Optimal Spares Allocation in an Exchangeable-Item Repair System with Tolerable Wait

Michael Dreyfuss\textsuperscript{a}, Yahel Giat\textsuperscript{a,*}

\textsuperscript{a}Department of Industrial Engineering, Jerusalem College of Technology, Jerusalem, Israel

Abstract

In a multi-location, exchangeable-item repair system with spares, the expected waiting time and the fill rate measures are oftentimes used as the optimization criteria for the spares allocation problem. These measures, however, do not take into account that customers will tolerate a reasonable delay and therefore, the firm does not incur reputation costs if customers wait less than their tolerable wait. Accordingly, we generalize the expected waiting time and fill rate measures to reflect customer patience. These generalized measures are termed the truncated waiting time and the window fill rate, respectively. We develop efficient algorithms to solve the problem for each of the criteria and demonstrate how incorporating customer patience provides considerable savings and profoundly affects the optimal spares allocation.

Keywords: Inventory, logistics, truncated waiting time, window fill rate, optimization criteria

1. Introduction

Exchangeable-item repair systems have been investigated and applied in many civil and military settings. In these systems, customers bring a failed item and exchange it for a serviceable item that is available on stock. The failed item itself is repaired on site after which it is returned to stock. To in-
crease the availability of serviceable items in stock, spare items are placed in stock. If the system has multiple locations into which customers may arrive, then in addition to determining the number of spares to be purchased, managers must also decide how to allocate the spares in the various locations with the goal of optimizing the predetermined service measure.

The target level of service is by itself an important managerial consideration and lies at the heart of this paper. Researchers frequently assume that the firm’s objective is to minimize the expected number of backorders, which is equivalent to minimizing the expected waiting time, (e.g., Wong et al. 2006, Van der Heijden et al. 2013). The reasoning behind this is quite straightforward, as longer waiting times are associated with negative customer satisfaction and result with reputation losses to firms. The prevalent use of this measure is further facilitated by the fact that expected waiting time is easy to compute and is a convex function of the number of spares and therefore a simple greedy algorithm can attain the optimum efficiently. In reality, however, expected waiting time is rarely an accurate proxy for the firm’s costs. For example, in many cases firms are obliged, either through government regulation or by contractual commitment, to reduce the waiting time to below a threshold. From the customers’ standpoint, too, there is a certain tolerable or acceptable period of wait, which may depend on their level of patience or expectation (see Durrande-Moreau 1999). In these cases, the objective should be to minimize the expected wait beyond this tolerable wait.

In other situations, the cost to the firms does not depend on the time the customer waits but on the number of customers who wait. When this is the case, the fill rate measure, i.e., the portion of customers who are served upon arrival, is the appropriate criterion. Furthermore, if the firm is penalized only for the number of customers who wait beyond the tolerable wait then the fill rate measure should be generalized to reflect the firm’s correct objective. Accordingly, given a spares allocation \( \vec{n} \), we consider in this paper two measures of service performance to serve as the optimality criteria for the spares allocation problem:
• The truncated waiting time, $W(\bar{n}, t)$: The expected time waited beyond $t$ units of time. For $t = 0$, this measure is the expected waiting time.

• The window fill rate, $F(\bar{n}, t)$: The expected fraction of customers who are served within $t$ units of time. For $t = 0$, this measure is the fill rate.

The use of the expected waiting time and fill rate as a criterion for optimality has been considered repeatedly by researchers and dates back to Sherbrooke (1968). There is, however, a difficulty with the treatment of the fill rate. The fill rate measure is concave in the number of spares in each location only if each location has been allocated sufficiently many spares. Therefore, the researchers who considered the fill rate as a criterion for optimality imposed the strict limitation of considering only the region for which the fill rate is concave. The goals of this paper are therefore twofold. The first, to develop an algorithm that efficiently computes the spares allocation that maximizes the fill rate. More importantly, we aspire to understand how taking into account customer patience, (i.e., the tolerable wait), affects the optimal allocation.

To achieve these goals we build on Berg and Posner (1990) who derive a mathematical expression for the window fill rate. We show that the window fill rate is either concave, or initially convex and then concave in the number of spares in each location. Further, we show that for a sufficiently large tolerable wait, the window fill rate is concave shaped. These observations allow us to make a number of contributions to current theory in exchangeable-item repair systems:

First, we are novel in our generalization of the expected waiting time measure and develop a mathematical expression for the truncated waiting time. Since this expression depends on the window fill rate, we exploit the mathematical properties of the window fill rate to show that the truncated waiting time is decreasing and convex in the number of spares and therefore optimality can be attained efficiently using a greedy algorithm.

Second, we develop an algorithm to efficiently derive a near optimal solution to the problem of optimizing the window fill rate. Using the properties of the
window fill rate, we apply a special case of [Hackman and Platzman 1990] to define a concave covering function of each of the locations’ window fill rate, and optimize these covering functions. We find that if there are few spares and the tolerable wait is short, then spares must be allocated to only part of the locations leaving the others with no spares. We characterize the a priori bound for the distance from optimum of the suggested algorithm. For the cases that the solution is sub-optimal for the original problem, we characterize the a posteriori distance from optimum. We show that the distance from optimum depends on the functional value of only one location and therefore as the scale of the problem increases the a priori and a posteriori distances from optimum decrease.

Our paper also provides practical insights to managers, which we demonstrate through a large-scale numerical example. This example underscores the importance of defining correctly the criterion of optimality. In particular, we demonstrate how costly it is for companies who neglect to take into account the customers’ patience level. Second, we show how the optimal allocation for the truncated waiting time changes with the tolerable wait. As the tolerable wait increases, the relative importance of high-arrival locations is mitigated and these locations are allotted less spares. Third, we show that the fill rate is more sensitive to customer patience than the expected waiting time. This implies that accurately estimating the tolerable wait is essential particularly when the fill rate is the optimality criterion. Finally, we show that the window fill rate criteria creates two classes of locations, so that one class receives spares and the other does not. As a consequence, in this case, managers should develop two different policies with respect to their service time to customers.

