From Low-Cost Airline Baggage to By-Minute Car Rental, the Impact of Penalty Cost on Customers’ Booking Decisions

Daniel Zhuoyu Long\textsuperscript{1} • Rowan Wang\textsuperscript{2} • Chi Xie\textsuperscript{1}

\textsuperscript{1} Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong
\textsuperscript{2} Lee Kong Chian School of Business, Singapore Management University, Singapore 178899

zylong@se.cuhk.edu.hk • rowanwang@smu.edu.sg • cxie@se.cuhk.edu.hk

Abstract

This paper studies two problems that are both related to the impact of shortage penalty cost on customers’ booking decisions. The first problem looks at low-cost airline passengers’ baggage allowance purchasing decisions. The second problem looks at by-minute car rental customers’ rental booking decisions. Both of the problems are newsvendor type of problems. However, from practical observations, customers’ baggage booking allowance (rental booking time duration) does not monotonically increase with the baggage overweight (return delay) penalty cost (i.e., the shortage cost), which is unexpected for newsvendor systems. In this paper, we explain this interesting phenomenon by looking into the influence of shortage penalty cost on customers’ baggage (trip) planning and weight (time) estimation, which in turn, affects the actual baggage weight (usage time) distribution. We then incorporate this into stochastic decision models and analytically characterize how important quantities (e.g., customers’ booking allowance, customers’ total cost, and firm’s profit) change with the penalty cost. We prove that both the customers’ booking allowance and the firm’s profit are quasiconcave in the penalty cost. For the by-minute car rental problem, we also study the optimal delay penalty cost for the rental firm to maximize its profit. We show that, when a rental car is shared by more customers, a higher delay penalty cost is preferred.

Keywords: baggage allowance; car rental; shortage penalty cost; revenue management; sharing economy; optimal policy
1 Introduction

Our study is motivated by an interesting general phenomenon observed from two different case studies. The first case study is on passengers’ purchasing decision for baggage allowance at a major low-cost airline in Southeast Asia. As a low-cost carrier, the airline charges its passengers for checked baggage. There are two ways for passengers to purchase baggage allowance, booking online while booking the flight; or purchasing at the airport while checking in the flight. The price rate for purchasing at the airport ($p$) is higher than that for booking online ($b$). In the case that a passenger has booked baggage allowance online, but the actual weight of the checked baggage is higher than the booked allowance, the extra portion will be charged at the airport purchasing rate $p$. Thus, we can consider $p$ as the overweight penalty cost. Passengers usually book flights and baggages weeks or months ahead of the travel dates. Due to the uncertainty in weather conditions and trip plans, at the time of flight booking, the actual baggage weight needed is unknown and can be considered as a random variable (see detailed explanation later in the section). Passengers need to choose a baggage booking allowance that minimizes the total baggage cost (the sum of the online booking cost and the overweight penalty cost), facing the uncertainty in the actual baggage weight needed.

It is easy to see that, the baggage allowance booking problem is indeed a *newsvendor* type of problem, with $b$ as overage cost; $p - b$ as shortage cost; and the actual baggage weight needed as random demand. As we all know, for classic newsvendor problems, the optimal ordering quantity should always increase with the shortage cost. Therefore, with fixed $b$, our client airline expects higher baggage booking allowance from passengers, when the overweight penalty cost $p$ is set higher. However, unexpectedly, the real observation is that, passengers’ baggage booking allowance decreases with $p$ for some region of $p$.

Similar customer behavior phenomenon is observed in another case study on customers’ booking decision at a by-minute car rental company in the same geographical region. To use the by-minute rental car, customers need to book the car for an exact time duration, at the booking price rate ($b$). Customers are expected to return the car before the end of the booking time period. In the case that customers delay the return of the car, a delay cost will be charged at the penalty cost rate ($p$). $p$ is higher than $b$. Due to the uncertainty in traffic conditions and route plans, at the time of rental booking, the actual usage time duration needed to complete the trip is unknown and can be considered as a random variable (see detailed explanation later in the section). Customers need to choose a booking time duration that minimizes the total trip cost (the sum of the booking cost and the delay cost), facing the uncertainty in the actual usage
time duration needed. This is again a newsvendor type of problem. Similar to what happens in
the airline baggage case, from the rental firm’s observation, customers’ booking time duration
decreases with the penalty cost \( p \) for some region of \( p \).

The two similar observations above draw great attention from us. Why passengers’ baggage
booking allowance (booking time duration) does not monotonically increase with the penalty
cost? What makes the customers’ behavior different from the newsvendor prediction? How
does the overweight quantity (delay duration) change with the penalty cost? How these affect
the airline’s (rental firm’s) profit?

To understand these interesting and unexpected phenomena, we conducted interviews with
a selected group of our client airline’s passengers who have booked multiple flights in the past
under different overweight penalty costs and their baggage booking records exhibit the non-
monotonic pattern. A major finding from the interviews is that, the overweight penalty cost
influences customers’ estimation on the baggage weight needed, which in turn affects the \( \text{variance} \) of the distribution of the actual baggage weight. When \( p \) gets higher, passengers will
be more serious on baggage planning and weight estimation, and the variance of the actual
baggage weight distribution decreases (see detailed explanation later in the section). Thus,
different from the classic newsvendor type of problems where the demand distribution is fixed
and the optimal solution monotonically increases with the shortage cost, our baggage allowance
booking problem has the unique feature that the variance of the “demand” distribution depends
on the shortage cost, and therefore the optimal solution may not monotonically increase with
the shortage cost. The same logic applies to the car rental booking problem.

