = 1, 2, \cdots, K$ around $\bar{v}$ and recalling $u^n_i = \bar{v} - v^n_i$ we have

$$ S^n_K(v^n) = n\Lambda \left[ (V(\bar{v}) + V'(\bar{v}) u^n_1) - \alpha W^n_K V(\bar{v}) \alpha V'(\bar{v}) \sum_{i=1}^{K-1} W^n_i (u^n_i - u^n_{i+1}) \right. $$

$$ \left. - \frac{V''(\bar{v})}{2} \sum_{i=1}^{K-1} W^n_i \left( (u^n_i)^2 - (u^n_{i+1})^2 \right) - \alpha W^n_K (V'(\bar{v}) u^n_K) + \epsilon^n \right] $$

(EC.26)

where

$$ \epsilon^n = \mathcal{O} \left( \sum_{i=1}^{K-1} W^n_i (u^n_{i+1})^3 \right) + \mathcal{O} \left( nW^n_K (u^n_K)^2 \right) + \mathcal{O}(n(u^n_1)^2), $$

(EC.27)

Collecting terms we have

$$ S^n_K(v^n) = n\Lambda \left[ (V(\bar{v}) + V'(\bar{v}) u^n_1) - \alpha W^n_K V(\bar{v}) \alpha V'(\bar{v}) \sum_{i=1}^{K-1} W^n_i (u^n_i - u^n_{i+1}) \right. $$

$$ \left. - \frac{V''(\bar{v})}{2} \sum_{i=1}^{K-1} W^n_i \left( (u^n_i)^2 - (u^n_{i+1})^2 \right) - \alpha W^n_K (V'(\bar{v}) u^n_K) \right] $$

(EC.28)

We next apply Taylor expansion to $W^n_i$. First, using $\Lambda F(\bar{v} + u^n_K) = \Lambda F(\bar{v}) + \Lambda f(\bar{v}) u^n_K + \Lambda O((u^n_K)^2)$ and recalling that $\Lambda F(\bar{v}) = 1$, we have

$$ W^n_K = \frac{1}{n \left( 1 - \Lambda F(\bar{v} + u^n_K) \right)} = \frac{1}{n\Lambda f(\bar{v}) u^n_K + n\Lambda O((u^n_K)^2)} $$

(EC.29)

Next, taking further the Taylor expansion of $W_K$ at $n\Lambda f(\bar{v}) u^n_K$ we get

$$ W^n_K = \frac{1}{n\Lambda f(\bar{v}) u^n_K} + \frac{1}{n\Lambda} \sum_{i=1}^{\infty} (-1)^i \frac{\mathcal{O}((u^n_K)^2)^i}{(f(\bar{v}) u^n_K)^{i+1}} $$

$$ = \frac{1}{n\Lambda f(\bar{v}) u^n_K} + \mathcal{O} \left( \frac{1}{n\Lambda} \right). $$

In the last equality we use Lemma EC.2 by which $u^n_K = o(1)$ so that there exist $L, \tilde{L} > 0$ such that

$$ \sum_{i=1}^{\infty} (-1)^i \frac{\mathcal{O}((u^n_K)^2)^i}{(f(\bar{v}) u^n_K)^{i+1}} \leq \sum_{i=1}^{\infty} \left| \frac{\mathcal{O}((u^n_K)^2)^i}{(f(\bar{v}) u^n_K)^{i+1}} \right| \leq \sum_{i=1}^{\infty} \frac{L^i (u^n_K)^{2i}}{(f(\bar{v}) |u^n_K|)^{i+1}} $$

$$ = \frac{L}{f(\bar{v})^2} \sum_{i=1}^{\infty} \frac{|Lu^n_K|^{i-1}}{f(\bar{v})} \leq \tilde{L}, $$
The last inequality follows from the fact that \( u^n_K = o(1) \) so that \( Lu^n_K \leq 1/2 \) for all \( n \) sufficiently large. For the Taylor expansion of \( W^n_i \) (\( i < K \)) we have

\[
W^n_i = \frac{1}{(1 - \Lambda f(\bar{v} + u^n_i)) (1 - \Lambda f(\bar{v} + u^n_{i+1}))} = \frac{1}{n \Lambda f(\bar{v}) u^n_i + n \Lambda O \left((u^n_i)^2\right)} \left(\Lambda f(\bar{v}) u^n_{i+1} + \Lambda O \left((u^n_{i+1})^2\right)\right).
\]