2. Literature Review

Exchangeable-item inventory systems have been extensively researched in various settings [Sherbrooke 2004, Muckstadt 2005]. For a recent survey of this field of research see [Basten and van Houtum 2014]. There are two popular
service measures for these systems, the fill rate, i.e., the probability for a random customer to be served upon arrival and the expected backorders, i.e., the number of customers waiting to be served. See, for example, Sherbrooke (1968), Wong et al. (2006), Caggiano et al. (2007) and Kranenburg and van Houtum (2009), Basten et al. (2012), Tsai and Zheng (2013), Ghaddar et al. (2015). Berg and Posner (1990) extend the fill rate measure by deriving the waiting time distribution so that one can measure the probability to be served within any time window $t$. In our paper, we denote the waiting time distribution as the window fill rate. This measure is similar to Song (1998) who compute the order fill rate for a multi-item system with lost sales, Caggiano et al. (2009) who derive approximations for the channel fill rate and Dreyfuss and Posner (2015) who derive the window fill rate for a system with batch arrivals.

The second popular service measure is the expected backorder measure, which, by Little’s Law, is equivalent to the expected waiting time. The truncated waiting time incorporates customer patience into the expected waiting time. To the best of our knowledge, there is no research considering the truncated waiting time as a service measure. Therefore, while there are many papers who optimize the spares allocation according to the expected waiting time (e.g., Wong et al. 2006, Kranenburg and van Houtum 2009, Basten et al. 2012), our paper is novel in that it allows to optimize using the truncated waiting time.

In spite of the prevalent use of fill rate as a service measure, only few papers consider it as a criterion for the optimal spares allocation due to its non-concave form. The few papers who discuss the fill rate as a criterion for optimality, limit the search space to its concave region only (e.g., Sherbrooke 1968, Song and Yao 2002, Muckstadt 2005, Larsen and Thorstenson 2014, Basten and van Houtum 2014). Caggiano et al. (2007) is the only paper that attempts to optimize according to the generalization of the fill rate. Their model is a multi-item, multi-echelon model that allows different delivery times to different locations and items. We differ from their paper in a number of ways. First, we do not assume deterministic lead times. Instead, in our model the lead time or repair time is stochastic with a general distribution. Second, our window fill rate
can be computed for any time window and not only the lead times. Third, our optimization procedure uses exact values of the window fill rate and not approximations. Finally, Caggiano et al. (2007) assume that each location receives sufficiently many parts so that the fill rates are typically concave. In contrast, we do not limit the spares search space. In fact, we show that limiting the search space may decrease the fill rate, since it may be optimal that certain locations receive no spares at all.

Our finding that the window fill rate’s functional form is generally S-shaped ties our research to optimization models of additively separable S-shaped functions. A recent paper by Udell and Boyd (2013) uses a branch and bound algorithm to iteratively compute converging upper and lower bounds on the optimal solution. Our approach to maximizing the fill rate is a special case of Hackman and Platzman (1990) who develop an efficient algorithm to attain near-optimal solution to additively-separable objective function. This method requires maximizing a concave covering function of the original objective function. We exploit the fact that each window fill rate is either S-shaped or concave to compute the a priori bound to the distance from optimality.

3. The Model

We consider an exchangeable-item repair system with $L$ locations. A customer that has a failed item arrives to any one of these repair locations. Upon arrival, the item is sent for repair and once the failed item returns from repair it is added to the location’s stock. To reduce customer waiting time, the system keeps a number of spares so that if there is a spare item available on stock it is given immediately to the client in exchange of the arriving failed item. Customers are served according to a first-come, first-serve policy and leave the system once they receive an item.

For each location $l$, $l = 1, ..., L$, we assume that customer arrival rate follows an independent Poisson process with parameter $\lambda_l$. Each location has ample servers with i.i.d. repair time and therefore functions as an $M/G/\infty$ queue. The
ample server assumption implies that repair is carried out independently. Let \( R_l(t) \) denote the cumulative probability of an item to be repaired by time \( t \) and let \( r_l \) denote the mean repair time.

3.1. The Basic Service Measures

The expected waiting time is a popular performance measure in inventory systems. Following standard formulas (see Equation (2) in Sherbrooke (1968)), if location \( l \) has \( n \) spares then the expected number of backorders at a random point of time is \( \sum_{i=1}^{\infty} i \cdot \frac{(\lambda r_l)^{i+n}}{(i+n)!} e^{-\lambda r_l} \). By Little's Law the expected waiting time at the location is

\[
W_l(n) = \frac{1}{\lambda_l} \sum_{i=1}^{\infty} i \cdot \frac{(\lambda r_l)^{i+n}}{(i+n)!} e^{-\lambda r_l},
\]

which is a decreasing, convex function of the number of spares, \( n \). Let \( \vec{n} = (n_1, n_2, \ldots, n_L) \) denote a spares allocation and let \( \lambda = \sum_{l=1}^{L} \lambda_l \) denote the total customer arrival rate to the system. The expected waiting time of a random customer in the (entire) system, \( W(\vec{n}) \), is the weighted average of the local expected waiting times. Therefore, if the firm chooses the expected waiting time to be the criterion for optimality then its problem is

\[
\min_{\vec{n} \geq 0} W(\vec{n}) = \frac{1}{\lambda} \sum_{l=1}^{L} \lambda_l W_l(n_l) \quad \text{s.t.} \quad \sum_{l=1}^{L} n_l = N, \tag{1}
\]

where \( N \) is total number of spares. Since each local waiting time depends only on the number of spares in the location, the system’s waiting time is additively separable in each \( n_l \). Therefore, \( W(\vec{n}) \) is decreasing and convex in each \( n_l \) and the problem (1) can be solved efficiently using a greedy algorithm (see Sherbrooke 1968).

The waiting time measure is an appropriate criterion when the cost to the firm is proportional to the length of the customers’ wait. In many other situations, firms incur costs that are proportional to the number of customers who are waiting regardless of the length of their wait. The service measure that addresses this situation is the fill rate, i.e., the probability that a random cus-
tomer entering the system is served immediately and the firm’s problem now is to maximize the fill rate subject to the budget constraint.