In this paper, we first build a stochastic decision model to study the low-cost airline baggage
problem. We solve an individual passenger’s optimization problem to minimize her expected
total baggage cost. Given overweight penalty cost \( p \), we derive explicit expressions for the
optimal booking allowance, the corresponding overweight quantity, and the corresponding total
baggage cost (airline’s profit). We prove that the optimal booking allowance is quasiconcave
in \( p \); the corresponding expected overweight quantity is decreasing in \( p \); and the expected
airline’s profit is quasiconcave in \( p \). We then build another model to study the by-minute
car rental problem. We consider the settings where the delayed car return from a previous
customer would affect the car usage of a following customer, and the rental firm needs to pay
a delay compensation to the affected customer. This interdependence among customers makes
the system much more complex to analyze. We provide analytical formulations to compute the
delay compensation and the rental firm’s profit for the system with two customers. We show that
the expected delay compensation associated with the optimal booking duration is decreasing in \( p \); and the expected rental firm’s profit is quasiconcave in \( p \). We compare a system consisting of only one customer with a system consisting of two customers, and show that, when the penalty cost is low, having an additional customer could in fact bring negative profit to the firm. We then prove that, for the rental firm to maximize its profit, the optimal choice of penalty cost for the two-customer system should be higher than that for the one-customer system. Finally, we numerically validate the robustness of the above insights for systems with more than two customers.

In the following, we discuss the low-cost airline baggage case and the by-minute car rental case in detail and explain the settings of our study.

**Low-Cost Airline Baggage Case in Detail**

**Why the actual baggage weight needed is uncertain at the time of booking?** From our client airline’s internal marketing research reports (not available to public), the majority of their passengers are non-business travelers (e.g., casual tourists or students), and can be considered as cost savers. This group of passengers exhibit certain features on their flight and baggage booking behaviors. (1) The passengers generally book flights weeks or months ahead of the travel dates, as the airfare is cheaper for early bookings. (2) The passengers generally book baggage allowance online, as the price for booking online is lower than that for purchasing at airport. (3) The passengers generally do not book for very high baggage allowance (e.g., more than 30 kg) online, which is an expected behavior from cost savers. (4) Baggage overweight issues quite often happen to the passengers. The major reason is that, since the flight and baggage are booked a long time ahead of the travel dates, there is high level uncertainty in the goods to carry and the baggage weight. For example, tourists usually need to bring multiple pieces of different clothes and shoes, and the selection depends heavily on the travel duration, activities, and weather conditions. According to our client airline’s customer survey results, while booking the flight months before the travel dates, a large portion of the passengers have not yet decided the total travel duration. (Many flight bookings at low-cost airlines in Southeast Asia are for one-way flights, since the round-trip fare is just equal to the sum of the two separated one-way fares, and discounts are given to selected flights from time to time. This is different from the pricing strategy at most of the full service airlines, where the round-trip fare is usually cheaper than the sum of the two separated one-way fares.) Many of our client airline’s passengers travel to multiple destinations before returning home. After booking the flights, passengers
read information and travel guides about the destinations, and sports or social activities (e.g., hiking or concert) could be added to the trip plan from time to time. Many tourist destinations (e.g., island or mountain area) have unstable weather conditions, and passengers may prefer to pack clothes only after receiving accurate weather forecasting right before the departure dates. For all the above reasons, as we can see, there is high level uncertainty in the goods to carry and the baggage weight. Thus, at the time of booking, the actual baggage weight needed is unknown and can be considered as a random variable.

Why the airport purchasing price \( p \) (instead of the online booking price \( b \)) is considered by the airline as an important decision variable? For a regional low-cost airline, baggage fee contributes to an important portion of its revenue. The airline wants to maximize its baggage revenue from each passenger, which is the sum of the revenue from online booking and the revenue from airport purchasing (overweight penalty). Among the two baggage prices, the online booking price \( b \) and the airport purchasing price \( p \), the airline quite often changes the later one but not the earlier one. In fact, among the few leading low-cost carriers in the region, the baggage online booking prices are almost the same, but the baggage airport purchasing prices can be quite different. The reason is that, the baggage online booking fee is paid at the same time with flight booking, and the sum is viewed, by the passengers, as the airfare, to compare with that from full service airlines which offer free checked baggage allowance. Therefore, customers are very sensitive to the change of \( b \). From our client airline’s past practice, when \( b \) increases, the flight booking demand drops remarkably. On the other hand, the baggage airport purchasing fee (the overweight penalty cost) is not paid at the time of flight booking, and it can be viewed as an indirect cost to passengers. From our client airline’s past practice, the change in \( p \) has relatively low impact on the flight booking demand. Of course, when \( p \) gets extremely high, say when \( p > \hat{p} \) for a very large \( \hat{p} \), baggage overweight will become too risky for passengers, and passengers will stop booking the flights. The value of \( \hat{p} \) can be observed and verified from flight booking data. As a result, our client airline wants to maximize its baggage revenue through an optimal choice of \( p \), but keeps \( b \) at the market price.