We take one more Taylor expansion to get

\[
W^n_i = \frac{1}{n (\Lambda f(\bar{v}))^2 u^n_i u^n_{i+1}} - \frac{f'(\bar{v})}{2n \Lambda^2 f(\bar{v})^3} \left(\frac{1}{u^n_i} + \frac{1}{u^n_{i+1}}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{1}{n u^n_i}\right) + O\left(\frac{1}{n u^n_{i+1}}\right)
\]

Since \( u^n_{i+1} > u^n_i \), \( O\left(\frac{1}{n u^n_i}\right) = O(1/n) \) so that

\[
W^n_i = \frac{1}{n (\Lambda f(\bar{v}))^2 u^n_i u^n_{i+1}} - \frac{f'(\bar{v})}{2n \Lambda^2 f(\bar{v})^3} \left(\frac{1}{u^n_i} + \frac{1}{u^n_{i+1}}\right) + O\left(\frac{1}{n u^n_i}\right)
\]

Plugging these back into (EC.27) we get

\[
e^n = O\left(\sum_{i=1}^{K-1} \left(\frac{u^n_{i+1}}{u^n_i}\right)^2\right) + O(n (u^n_i)^2) + O(1)
\]

and into (EC.28), hence objective function becomes (by also using the fact that \( u^n_{i+1} > u^n_i \))

\[
S^n_K(v^n) = n \Lambda \left((V(\bar{v}) + V'(\bar{v}) u^n_i) - \frac{\alpha V(\bar{v})}{n \Lambda f(\bar{v}) u^n_K} - \frac{\alpha V'(\bar{v})}{\Lambda f(\bar{v})} \sum_{i=1}^{K-1} \left(\frac{1}{u^n_{i+1}} - \frac{1}{u^n_i}\right)\right)
\]

\[
\frac{\gamma(\bar{v})}{2 \Lambda (f(\bar{v}))^2} \left(V'(\bar{v}) \frac{f'(\bar{v})}{f(\bar{v})} - V''(\bar{v})\right) = \frac{\alpha}{2 \Lambda f(\bar{v})} \frac{\bar{F}(\bar{v})}{2 f(\bar{v})},
\]

and

\[
\beta(\bar{v}) = \alpha V(\bar{v}) + \frac{\alpha V'(\bar{v})}{\Lambda (f(\bar{v}))^2} = \alpha \frac{V(\bar{v})}{f(\bar{v})} - \alpha \frac{\bar{v} \bar{F}(\bar{v})}{f(\bar{v})} = \alpha \frac{V(\bar{v} - \bar{v} \bar{F}(\bar{v}))}{f(\bar{v})},
\]

and hence

\[
\frac{\beta(\bar{v})}{\gamma(\bar{v})} = 2 \frac{V(\bar{v}) - \bar{v} \bar{F}(\bar{v})}{\bar{F}(\bar{v})}.
\]

(EC.30)
Proof of Lemma [EC.1] To simplify notation we replace \( m(x) := MRL(x) \). Using the relationship
\[
h(x) = \frac{m'(x) + 1}{m(x)},
\]
between the MRL and the hazard function, the inequality [EC.7] is equivalent to
\[
\frac{m'(x)}{m(x)} \geq -\frac{h'(x)}{h(x)} \tag{EC.32}
\]
Taking a derivative in (EC.31) we further have
\[
h'(x) = \frac{m''(x) m(x) - m'(x) (m'(x) + 1)}{m^2(x)}.
\]
so that inequality (EC.32) is equivalent to
\[
m'(x) \geq -\frac{m''(x) m(x) - m'(x) (m'(x) + 1)}{(m'(x) + 1)} \tag{EC.33}
\]
Any MRL has \( m(x) \geq 0, m'(x) \geq -1 \); see Lai and Xie (2006). Therefore, we can multiply both sides of (EC.33) by \( m'(x) + 1 \geq 0 \) and say the inequality there holds if and only if
\[
m''(x) m(x) \geq 0,
\]
which holds for all \( x \) if and only if \( m(\cdot) \) is convex. \( \blacksquare \)

Proof of Lemma [EC.4] We provide the detailed proof for the case that \( j = 2 \). The other cases follow identically.