3.2. Generalizing the Service Measures

The two service measures described above assume that the firm is penalized even when a customer waits a very small amount of time. In reality, however, this assumption is rarely true. Most service contracts state a period of time during which the firm is supposed to render the service to the customer. In this case, the penalty takes place only if the wait is longer than that stated time. Even without such contractual commitment, researchers find that customers have an “acceptable” or “reasonable” waiting time to be served (Durrand-Moreau, 1999). It is therefore necessary that the waiting time and the fill rate be generalized to consider only the customers who wait beyond the tolerable wait, \( t \). We denote this generalization by adding a time index to the expected waiting time and the fill rate functions. Accordingly, we let \( F_l(n,t) \) denote the probability that a random customer waits less than \( t \) units of time in location \( l \), with \( n \) spares. There is no agreed-upon term for \( F_l(n,t) \) in the literature. In Berg and Posner (1990) it is called the customer delay distribution, Song (1998) uses the term order fill rate, which is appropriate for the order to assembly systems that she discusses, Caggiano et al. (2007) uses the term channel fill rate and Sherbrooke (1968) simply refers to it as the fill rate. To distinguish it from the standard fill rate we use the term window fill rate \( F_l(n,t) \) as it measures the fill rate during a time window of \( t \).

Let \( \hat{Y}_l(t) \sim Skellam(\lambda_l \int_x^{\infty} (1 - R_l(x))dx, \lambda_l \int_0^t R_l(x)dx) \) (see Skellam 1946). That is, \( \hat{Y}_l(t) \) is the difference of two Poisson variables such that the positive Poisson has a parameter \( \lambda_l \int_x^{\infty} (1 - R_l(x))dx \) and the negative Poisson has a parameter \( \lambda_l \int_0^t R_l(x)dx \). The fill rate window for a single location, \( F_l(n,t) \), is derived in Berg and Posner (1990) equation (12) and is given by

\[
F_l(n,t) = Pr[\hat{Y}_l(t) \leq n - 1] + R_l(t)Pr[\hat{Y}_l(t) = n].
\]  

The window fill rate of the entire system, \( F(\vec{n},t) \) is the weighted average of the
local window fill rates. Therefore, if the firm’s costs are proportional to the number of customers waiting more than $t$, the firm’s objective is

$$\max_{\vec{n} \geq 0} F(\vec{n}, t) := \sum_{l=1}^{L} \frac{\lambda_l}{\Lambda} F_l(n_l, t) \quad \text{s.t.} \quad \sum_{l=1}^{L} n_l = N. \quad (3)$$

We characterize the shape of the window fill rate in the following proposition:

**Proposition 1.** $F_l(n, t)$ is strictly increasing in $n$. If $t \geq r_l$ then $F_l(n, t)$ is concave in $n$. Otherwise, $F_l(n, t)$ is either concave in $n$ or initially convex and then concave (S-shaped) in $n$.

Proofs are provided in the Appendix.

If each of the local window fill rates is concave, then the window fill rate $F(\vec{n}, t)$ is a sum of $L$ separable concave functions and a greedy algorithm can be used to solve the problem. By Proposition 1, this is certain to happen for a sufficiently large tolerable wait such that $t \geq \max_{0 \leq l \leq L} r_l$. In Section 4, we discuss the more challenging case in which there is at least one location that is S-shaped.

The argument that customers tolerate a certain amount of wait applies to the expected waiting time measure, too. Managers should focus attention only on disappointed customers who had to wait more than $t$ since only they require compensation. We further assume that the compensation to customers is only for the wait time beyond the permissable wait time $t$, i.e., if a customer waited $x$, he is compensated only for $\max\{x - t, 0\}$. We use the term *truncated waiting time* to denote the generalization of the expected waiting time and denote this by adding a time index to the expected waiting time function as in the following proposition:

**Proposition 2.** The truncated waiting time for location $l$ with $n$ spares, given by

$$W_l(n, t) = W_l(n) - \int_{0}^{t} (1 - F_l(n, x)) \, dx,$$

(4)

is a decreasing convex function of $n$. 

9
The truncated waiting time of the entire system, \( W(\vec{n}, t) \) is the weighted average of the local waiting times. Therefore, if the firm’s costs are proportional to the wait time that is greater than \( t \), the firm’s objective is

\[
\min_{\vec{n} \geq 0} W(\vec{n}, t) := \sum_{l=1}^{L} \frac{\lambda_l}{\lambda} W_l(n_l, t) \quad s.t. \quad \sum_{l=1}^{L} n_l = N. \quad (5)
\]

By Proposition 2, the generalized waiting time is a sum of \( L \) separable convex functions and therefore a greedy algorithm can be used to reach optimum efficiently.

The following remarks discuss modeling extensions and variations and are useful for application purposes.

**Remark 1.** For simplicity, we assume that the tolerable wait is identical for all locations. This assumption can be easily relaxed, without changing the analysis. In many realistic situations, customers in different locations have different expectation about the acceptable wait for service. Accordingly, instead of a generic tolerable wait, \( t \), a location-dependent, \( t_l \), should be used for location \( l \), and \( \vec{t} = (t_1, ..., t_L) \) for the entire system.

**Remark 2.** In our model, the number of spares, \( N \), is given exogenously. Alternatively, this constraint may be removed so that managers minimize total costs, which comprise the cost of spares and the reputation costs due to wait. Edogenizing the number of spares simplifies the analysis, since now each location may be solved independently of the other locations.

**Remark 3.** Ordinarily, repair systems handle multiple item-types. Adjusting the model to this requires defining the budget in term of dollars instead of spares. Let \( M \) denote the number of item-types and \( c_m, m = 1, ..., M \) denote the cost of each type. The spares allocation is a two dimensional vector, \( \vec{n} \), where element \( n_{m,l} \) is the number of spares of type \( m \) in location \( l \). The objective function of the optimization problem is now separable with each \( n_{m,l} \) and therefore all the results of the basic model with respect to the formulations and characteristics of the service measures are replicated. The budget constraint, however, cannot
be expressed in the number of spares, but instead, as a dollar value. That is, the budget constraint in (1), (3) and (5) is now replaced with

\[ \sum_{l=1}^{L} \sum_{m=1}^{M} c_{m} n_{m,l} \leq B, \]

where \( B \) is the dollar value of the spares budget. The greedy algorithms should be reformulated to reflect the change of the budget constraint. That is, in the greedy iteration we search for the duple \((m, l)\) for which an additional dollar value of spare has the best “bang for the buck” and allocate to location \( l \) an additional \( m \)-type item (subject to the feasibility constraint that the item’s cost is less than the residual budget). Since spares are indivisible, the greedy algorithm may result with an unused residual budget and is suboptimal. In many practical problems, however, the distance from optimality is negligible (see, for example, Dreyfuss and Giat 2016).