Why the variance of the actual baggage weight distribution is affected by the overweight penalty cost? While booking the baggage allowance, passengers evaluate the potential cost associated with baggage overweight through the penalty cost. When the penalty cost is low, the potential cost of baggage overweight is limited, and customers will just roughly estimate the baggage weight needed. The resulting actual baggage weight distribution will have a high variance. On the other hand, when the penalty cost is high, the potential cost
of baggage overweight is significant, and customers will be much more serious on baggage planning and weight estimation. For example, passengers will search for more information on destination activities and weather conditions, and plan the clothes and shoes accordingly. Passengers will also decide whether certain things (e.g., book or camera) are needed for the trip, and consider where to put the items, in checked luggage or in carry-on. After detailed baggage planning, passengers will calculate (or estimate) the needed weight of checked baggage, and make the booking. Although randomness still exists, it is easy to understand that, the resulting baggage weight estimation will be much more accurate, and the associated actual baggage weight distribution will have a much lower variance.

By-Minute Car Rental Case in Detail

Why the actual usage time duration needed is uncertain at the time of booking?
An important observation from the rental vehicle GPS driving record at our client car rental company is that, a large portion of the rental demand is for relatively long-distance trips (e.g., more than 20 km) with multiple segments (i.e., the vehicle stops at multiple parking lots during a rental trip before it is returned). A plausible reason for this phenomenon is that, in big cities of Southeast Asia (where our client rental firm operates), taxi service is pervasive with reasonable price. For short-distance trips (e.g., less than 5 km), in general, it is cheaper and much more convenient to take a taxi than to rent a car. However, compare with taxi, rental car has two major advantages, among others. (1) Rental cars can be driven to suburb areas, while taxis can only be hired to travel within the city area. Taxi drivers generally reject the requests to go to suburb areas, since it is very difficult to get passengers on the way back to town. The only possibility is to hire the taxi for round trip, but the cost is usually very high. (2) Rental car drivers can store bags or luggages in the car during the whole trip, while taxi riders have to carry the items on hand in between two taxi rides. For example, a customer may want to visit several malls and grocery stores alone the way of a weekend shopping trip. If she drives the rental car, then she can put all the purchased items in the car while visiting the next store. On the other hand, if she takes taxi, then she has to carry all the purchased items on hand. Because of these two advantages, for long-distance trips with multiple subtasks, rental car could be preferred to taxi. Typical examples of this kind of trips include tourist trips to suburb areas and shopping trips to several stores. Usually, customers will not have serious and detailed plans for this kind of trips. Total trip durations could be flexible, and additional stops (tourist sites or malls) would be added along the trip. There are usually multiple route plans
to choose from, and the traffic conditions vary from time to time. For the above reasons, as we can see, there is high level uncertainty in the total trip duration of a rental. Thus, at the time of rental booking, the actual usage time duration needed is unknown and can be considered as a random variable.

Why the delay penalty cost $p$ is considered by the rental firm as an important decision variable, and why the variance of the actual usage time distribution is affected by the delay penalty cost? The explanations for these two questions are similar to those we have for the low-cost airline baggage case, and are therefore omitted for the sake of brevity.

**Related Literature**

Our work contributes to three streams of literature, revenue management, car rental (and sharing economy in general), and airline baggage policy.

First, both the low-cost airline baggage problem and the by-minute car rental problem can be viewed as a special newsvendor type of problem where the demand distribution depends on the shortage cost, and the shortage cost is considered as a decision variable for profit maximization. Although, to the best of our knowledge, the current work is the first one to study newsvendor type of problems with shortage cost dependent demand distribution, our paper is related to the revenue management literature, and particularly the literature of joint inventory-pricing control, in general. The incorporation of price effects in inventory problems can be traced back to Whitin (1955), who is the first one to study a retail setting where demand depends on selling price, and the firm decides the ordering quantity and the selling price simultaneously to maximize its profit. Since then, a substantial body of literature has extended this idea and studied the joint inventory-pricing problems in various settings. For some recent development, Pang et al. (2012) study the problem with positive order leadtime. Feng et al. (2014) analyze the system under the setting that the demand follows a generalized additive model. Chen et al. (2017) look at a supply chain model where a supplier prices a product and sells it to a multi-period newsvendor retailer. When the uncertain demand is only with limited distributional information, Adida and Perakis (2006) propose a robust optimization framework to search for the optimal policy.

Although prevalent in practice, the traditional car rental business has received relatively little attention in the operations management literature. Li and Pang (2017) recently model the operations in a single-station car rental firm using discrete-time Markov decision processes and propose solution procedures for the firm to decide when to accept or to reject customer orders. Different from a traditional rental car which can only be rented by one customer per
day, a by-minute rental car can be shared by multiple customers over the course of a day. This new operational model grows rapidly under the emerging business environment of sharing economy. Recently, there is a growing body of literature on sharing economy. Benjaafar et al. (2018) characterize the equilibrium for peer-to-peer product sharing systems where individuals can decide whether to own or to rent products. Taylor (2018) studies two main features of an on-demand sharing system, delay sensitivity and agent independence, and examines the impact of these two features on the price and wage for system users. Cachon et al. (2017) look at different pricing schemes and show the advantages of the widely used surge pricing policy. Hu and Zhou (2018) analyze the performance of the fixed commission contract which is used by many on-demand sharing platforms in practice. Feng et al. (2018) compare the new on-demand ride-hailing system with the traditional street-hailing system, and propose methods to improve the performance of the on-demand hailing system.