Consider the change of variables \( x_i = u_{i+1}/u_i \) and the optimization problem
\[
\min \sum_{i=1}^{K-1} \gamma(\bar{v}) x_i + \beta(\bar{v}) x_K \quad \text{subject to } x_i \geq 0 \text{ and } \prod_{i=1}^{K} x_i \geq \frac{1}{u_1} \tag{EC.34}
\]
Because \( \prod_{i=1}^{K} x_i \) is a jointly concave function (see Marcus and Lopes (1957)) and the objective function is linear, this is a convex minimization problem. It therefore has a unique solution which, we will show, is given by
\[
x_i^* = \varphi^{1/K} \left( \frac{1}{u_1} \right)^{1/K}, \quad i = 1, 2, \ldots, K - 1, \text{ and } x_K^* = \varphi^{-(K-1)/K} \left( \frac{1}{u_1} \right)^{1/K}.
\]
Denote the KKT multipliers for \( x_i \geq 0 \) by \( \mu_i \) and for \( \prod_{i=1}^{K} x_i \geq \frac{C}{u_1} \) by \( \eta \). We claim that \( \mu_i \equiv 0 \) and
\[
\eta := \frac{\gamma(\bar{v})}{\prod_{j \neq i} x_j^*}
\]
satisfy both the complementary slackness and first-order (stationarity) conditions. The complementary slackness conditions are

\[ \eta \left( \prod_{i=1}^{K} x^*_i - \frac{1}{u_1} \right) = 0 \text{ and } \mu_i x_i = 0, \ i = 1, \ldots, K. \]

They are both satisfied under our solution. For the first-order (stationarity) conditions, under \( \mu_i \equiv 0 \), we must check \( \gamma(\bar{v}) = \eta \prod_{j \neq i} x^*_j, \ i = 1, 2, \ldots, K - 1 \) (derivative with respect to \( x_i \)) and \( \beta(\bar{v}) = \eta \prod_{j \neq K} x^*_j \) (derivative with respect to \( x_K \)).

Recall that (see Theorem EC.1 and equation (EC.30))

\[ \frac{\beta(\bar{v})}{\gamma(\bar{v})} = 2 \frac{V(\bar{v}) - \rho(\bar{v})}{F(\bar{v})} = \varphi. \]

Since,

\[ \frac{\prod_{j \neq K} x^*_j}{\prod_{j \neq i} x^*_j} = \frac{x^*_i}{x^*_K} = \frac{x_i}{x_K} = \varphi, \]

we have that

\[ \eta = \frac{\gamma(\bar{v})}{\prod_{j \neq i} x^*_j} = \frac{\beta(\bar{v})}{\prod_{j \neq K} x^*_j}, \]

which means that the first order conditions are satisfied with the proposed \( \eta \).

Finally, notice that with the change of variable \( x_i = u_{i+1}/u_i \) the minimization problem (EC.34) is equivalent to the minimization problem

\[ \max_{u_{-1}} B(u_1, u_{-1}) = \max_{u_{-1}} \left\{ - \left( \frac{1}{u_K} \beta(\bar{v}) + \gamma(\bar{v}) \sum_{i=1}^{K} \frac{u_{i+1}}{u_i} \right) \right\}, \]

because by definition \( \prod_{i=1}^{K} \frac{u_{i+1}}{u_i} = \frac{1}{u_1} \). Thus, from the solution \( x^* \) we construct the solution \( u_{-1}^* = \left( \frac{u_1}{\varphi} \right)^{\frac{K-i}{K}} \). This solution also satisfies our requirement that \( u_1 < u_2^* < \cdots < u_K^* \).