4. Optimizing the Window Fill Rate

If the criterion of optimality is a window fill rate with a sufficiently short tolerable wait, then there is at least one location with an S-shaped function. Indeed, researchers have pointed out that the fill rate is concave only when the number of spares in the location is sufficiently large, a property that makes it difficult to be used as a criterion of optimality. To overcome this, they limit the feasible region to the concave region. In what follows, we take a different approach to solve this problem directly without placing any constraints on the search space. We develop a near optimal algorithm that solves (3) using the following steps:

1. We define \( H(\vec{n}, t) \), a concave covering function of \( F(\vec{n}, t) \).
2. We maximize \( H(\vec{n}, t) \).
3. We show the relationship between the optimal solution of the covering function and the optimal solution of our problem (3).
The algorithm we describe is a special case of Hackman and Platzman (1990). Our problem involves a single constraint and the functions have a particular S-shaped functional form. The S-shape form of the objective function guarantees that \( H \) is initially linear and then strictly concave, and results with a specific structure of the optimal solution as described in Proposition 3. The single constraint allows us to simplify the second step to using a greedy algorithm.

For each S-shaped location, let \( m^*_l \), the tangent point of location \( l \), denote the first integer such that 
\[
\frac{F_l(m^*_l, t) - F_l(0, t)}{m^*_l} > \Delta F_l(m^*_l, t),
\]
where \( \Delta F_l(n, t) \) is the first difference of \( F_l(n, t) \) and is given by
\[
\Delta F_l(n, t) := F_l(n+1, t) - F_l(n, t) = (1 - R_l(t)) P_r[\hat{Y}_l(t) = n] + R_l(t) P_r[\hat{Y}_l(t) = n+1].
\]

The line connecting the points \((0, F_l(0, t))\) and \((m^*_l, F_l(m^*_l, t))\), (see Figure 1), the dashed line denoted by “tangent line”), is a cover of \( F_l(n, t) \) and its slope is steeper than \( \Delta F_l(m^*_l, t) \). For the concave locations, we set the tangent point to zero. We note, that to simplify exposure, we omit the dependency of \( m^*_l \) on \( t \). For each location, let \( H_l(n, t) \) denote the concave cover of \( F_l(n, t) \) in the

Figure 1: The shape of the window fill rate in a convex-concave location

![Shape of the window fill rate](image_url)
following manner:

\[
H_l(n, t) = \begin{cases} 
F_l(0, t) + \frac{F_l(m^*_l, t) - F_l(0, t)}{m_l} \cdot n & \text{if } 0 \leq n \leq m^*_l - 1 \\
F_l(n, t) & \text{if } n \geq m^*_l
\end{cases}
\]

That is, for any \( n \) smaller than the tangent point, we replace \( F_l(n, t) \) with the straight line connecting the point \((0, F_l(0, t))\) and the point \((m^*_l, F_l(m^*_l, t))\), the tangent line in Figure 1. By construction, for all \( n \geq 0 \), \( H_l(n, t) \) is concave and \( H_l(n, t) \geq F_l(n, t) \).

Let \( H(\vec{n}, t) \) denote the weighted sums of all the locations’ functions \( H_l(n, t) \) as in (3). Since for each location \( l \) and any allocation \( \vec{n} = n_1, ..., n_L \), \( H_l(n_l, t) \geq F_l(n_l, t) \), the optimal solution of \( H(\vec{n}, t) \) is an upper bound for the optimal solution of \( F(\vec{n}, t) \). Further, since each \( H_l(n, t) \) is concave, we have that \( H(\vec{n}, t) \) is also concave and therefore we can use a greedy algorithm to find its optimal solution. In the subsequent analysis, we let \( \vec{f}^* := (f^*_1, ..., f^*_L) \) denote the solution to the original problem (3), and let \( \vec{h}^* := (h^*_1, ..., h^*_L) \) denote the solution to \( \max \vec{h} H(\vec{h}) \text{ s.t. } \sum_{l=1}^L n_l = N \). Here, too, to simplify exposure we, omit the dependency of the optimal solutions on \( t \).

Since, by construction, \( H(\vec{n}, t) \) is concave in \( n \), we can use a greedy algorithm to find the optimal allocation for \( H(\vec{n}, t) \), \( \vec{h}^* \). The greedy algorithm sorts all the S-shaped locations according to the slope of their tangent lines (see Figure 1). Spares are allocated to the first sorted location (i.e., the location with the steepest slope) until it has reached its tangent point. The next sorted location receives units until it has reached its tangent point and so forth. To be precise, we note that it is possible that between switching from one location to the next, locations that have already reached their tangent point will receive additional spares. However, once a location starts receiving spares, it will be the only one to receive spares until it has reached its tangent point. The outcome of this process can be grouped into two cases as described in the following proposition.

**Proposition 3.** \( \vec{h}^* \) satisfies one of the following two cases:

1. For every \( l = 1 \ldots L \), either \( h^*_l \geq m^*_l \) or \( h^*_l = 0 \).
There exists a single location, denoted by \( \hat{l} \) such that \( 0 < h^*_l < m^*_l \). For every other location \( l \neq \hat{l} \), either \( h^*_l \geq m^*_l \) or \( h^*_l = 0 \).

Once an S-shaped location begins to receive spares, it will be the only recipient of spares until it has reached its tangent point. The second case of Proposition 3 therefore happens only if spares are depleted while allocating spares to location \( \hat{l} \) and before reaching its tangent point.

The consequence of Proposition 3 is that we have at most one location with an “intermediate” allocation of spares, i.e., a nonzero allocation that is smaller than the location’s tangent point. This observation is crucial for the characterization of the optimal solution of (3) described in the following proposition.