Our model and results on the low-cost airline baggage case add to the study of commercial airlines’ baggage policy. The majority of the literature on airlines’ operations management focuses on either the airplanes (fleet assignment) or the passengers (revenue management). Only a few papers study baggage related issues. Wong et al. (2009) look into the issue of how to optimally allocate the limited belly space of an airliner between passenger baggage and air cargo. Through a multi-item newsvendor model, the authors show that, to maximize profit, airlines should reduce passenger baggage allowance limit and reserve more space for air cargo on wide-body aircrafts. Fragniere et al. (2012) study airline passengers’ perceived value of the baggage allowance as well as other services associated with a commercial flight. The results suggest that higher baggage allowance leads to higher perceived value but also higher fuel cost. Nicolae et al. (2017) investigate the impact of baggage policies. From empirical analysis, the authors show the relationship between the free/charged baggage policy and the airlines’ on-time departure performance.

The rest of the paper is organized as follows. In Section 2, we analyze the low-cost airline baggage problem. In Section 3, we study the by-minute car rental problem. In Section 4, we provide concluding comments. Throughout the paper, “increase/decrease” means “nondecrease/nonincrease”.

2 Low-Cost Airline Baggage Problem

We consider an individual passenger who travels with a low-cost airline which charges for checked baggage. There are two ways for the passenger to purchase baggage allowance, booking online
while booking the flight; or purchasing at the airport while checking in the flight. The baggage price is \( b \) per unit weight for online booking, and \( p \) per unit weight for airport purchasing. \( p \geq b \).

(In reality, usually the baggage allowance is sold in the unit of 5 kgs. We consider a continuous model to simplify the analysis. Our insights also hold for discrete models.) Let \( x \) denote the passenger’s baggage booking allowance and assume \( x \in \mathbb{R} \). In the event that the actual weight of the checked baggage is higher than \( x \), the extra portion will be charged at the rate of \( p \) per unit weight. We can view \( p \) as the penalty cost for shortage in booking. In this section, we mainly study the impact of the penalty cost \( p \) on passengers’ booking decisions, baggage costs, and firm’s profit. The booking price \( b \) is considered as a constant.

Passengers usually book flights and baggages weeks or months ahead of the travel dates. Due to the uncertainty in weather conditions and trip plans, at the time of flight booking, the actual baggage weight needed is stochastic. To capture the impact of the penalty cost on passengers’ booking decisions, a main feature incorporated into our model is that, the penalty cost affects the distribution of the actual baggage weight. When the penalty cost increases, the variance of the actual baggage weight distribution decreases (see explanation in the introduction section).

Let \( X(p) \) denote the random variable that describes the actual baggage weight. We represent \( X(p) \) as

\[
X(p) = \alpha(p) \frac{1}{\mu} + [1 - \alpha(p)] \tilde{X},
\]

with

\[
\alpha(p) = 1 - e^{-\beta(p-b)},
\]

where \( \mu, \beta > 0 \) are constants, and \( \tilde{X} \) follows an exponential distribution with rate \( \mu \). We see the \( X(p) \) in Equation (1) consists of two parts. For the first part, \( \frac{1}{\mu} \) represents the deterministic component; and for the second part, \( \tilde{X} \) represents the stochastic component. \( \alpha(p) \) in Equation (2) is the weight function determining the weights of the deterministic component and the stochastic component. \( \beta \) is the exponential coefficient determining the impact of \( p \) on the weight function. \( \alpha(p) \) increases in \( p \). This means that, as \( p \) increases, the weight of the deterministic component increases and the weight of the stochastic component decreases. Let \( E(\tilde{X}) \) denote the expectation of \( \tilde{X} \). \( E(\tilde{X}) = \frac{1}{\mu} \), and therefore, \( E(X(p)) = \frac{1}{\mu} \). We see that \( p \) affects the variance of \( X(p) \) while keeping the mean constant. The variance of \( X(p) \) decreases while \( p \) increases. The interpretation of Equation (1) is that, as we mentioned in the introduction section, with higher overweight penalty cost, passengers are more sensitive to overweight, and therefore will be more serious on baggage planning and weight estimation. As a result, the
variance of the baggage weight decreases. We emphasize here that the function forms of \(X(p)\), \(\alpha(p)\), and \(\tilde{X}\) are chosen for the ease of deriving analytical results. The main results and insights remain valid in general for other function forms with the same key properties (variance of \(X(p)\) decreases in \(p\); \(\alpha(p)\) increases in \(p\); and \(E(\tilde{X}) = \frac{1}{\mu}\)).

In the following, we describe the passenger’s decision making process. A passenger needs to take travel baggage with mean baggage weight \(\frac{1}{\mu}\). She checks the penalty cost \(p\) and plans the baggage. With higher \(p\), the baggage will be better planned, and the variance of the baggage weight will be lower. Then, based on the booking price \(b\), the penalty cost \(p\), and the distribution of the baggage weight \(X(p)\), the customer decides the booking allowance \(x(p)\) to minimize her expected total baggage cost.