**Proposition 4.** If \( \tilde{h}^* \) satisfies case 1 of Proposition 3, then the optimal solution to (3), \( \tilde{f}^* = \tilde{h}^* \). If \( \tilde{h}^* \) satisfies case 2 of Proposition 3, then (a) the optimal value of \( F \) is bounded above by \( H(\tilde{h}^*, t) \), (b) the distance from optimum is bounded by \( \frac{\lambda}{\lambda_H}(H(I(h^*_l, t) - F(I(h^*_l, t))) \), and (c) \( 0 \leq f^*_l \leq m^*_l \).

The implication of Proposition 4 is that for large-scale problems with many locations, we can efficiently find the optimal solution (if we are in case 1) or at least be near it (case 2). In case 2, we do not necessarily have the optimal solution. To establish the a priori upper bound to the distance from optimality we need to search on \( n \) and find the maximal distance between \( H_l(n, t) \) and \( F_l(n, t) \).

Since \( H_l(n, t) \) is linear between zero and the tangent point, whereas \( F_l(n, t) \) is initially strictly convex and the concave, it can be easily shown that the maximal distance is attained at a point that is less or equal to the inflection point, \( n^*_l \) (see Figure 1). The search for this point can be further expedited using numerical techniques exploiting the fact that \( \Delta F_l(n, t) \) is strictly unimodal. Among these maximal distances, we need to choose the location with the greatest maximum since only one location at most will in fact have a different value for \( H_l \) and \( F_l \). Formally, the a priori upper bound for the distance from optimality is:

\[
\max_{0 \leq l \leq L} \max_{0 < n \leq n^*_l} \frac{\lambda}{\lambda_H}(H_l(n, t) - F_l(n, t)).
\]
The a posteriori maximal distance from optimality is the difference between the bounds, \( H(h^*, t) - F(h^*, t) = \frac{1}{N}(H(h^*_l, t) - F(h^*_l, t)) \), which is the difference of the functions’ values for location \( \hat{l} \) only. The implication of this result is that the bigger the problem in terms of number of locations, both the a priori and a posteriori bounds are smaller. In fact, we can further decrease the a posteriori distance between the bounds by executing improvement procedures that examine whether decreasing the number of units in location \( \hat{l} \) and allocating them to other locations increases the window fill rate. Since in a typical practical problem \( h^*_l \ll N \), the computation time of these procedures is negligible.

By statement (c) of Proposition 4, to reach optimality requires also examining all the possibilities of increasing the number of spares in location \( \hat{l} \) until its tangent point. Unfortunately, in the general case this requires examining all the possible permutations involving all \( N \) units. While from a theoretic standpoint this observation restricts attaining optimality in polynomial time, from a practical standpoint its effect is negligible.

4.1. The Optimal Solution

The greedy algorithm minimizing the truncated waiting time tends to distribute spares among all the locations. This is not generally true when maximizing the window fill rate. For simplicity, assume that the tolerable wait is sufficiently small so that tangent point of all the locations is strictly positive. If spares are rare or expensive so that the number of spares is small, then it is optimal to pool those spares in few locations instead of distributing them among the entire system. From a practical managerial perspective, this result is counter-intuitive as managers are usually resistant to neglecting part of the system. Only if the number of spares is sufficiently high so that all the location are beyond their tangent point then the solution tends to distribute spares among the entire system.

For example, consider a simple symmetric system comprising fifty identical locations, each with a tangent point of five units. If there are one hundred spares then maximizing the window fill rate is attained when only twenty loca-
tions receive spares (five spares for each location) and the other eighty locations receive none. In contrast, minimizing the truncated waiting time is achieved by distributing the spares equally to all the locations (two spares for each location). In the following section, we numerically illustrate this intuition and provide additional practical implications of the optimal solution.

5. Numerical Example

The following numerical example is motivated by the failed attempt to establish a battery-swapping system for electric vehicles in Israel. This large-scale example demonstrates the effects of customer patience and illuminates how the optimal solutions for the window fill rate and the truncated waiting time are profoundly different. For a detailed modeling of the battery-swapping problem we refer the reader to Avci et al. 2014.

The parameters of the example are: $L = 200$ stations; $N = 5000$ spare units; $\lambda_l = 10 + 0.25 \cdot l$ customers per hour, $R \sim N(45, 10^2)$ minutes.

The criteria for optimality are the truncated waiting time and the window fill rate. We optimize these measures for tolerable waits of zero ($W_{FR}$), ten ($W_{10}, F_{10}$) and fifteen ($W_{15}, F_{15}$) minutes. We further assume that the “correct” tolerable time is ten minutes. Therefore, for example, if managers allocated spares to maximize the fill rate, the firm’s true costs are nevertheless measured according to the window fill rate for ten minutes.

Table 1 details performance statistics for each of the six optimality criteria. The performance statistics are the same measures that we use for the optimality criteria, i.e., the truncated waiting time and window fill rate for zero, ten and fifteen minutes. We see that different criteria lead to significantly different optimal values of the objective function. In what follows, we also show that the optimal spares allocations dramatically differs.

5.1. Optimizing the Window Fill Rate

When service is measured according to the number of customers who recharge their batteries within ten minutes, $F_{10}$ should be optimized. Therefore, if the
Table 1: Performance Statistics for each of the Optimality Criteria

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$W$</th>
<th>$W_{10}$</th>
<th>$W_{15}$</th>
<th>$FR$</th>
<th>$F_{10}$</th>
<th>$F_{15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Criterion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$W$</td>
<td>4.649</td>
<td>0.710</td>
<td>0.171</td>
<td>36.97%</td>
<td>82.64%</td>
<td>94.39%</td>
</tr>
<tr>
<td>$W_{10}$</td>
<td>4.743</td>
<td>0.644</td>
<td>0.111</td>
<td>35.37%</td>
<td>82.10%</td>
<td>95.02%</td>
</tr>
<tr>
<td>$W_{15}$</td>
<td>4.876</td>
<td>0.662</td>
<td>0.104</td>
<td>34.67%</td>
<td>81.17%</td>
<td>94.90%</td>
</tr>
<tr>
<td><strong>Optimality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$FR$</td>
<td>11.655</td>
<td>8.945</td>
<td>7.667</td>
<td>69.48%$^a$</td>
<td>74.32%</td>
<td>74.44%</td>
</tr>
<tr>
<td>$F_{10}$</td>
<td>5.940</td>
<td>2.886</td>
<td>2.321</td>
<td>49.14%</td>
<td>85.29%$^b$</td>
<td>90.87%</td>
</tr>
<tr>
<td>$F_{15}$</td>
<td>4.757</td>
<td>0.645</td>
<td>0.110</td>
<td>35.30%</td>
<td>82.00%</td>
<td>95.02%$^c$</td>
</tr>
</tbody>
</table>

Notes: The table displays the performance statistics when we optimize according to each of the criteria. $W$, $W_{10}$ and $W_{15}$ denote the truncated waiting time for 0, 10 and 15 minutes, respectively. $FR$, $F_{10}$ and $F_{15}$ denote the window fill rate for 0, 10 and 15 minutes, respectively. $^a$ The upper bound is displayed. The distance between the bounds is less than $4.4 \times 10^{-5}$%. $^b$ The upper bound is displayed. The distance between the bounds is less than 0.046%. $^c$ The optimal solution is displayed.
firm incorrectly implements the FR criterion, it will have only 74.32% of its customers receiving service within ten minutes (see the F10 column in Table 1). However, optimizing according to the window fill rate for ten minutes, increases this measure to 85.29%. The number of customers per day is $\sum \lambda_l \cdot 60 \cdot 12 = 84,300$. Therefore, an increase of 85.29% - 74.32% = 10.97% is translated into 9,248 fewer frustrated customers every day!

In Figure 2 we compare the spares allocation for four optimality criteria $(W, FR, F10, F15)$. The optimal spares allocation is starkly different for each of the optimization criteria. For sufficiently small tolerable wait times (i.e., $FR, F10$), locations with few arrivals are not allocated any spares because of the window fill rate’s tendency to initially pool spares rather than distribute them among all locations. As the tolerable wait increases, more locations receive spares because the relative advantage of spares decreases with $t$ and therefore spares can be moved from the high arrival rate locations to other locations.
<table>
<thead>
<tr>
<th>Location</th>
<th>Number of Spares</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>60</td>
</tr>
</tbody>
</table>

In Figure 3, we show how the number of spares, \( N \), affects the optimal allocation when the firm optimizes \( F_{10} \). When there are only 2000 spares, only the top fifty locations receive spares. Increasing the number of spares has two effects. First, more locations receive spares. Second, locations that previously received spares now receive slightly more. If there are sufficiently many spares (e.g., \( N = 6000 \)), all the locations receive spares.

The low-arrival locations receive no spares and therefore have a very low window fill rate. For these locations, it is recommended that the firm does not commit to a ten minute exchange time. To compensate customers for the longer wait, the system’s managers could offer customers, for example, discounted snacks or drinks. Using behavioral research about customer waiting experience (e.g., Maister 1985), managers may provide customers incentives to agree to longer than usual waiting times.
5.2. Optimizing the Truncated Waiting Time

If managers estimate that the firm’s losses are proportional to the \textit{wait time} they should minimize $W_{10}$. According to Table 1, optimizing the truncated waiting time instead of the expected waiting time decreases the truncated waiting time, $W_{10}$, by $0.710 - 0.644 = 0.066$ minutes, a 9.3% improvement. Since the system serves 84,300 customers daily, this difference translates to an annual savings of more than 33,800 hours.

In Figure 4, we compare the optimal spares allocation for different values of $t$ when the firm’s losses are proportional to the wait time. The criteria for optimality that we examine are the truncated waiting time for zero, ten and fifteen minutes. The slope of the graph decreases as $t$ increases. This implies that for a higher value of $t$, the importance of the high rate locations decreases compared to the low rate locations. This result replicates the findings described when the window fill rates are the criteria for optimality.

Changes in the tolerable wait affect the fill rate and the waiting time optimizations differently. With the fill rate, the time effect is initially profound. Suppose the amount of spares in the budget is such that we can have only few of the locations at their tangent points. Therefore, when the tolerable wait is zero, many of the locations will have zero spares. Increasing $t$ changes the graph dramatically since more and more locations receive spares. Once $t$ is sufficiently large, an increase in $t$ results with only a subtle decrease in the graph’s slope. In contrast to the window fill rate, with the truncated waiting time, a change in $t$ has only the gradual effect of decreasing the slope of the spares allocation. From a practical perspective, these observations dictate that managers should invest more in estimating correctly the customers’ tolerance when fill rate is the objective compared to when the expected waiting time is the optimality criterion.
Figure 4: The optimal spares allocation for different optimality criteria when the spares budget is $N = 5000$ batteries.

$W$, $W_{10}$ and $W_{15}$ are the truncated waiting time for 0, 10 and 15 minutes, respectively.

6. Conclusions

In this paper, we propose that managers take into account the tolerable wait time and use the window fill rate or the truncated waiting time as the criterion for optimization in lieu of the fill rate and the expected waiting time. For each of the proposed performance measures, we derive the objective function for the spares allocation problem in an exchangeable-item, multiple-location repair system with ample servers. We characterize the window fill rate’s functional form and use it to show that the truncated waiting time is a convex function of the number of spares in each location. Therefore, if the firm’s costs are proportional to the customers’ wait beyond the tolerable wait, a greedy algorithm may be used to find the optimal allocation.

If, in contrast, the costs of the firm are proportional to the number of customers who wait more than the tolerable wait, then the window fill rate should serve as the objective function. If customers’ patience is high, i.e., they tolerate a sufficiently long wait, the window fill rate is concave and a greedy algorithm...
maximizes the window fill rate. Otherwise, the window fill rate is initially con-

vex and then concave. We define a concave covering function for which the
optimal solution can be computed efficiently. We use this solution to find lower
and upper bounds to the optimal objective function and characterize when the
optimal solution for the covering function is also the optimal solution to the
original problem.