The expected total baggage cost consists of two parts, the booking cost and the penalty cost. Notice that \(O(p) = E([X(p) - x(p)]^+) = E(\max\{X(p) - x(p), 0\})\) represents the expected overweight quantity. Define \(x^*(p)\) as the optimal booking allowance to minimize the expected total baggage cost; \(O^*(p)\) as the corresponding expected overweight quantity; and \(C^*(p)\) as the corresponding expected total baggage cost. We have

\[
C^*(p) = bx^*(p) + pO^*(p).
\]

We also consider \(C^*(p)\) as the expected airline’s profit on baggage (from a passenger) given penalty cost \(p\).

In the following theorem, we show how the booking allowance, the overweight quantity, and the customer’s total baggage cost (the airline’s profit) are affected by the penalty cost.

**Theorem 1.** Given penalty cost \(p\), the optimal booking allowance \(x^*(p) = \frac{1}{\mu}[1 - e^{-\beta(p-b)} + e^{-\beta(p-b)} \ln(\frac{p}{b})]\), the corresponding expected overweight quantity \(O^*(p) = \frac{b}{p\mu}e^{-\beta(p-b)}\), and the corresponding expected airline’s profit \(C^*(p) = \frac{b}{\mu}[1 + e^{-\beta(p-b)} \ln(\frac{p}{b})]\). In addition,

- (a1) \(x^*(p)\) is quasiconcave in \(p\),
- (a2) \(\lim_{p \to \infty} x^*(p) = \frac{1}{\mu}\),
- (b1) \(O^*(p)\) is decreasing in \(p\),
- (b2) \(\lim_{p \to \infty} O^*(p) = 0\),
- (c1) \(C^*(p)\) is quasiconcave in \(p\),
- (c2) \(\lim_{p \to \infty} C^*(p) = \frac{b}{\mu}\).

**Proof of Theorem 1:** For newsvendor type of problems, the optimal solution \(x^*(p)\) satisfies
\[
\frac{p-b}{p} = \Pr \{ X(p) \leq x^*(p) \}. \]

From Equation (1), we should have
\[
\frac{p-b}{p} = \Pr \left\{ \alpha(p) \frac{1}{\mu} + [1 - \alpha(p)] \bar{X} \leq x^*(p) \right\} = \Pr \left\{ \bar{X} \leq \frac{x^*(p) - \alpha(p)}{1 - \alpha(p)} \right\} = 1 - e^{-\mu} \frac{x^*(p) - \alpha(p)}{1 - \alpha(p)},
\]

where the last equality comes from the fact that \( \bar{X} \) follows an exponential distribution with rate \( \mu \). It is then not hard to obtain that
\[
x^*(p) = \frac{1}{\mu} \left[ \alpha(p) + [1 - \alpha(p)] \ln(\frac{p}{b}) \right] = \frac{1}{\mu} \left[ 1 - e^{-\delta(p-b)} + e^{-\delta(p-b)} \ln(\frac{p}{b}) \right].
\]

To show Property (a1), take the derivative of \( x^*(p) \) with respect to \( p \), then we get
\[
x^*(p) = \frac{1}{\mu} \left( e^{-\delta(p-b)} \delta \right)
\]

where \( \delta(p) \) is defined as \( \delta(p) = \beta + \frac{1}{p} \ln \left( \frac{p}{b} \right) \). Now, notice that (1) \( \delta(p) \) is decreasing in \( p \); (2) \( \delta(b) = \beta + \frac{1}{b} > 0 \); and (3) \( \lim_{p \to \infty} \delta(p) = -\infty \). There exists a \( p^*_b > b \) such that \( \delta(p) \geq 0 \) for \( p \in [b, p^*_b] \) and \( \delta(p) < 0 \) for \( p \in (p^*_b, +\infty) \). Note that \( e^{-\delta(p-b)} > 0 \), and therefore the sign of \( x^*(p) \) follows that of \( \delta(p) \). Thus, we have \( x^*(p) \geq 0 \) for \( p \in [b, p^*_b] \) and \( x^*(p) < 0 \) for \( p \in (p^*_b, +\infty) \). That is, \( x^*(p) \) is quasiconcave in \( p \).

To prove Property (a2), notice that,
\[
\lim_{p \to \infty} x^*(p) = \lim_{p \to \infty} \frac{1}{\mu} \left( 1 - e^{-\delta(p-b)} + e^{-\delta(p-b)} \ln(\frac{p}{b}) \right) = \frac{1}{\mu} \left[ 1 + \lim_{p \to \infty} e^{-\delta(p-b)} \ln(\frac{p}{b}) \right].
\]

By L’Hospital’s rule,
\[
\lim_{p \to \infty} e^{-\delta(p-b)} \ln(\frac{p}{b}) = \lim_{p \to \infty} \frac{d \ln(\frac{p}{b})}{dp} = \lim_{p \to \infty} \frac{1}{\beta e^{\delta(p-b)}} = \lim_{p \to \infty} \frac{1}{\beta e^{\delta(p-b)}} = 0.
\]

Thus, we have \( \lim_{p \to \infty} x^*(p) = \frac{1}{\mu} \).

When the booking allowance equals \( x^*(p) \), the corresponding expected overweight quantity \( O^*(p) \) can be computed as
\[
O^*(p) = E([X(p) - x^*(p)]^+)
\]
\[
= [1 - \alpha(p)]E([\bar{X} - \frac{1}{\mu} \ln(\frac{p}{b})]^+)
\]
\[
= [1 - \alpha(p)] \int_{-\frac{1}{\mu} \ln(\frac{p}{b})}^{\infty} \left[ t - \frac{1}{\mu} \ln(\frac{p}{b}) \right] e^{-\mu t} dt
\]
\[
= [1 - \alpha(p)] \left\{ -te^{-\mu t} - \frac{1}{\mu} e^{-\mu t} + \frac{1}{\mu} \ln(\frac{p}{b})e^{-\mu t} \right\} \bigg|_{-\frac{1}{\mu} \ln(\frac{p}{b})}^{\infty}
\]
\[
= [1 - \alpha(p)] \left\{ \frac{1}{\mu} \ln(\frac{p}{b}) \cdot \frac{b}{p} + \frac{1}{\mu} \cdot \frac{b}{p} - \frac{1}{\mu} \ln(\frac{p}{b}) \cdot \frac{b}{p} \right\}
\]
\[
= [1 - \alpha(p)] \frac{b}{p} \mu
\]
\[
= \frac{b}{p} e^{-\beta(p-b)}.
\]

It is then easy to see that Properties (b1)-(b2) are true.
For the corresponding expected airline’s profit,

\[
C^*(p) = bx^*(p) + pO^*(p) \\
= \frac{b}{\mu} [\alpha(p) + [1 - \alpha(p)] \ln(\frac{p}{b})] + p[1 - \alpha(p)] \cdot \frac{b}{p\mu} \\
= \frac{b}{\mu} [1 + [1 - \alpha(p)] \ln(\frac{p}{b})] \\
= \frac{b}{\mu} [1 + e^{-\beta(p-b)} \ln(\frac{p}{b})].
\]

For Property (c1), take the derivative of \(C^*(p)\) with respect to \(p\), then we get

\[
C^*'(p) = \frac{b}{\mu} e^{-\beta(p-b)} \frac{1}{p} - \beta \ln(\frac{p}{b}) = \frac{b}{\mu} e^{-\beta(p-b)} \epsilon(p),
\]

where \(\epsilon(p)\) is defined as \(\epsilon(p) = \frac{1}{p} - \beta \ln(\frac{b}{p})\). Now, notice that (1) \(\epsilon(p)\) is decreasing in \(p\); (2) \(\epsilon(b) = \frac{1}{b} > 0\); and (3) \(\lim_{p \to \infty} \epsilon(p) = -\infty\). There exists a \(p^*_C > b\) such that \(\epsilon(p) \geq 0\) for \(p \in [b, p^*_C]\) and \(\epsilon(p) < 0\) for \(p \in (p^*_C, +\infty)\). Note that \(e^{-\beta(p-b)} > 0\), and therefore the sign of \(C^*'(p)\) follows that of \(\epsilon(p)\). Thus, we have \(C^*'(p) \geq 0\) for \(p \in [b, p^*_C]\) and \(C^*'(p) < 0\) for \(p \in (p^*_C, +\infty)\). That is, \(C^*(p)\) is quasiconcave in \(p\). Property (c2) can be proved in the same way as the proof of Property (a2).

Figure 1 illustrates the results in Theorem 1 for an example system with \(\mu = 1\), \(\beta = 0.1\), and \(b = 2\).

Figure 1: Optimal booking allowance, expected overweight quantity, and expected airline’s profit, as a function of penalty cost

For any given value of the penalty cost, Theorem 1 provides explicit representations of the corresponding optimal booking allowance, the associated expected overweight quantity, and the associated expected airline’s profit. It also describes how these quantities change with the penalty cost. First, we want to emphasize Property (a1). Notice that, intuitively, as if for a classic newsvendor problem, the optimal ordering quantity should always increase with the shortage cost. However, according to Property (a1), the optimal booking allowance first increases and then decreases with the penalty cost. We explain this as follows. First, it is
easy to see that, initially, for $p = b$, it is optimal for the customer to book for zero allowance and just pay the penalty at airport. As $p$ increases, the customer books for higher allowance to protect herself from the overweight penalty cost. Meanwhile, as $p$ increases, the customer gets more serious on baggage planning and does a better weight estimation. As a consequence, the variance of the baggage weight decreases, and the possibility and magnitude of overweight decrease. At some point, after the possibility and magnitude of overweight become sufficiently low, the customer no long needs to book for higher allowance for overweight protection. The booking allowance starts to decrease. As a limit, when $p$ approaches infinity, the baggage weight becomes deterministic, and the customer just needs to book for the exact weight, which equals $\frac{1}{\mu}$. This leads to Property (a2). Property (b1) states that the expected overweight quantity associated with the optimal booking allowance decreases with the penalty cost. There are two factors here. First, as we mentioned above, when $p$ increases, the variance of the baggage weight decreases, and the possibility and magnitude of overweight decrease. Second, from Property (a1), when $p$ increases, the optimal booking allowance first increases and then decreases. When $p$ is low, it is easy to see that, the combined force of the above two factors makes the expected overweight quantity decreasing in $p$. When $p$ is high, according to Property (b1), the first factor dominates the second one, and therefore the expected overweight quantity still decreases with $p$. As a limit, when $p$ approaches infinity, the expected overweight quantity approaches zero. This leads to Property (b2). Property (c1) says that the expected airline’s profit associated with the optimal booking allowance first increases and then decreases with the penalty cost. Notice that the airline’s profit consists of two parts, the booking revenue and the overweight penalty. From Property (a1), the booking revenue first increases and then decreases with the penalty cost. From Property (b1), the overweight penalty decreases with the penalty cost. According to Property (c1), when $p$ is low, the change in booking revenue dominates the change in overweight penalty, and therefore the airline’s profit first increases and then decreases with the penalty cost. As a limit, when $p$ approaches infinity, from Property (a2), the expected booking revenue approaches $\frac{b}{\mu}$; and from Property (b2), the expected overweight penalty approaches zero. As a consequence, the expected airline’s profit approaches $\frac{b}{\mu}$. This leads to Property (c2).