We complement our theoretical analysis with a numerical example motivated
by the attempt to establish a battery-swapping network for electric vehicles in
Israel. Similar attempts are considered by other companies in the electric car
industry and in other industries as well. In addition to the insights that are
gleaned from the numerical example, it demonstrates that the model can be
easily implemented to solve the spares allocation problem even in very large-

scale problems.

The system we consider is structured as a single echelon problem. One may
consider extending the model to a multi-echelon repair system. Such systems
are quite common and allow firms to save resources so that lower echelon loca-
tions handle only relatively simple repair jobs whereas the higher echelons are
equipped to handle complex repairs. Taking into account customer patience by
applying the window fill rate or the truncated waiting time will help managers
decrease costs and improve customer satisfaction.

Appendix

Proof of Proposition 1: The first difference of \( F_l(n, t) \) is given by (6). Recall, \( R_l(t) \) is the cumulative repair time distribution and therefore, \( R_l(t) > 0 \)
and \( 1 - R_l(t) > 0 \). Consequently, (6) is strictly positive for all \( n \) and therefore
the first statement of the proposition is true. The second difference is given by

\[
\Delta^2 F_l(n, t) := \Delta F_l(n + 1, t) - \Delta F_l(n, t) = \\
(1 - R_l(t))(Pr[\hat{Y}_l(t) = n+1] - Pr[\hat{Y}_l(t) = n]) \\
+ R_l(t)(Pr[\hat{Y}_l(t) = n+2] - Pr[\hat{Y}_l(t) = n+1]). \tag{7}
\]
We begin by showing $\Delta^2 F_l(n, t)$ is either negative for all $n \geq 0$ or initially positive and then negative. The Skellam distribution is strongly unimodal (see Karlis and Ntzoufras 2006), and therefore it is initially increasing and then decreasing. Let $\hat{n}^*_t(t)$ denote the first point for which $Pr[\hat{Y}_l(t) = n + 1] - Pr[\hat{Y}_l(t) = n] \leq 0$. Assume for now that $\hat{n}^*_t(t) \geq 2$. By definition, $Pr[\hat{Y}_l(t) = \hat{n}^*_t(t) + 1] - Pr[\hat{Y}_l(t) = \hat{n}^*_t(t)] \leq 0$ and $Pr[\hat{Y}_l(t) = \hat{n}^*_t] - Pr[\hat{Y}_l(t) = \hat{n}^*_t(t) - 1] > 0$. Therefore, by (7), $\Delta^2 F_l(\hat{n}^*_t(t), t) \leq 0$ (it is a sum of a non-positive and a strictly negative) and $\Delta^2 F_l(\hat{n}^*_t(t) - 2, t) > 0$ (because both components are strictly positive) and therefore $\Delta^2 F_l(n, t)$ is initially positive and then negative. If $\hat{n}^*_t = 0$ then $\Delta^2 F_l(n, t)$ is non-positive for $n = 0$ and strictly negative for all $n > 0$. Finally, if $\hat{n}^*_t(t) = 1$, then $\Delta^2 F_l(n, t)$ is positive for $n = 0$, non-positive for $n = 1$ and negative for all $n > 1$. It is now left to show that if $r_l \geq t$ then $\Delta^2 F_l(n, t) < 0$ for all $n$. Let $x_l(t) = \lambda_l \int_{x=t}^{\infty} (1 - R_l(x)) dx$ and $y_l(t) = \lambda_l \int_{x=0}^{t} R_l(x) dx$. By definition, $\hat{Y}_l(t) \sim Skellam(x_l(t), y_l(t))$. The mean of $\hat{Y}_l(t)$ is the difference of the means of the two Poisson random variables that compose $\hat{Y}_l(t)$, and is given by

$$x_l(t) - y_l(t) = \lambda_l \int_{x=t}^{\infty} (1 - R_l(x)) dx - \lambda_l \int_{x=0}^{t} R_l(x) dx$$

$$= \lambda_l \left( \int_{x=0}^{\infty} (1 - R_l(x)) dx - \int_{x=0}^{t} (1 - R_l(x)) dx + \int_{x=0}^{t} R_l(x) dx \right) = \lambda_l (r_l - t).$$

In the above, $r_l$ is the mean repair time and therefore $r_l = \int_{x=0}^{\infty} (1 - R_l(x)) dx$. When $r_l = t$, then $x_l(t) = y_l(t)$ and the Skellam distribution is symmetric (Karlis and Ntzoufras 2006). When $t > r_l$ then $x_l(t) < y_l(t)$ and the distribution of $\hat{Y}_l(t)$ is shifted to the left. Therefore, when $r_l \leq t$ for all $n \geq 0$, $Pr[\hat{Y}_l(t) = n + 1] - Pr[\hat{Y}_l(t) = n] < 0$ and both components of the right hand side of (7) are negative. Thus, $\Delta^2 F_l(n, t) < 0$ and $F_l(n, t)$ is strictly concave. □

The following lemma is needed to prove Proposition 2.

**Lemma 1.** The inflection point, $n^*(t)$, is nonincreasing with $t$.

**Proof of Lemma 1.** In the proof we use the notation in the proof of Proposition 1. Recall, $\hat{n}^*_t(t)$ is the first point for which $Pr[\hat{Y}_l(t) = n + 1] - Pr[\hat{Y}_l(t) =$
When \( t \) is increasing, \( x(t) \) is decreasing and \( y(t) \) is increasing. Therefore, by the properties of the Skellam distribution (see Karlis and Ntzoufras (2006)), the distribution of \( \hat{Y}_i(t) \) is shifted to the left and the mode of the distribution is nonincreasing, i.e., \( \hat{n}_1^* \) is nonincreasing with \( t \). We now show that the inflection point, \( n_1^* \), is either \( \hat{n}_1^* \) or \( \hat{n}_1^* - 1 \). If \( \hat{n}_1^* \geq 2 \) then by (7), \( \Delta^2 F_l(\hat{n}_1^*) \leq 0 \) and \( \Delta^2 F_l(\hat{n}_1^* - 2) > 0 \). If \( \Delta^2 F_l(\hat{n}_1^* - 1) > 0 \) then \( n_1^* = \hat{n}_1^* \), otherwise \( n_1^* = \hat{n}_1^* - 1 \).