3 By-Minute Car Rental Problem

In Section 2, we showed how an individual passenger’s baggage booking decision and the airline’s profit are affected by the penalty cost. We explained the interesting and counterintuitive phenomenon that the passenger’s booking allowance and the airline’s profit do not monoton-
ically increase with the penalty cost. As we mentioned in Section 1, similar phenomenon is also observed in the by-minute car rental industry. In this section, we study how customers’ rental booking decisions and the firm’s profit are affected by the penalty cost, in the by-minute car rental setting. Notice that, for the low-cost airline baggage system, the passengers are independent with each other, and therefore the analysis of an individual passenger’s decision is sufficient. However, for the by-minute car rental system, a rental car can be rented by different customers, and one customer’s usage of the car could affect the other customers. For example, a following customer’s usage of the car could be delayed if a previous customer does not return the car on time. This interdependence among customers leads to more interesting but also more challenging modeling and analysis. For the ease of reading, we use similar notations and present results in similar fashions as in Section 2.

In this model, we consider a rental car which can be rented by different customers over the course of a day. For example the car can be booked by customer 1 from 10:00 to 12:00; customer 2 from 12:00 to 14:40; and customer 3 from 14:40 to 16:00. There is no time overlap between two bookings. We order the customers by the sequence that they use the car (e.g., customer 1 is the first to use the car). Assume that on a certain day, the car is fully booked by \( N \) customers in total; i.e., there is no time gap between two consecutive bookings. Let \( x_n \) denote the booking time duration of the \( n^{th} \) customer and assume \( x_n \in \mathbb{R} \). The booking price is \( b \) per unit time. In the event that customer \( n \) occupies the car for a time period longer than \( x_n \), a penalty cost of \( p \) per unit time delay will be charged to that customer. \( p \geq b \). If the car is not available at the beginning of the booking time window of customer \( n \) (due to the delayed return of the previous customers), a delay compensation of \( c \) per unit time delay will be paid to the affected customer by the rental firm. We assume that \( c = \gamma p \) with \( \gamma \geq 1 \). This avoids the unrealistic case that customers’ delay in returning the car is favorable to the rental firm.

For a travel need from a customer, the time duration to complete the trip is stochastic. As in Section 2, we assume that the penalty cost affects the distribution of the time duration to complete a trip. When the penalty cost increases, the variance of the time duration to complete the trip decreases.

Let \( X_n(p) \) denote the random variable that describes the time duration to complete the trip of customer \( n \). We represent \( X_n(p) \) as

\[
X_n(p) = \alpha(p) \frac{1}{\mu} + [1 - \alpha(p)] \tilde{X}_n,
\]

where \( \alpha(p) \) is given in Equation (2), and \( \tilde{X}_n \) follows an exponential distribution with rate \( \mu \). The interpretation of Equation (3) is that, as we mentioned in the introduction section, with higher
delay penalty cost, customers are more sensitive to delay, and therefore will be more serious on trip planning and time estimation. As a result, the variance of the trip duration decreases. Similar to Section 2, customer \( n \) decides the booking time duration \( x_n(p) \) to minimize her expected total trip cost.

We assume that \( \{x_n \mid n = 1, \ldots, N\} \) are independent and \( \{X_n(p) \mid n = 1, \ldots, N\} \) are independent. In other words, a customer’s booking and usage of the car will not be affected by the booking and usage from the other customers. This means that, even if a customer receives the car later than her booking time (due to the delayed return of previous customers), she will still use the car, and her time duration to complete the trip remains unchanged. On the other hand, even if the car becomes available earlier, a customer will not change her booking time. It is easy to see that, under this setting, the optimal booking time duration for every individual customer is the same and follows the expression in Theorem 1. Properties (a1)-(b2) still hold. We can then delete the subscript \( n \) in \( x_n \) and \( X_n \) for the ease of reading.