Now, let \( t_1 > t_0 \). We need to show that \( n_1^*(t_1) \leq n_1^*(t_0) \), which requires showing that \( \Delta^2 F_l(n_1^*(t_0), t_1) \) is nonnegative and the second component is positive. Assume therefore that \( n_1^*(t_0) = \hat{n}_1^*(t_0) - 1 \). Now, the first component of \( \Delta^2 F_l(n_1^*(t_0), t_0) \), is negative and the second component is positive. However, by changing the time to \( t_1 \), \( R_l(t_1) \geq R_l(t_0) \) and so the negative component is smaller (in absolute terms) and the positive component is greater and thus we have that \( \Delta^2 F_l(n_1^*(t_0), t_1) > \Delta^2 F_l(n_1^*(t_0), t_0) > 0 \).

**Proof of Proposition 2.** Let \( f_W(n, t) = \frac{\partial W(n, t)}{\partial x} \). In words, \( f_W(n, t) \) is the probability density function of the customer’s waiting time. By definition, the truncated waiting time is

\[
W_i(n, t) = \int_0^\infty \max\{x-t, 0\} f_W(n, x) dx = \int_t^\infty (x-t) f_W(n, x) dx = \int_t^\infty (1-F_l(n, x)) dx.
\]

(8)

The last equality in the above is attained by integration by parts. The expected waiting time

\[
W_i(n) = W_i(n, 0) = \int_0^\infty x f_W(n, x) dx = \int_0^\infty (1-F_l(n, x)) dx
\]

(9)

Substituting (9) in (8) gives (7).

By (8), the first difference of \( W_i(n, t) \) is

\[
\Delta W_i(n, t) := W_i(n+1, t) - W_i(n, t) = -\int_t^\infty \Delta F_l(n, x) dx.
\]
By Proposition 1, $\Delta F_l(n, t) > 0$ for all $n$ and therefore $\Delta W_l(n, t) < 0$ for all $n$ and therefore $W_l(n, t)$ is decreasing in $n$. The second difference of $W_l(n, t)$ is

$$\Delta^2 W_l(n, t) := \Delta W_l(n + 1, t) - \Delta W_l(n, t) = - \int_t^\infty \Delta^2 F_l(n, x)dx. \quad (10)$$

to show convexity we need to show that $\Delta^2 W(n, t) > 0$. By Proposition 1, $\Delta^2 F_l(n, t)$ is either negative for all $n$ or initially positive (for lower $n$-values) and then negative. Specifically, for any tolerable wait $x$, if $n < n^*(t)$ then $\Delta^2 F_l(n, x) \geq 0$ and if $n > n^*(t)$ then $\Delta^2 F_l(n, x) < 0$. If $n > n^*(t)$, then by Lemma 1 for all $x \geq t$, $n > n^*(t)$ and therefore for all $x \leq t$ we have that $\Delta^2 F_l(n, x) \geq 0$. Thus, $\int_t^\infty \Delta^2 F_l(n, x)dx \geq 0$. Therefore, $\Delta^2 F_l(n, x) = - \int_t^\infty \Delta^2 F_l(n, x)dx - \int_0^\infty \Delta^2 F_l(n, x)dx = \int_0^\infty \Delta^2 F_l(n, x)dx = \Delta^2 F_l(n, 0) > 0$.

In the above, the strict inequality is true since $W(n)$ is convex. □

Proof of Proposition 3: The proposition is a summary of the observations proceeding it. □

The following lemma is needed to prove Proposition 1.

Lemma 2. $H(\tilde{h}^*) \geq F(\hat{f}^*)$.

Proof of Lemma 2: Since $\tilde{h}^*$ is optimal for $H$, $H(\tilde{h}^*) \geq H(\hat{f}^*)$. $H$ is a covering function of $F$ and therefore, in particular, $H(\hat{f}^*) \geq F(\hat{f}^*)$. Thus, $H(\tilde{h}^*) \geq F(\hat{f}^*)$. □

Proof of Proposition 4: By construction, for every location, $H_l(0, t) = F_l(0, t)$ and $H_l(n, t) = F_l(n, t)$ for all $n \geq m_l^*$. Therefore, if $\tilde{h}^*$ satisfies case 1 then $H(\tilde{h}^*, t) = F(\tilde{h}^*, t)$ and by Lemma 2 we have that $F(\tilde{h}^*, t) \geq F(\hat{f}^*, t)$. This immediately implies that $\tilde{h}^*$ is the optimal solution to (3), that is, $\hat{f}^* = \tilde{h}^*$. If $\tilde{h}^*$ satisfies case 2, then by Lemma 2 $H(\tilde{h}^*, t)$ is an upper bound. The
distance from optimality is $H(\vec{h}^*, t) - F(\vec{f}^*, t)$. Since for every location $l \neq \hat{l}$, $H_l(h_l^*, t) = F_l(f_l^*, t)$, this distance reduces to $\frac{\lambda}{N}(H_l(h_l^*, t) - F_l(h_l^*, t))$. 

It is left to show the last statement of the proposition. Let $N_1 = N + m_l^* - h_l^*$ (N1 is set so that the we have sufficiently many spares to bring the $\hat{l}$ location to its tangent point). Similar to the arguments preceding the proof of Proposition 3, the optimal solution to the problem with $N_1$ spares is identical to the current solution with the exception that location $\hat{l}$ now has $m_l^*$ spares. This solution is optimal for $F(\vec{n}, t)$ and for $H(\vec{n}, t)$ as it matches case 1 of Proposition 3. Now, assume in contrast, that for the original problem with $N$ spares, $f_l^* > m_l^*$. We will show that this contradicts the optimal solution for $N_1$ spares. Since $f_l^* > m_l^*$ there is at least one location $l^*$ that surrendered a spare to location $\hat{l}$, which implies that the improvement $F_l(m_l^* + 1, t) - F_l(m_l^*, t)$ is greater than the contribution attained by the an additional spare to location $l^*$. However, if that is true, then when there are $N_1$ spares available it is impossible for location $\hat{l}$ to have only $m_l^*$ since by transferring a spare from $l^*$ to location $\hat{l}$ we can increase $F$. □

References