Although the structure of the total cost and the optimization problem for each individual customer are the same as in the low-cost airline baggage problem, the structure of the rental firm’s profit could be quite different. We first illustrate this using a system with two customers (if the car is only rented by one customer, the system is equivalent to the one for the low-cost airline baggage problem). Notice that, for the system with two customers, the rental firm needs to pay a delay compensation to customer 2 if customer 1 delays the return of the car. Therefore the expected rental firm’s profit is equal to the sum of the expected total trip costs of the two customers minus the expected delay compensation paid to customer 2. Define \( D^*(p) \) and \( P^*(p) \) as the expected delay compensation and the expected rental firm’s profit corresponding to the optimal booking time duration given penalty cost \( p \), respectively. We have

\[
P^*(p) = 2C^*(p) - D^*(p),
\]

with

\[
D^*(p) = \gamma p O^*(p),
\]

where \( C^*(p) \) and \( O^*(p) \) are defined in Section 2.

In the following theorem, we show how the delay compensation and the rental firm’s profit are affected by the penalty cost.

**Theorem 2.** For the system with two customers, the expected delay compensation \( D^*(p) = \frac{\gamma b}{\mu} e^{-\beta(p-b)} \), and the expected rental firm’s profit \( P^*(p) = \frac{b}{\mu}(2 + e^{-\beta(p-b)}[2 \ln(\frac{p}{b}) - \gamma]) \). In addition, (a1) \( D^*(p) \) is decreasing in \( p \),
(a2) \( \lim_{p \to \infty} D^*(p) = 0; \)

(b1) \( P^*(p) \) is quasiconcave in \( p \),

(b2) \( \lim_{p \to \infty} P^*(p) = \frac{2b}{\mu} \).

Proof of Theorem 2: The representations of \( D^*(p) \) and \( P^*(p) \) can be easily obtained from the results in Theorem 1. Properties (a1)-(a2) are trivial to show. For Property (b1), take the derivative of \( P^*(p) \) with respect to \( p \), then we get

\[
P^*(p) = \frac{2b}{\mu} e^{-\beta(p-b)} \varepsilon(p),
\]

where \( \varepsilon(p) \) is defined as \( \varepsilon(p) = \frac{3\gamma}{2} + \frac{1}{b} - \beta \ln(b \gamma) \). The rest of the proof is similar to the proof of Property (a1) in Theorem 1. Property (b2) can be proved in the same way as the proof of Property (a2) in Theorem 1.

Figure 2 illustrates the results in Theorem 2 for an example system with \( \mu = 1, \beta = 0.1, b = 2, \) and \( \gamma = 1.5. \)

Figure 2: Expected delay compensation, and expected rental firm’s profit, as a function of penalty cost

Theorem 2 provides, for the system with two customers, explicit representations of the expected delay compensation and the expected rental firm’s profit. It also describes how these quantities change with the penalty cost. The explanation of the properties here is similar to that after Theorem 1 in Section 2, and is therefore omitted for the sake of brevity.

Notice that, comparing the two-customer system with the one-customer system, intuitively, with one more customer, the rental firm’s total profit should increase. However, according to Figure 2, for small \( p \), the profit from the two-customer system is in fact lower than that from the one-customer system. This is because, when the penalty cost is low, the first customer’s return delay could be significant, and the revenue from the second customer may not be enough to cover the delay compensation. This implies that, from the rental firm’s perspective, when the system has more customers, a higher penalty cost may be preferred, in order to control the
return delay and the compensation. We formally state these in the following Theorem.

**Theorem 3.** (a) Define \( p = \bar{p} \) to be the unique root of \[ \gamma - \ln(p(b))e^{-\beta(p-b)} - 1 = 0 \] with \( p \geq b \) and \( \gamma > 1 \). When \( p < \bar{p} \), the rental firm’s profit from the two-customer system is less than that from the one-customer system.

(b) Define \( \epsilon_1 \) and \( \epsilon_2 \) to be the optimal penalty cost to maximize the expected rental firm’s profit for the one-customer system and the two-customer system, respectively. We have \( \epsilon_2 > \epsilon_1 \).

Proof of Theorem 3: (a) To compare the rental firm’s profit from the two-customer system with that from the one-customer system, define \( f(p) = C^*(p) - P^*(p) = \frac{b}{\mu}(\gamma - \ln(p(b))e^{-\beta(p-b)} - 1) \). We see that the two-customer system has a higher profit if and only if \( f(p) < 0 \). Let \( \bar{p} = \min\{p \mid f(p) \leq 0\} \). \( f(\bar{p}) \leq 0 \Rightarrow \gamma - \ln(\bar{p})e^{\beta(\bar{p}-b)} \Rightarrow \gamma e^{\beta(\bar{p}-b)} \geq \frac{b}{\mu} \Rightarrow \gamma \geq \frac{b}{\mu} e^{-\beta(\bar{p}-b)}(\gamma - \ln(\bar{p})) \Rightarrow f(p) = \frac{b}{\mu}e^{-\beta(\bar{p}-b)}[\gamma - \ln(\bar{p})] - e^{\beta(\bar{p}-b)}] < 0 \). Next, using L’Hospital’s rule, it is not hard to show that \( \lim_{p \to \infty} f(p) < 0 \). Now, since \( f(p) \) is continuous, and \( f(b) = \gamma - 1 > 0 \), \( f(p) = 0 \) must have roots. By the definition of \( \bar{p} \) and the fact that \( f(p) < 0 \), \( \forall p > \bar{p} \), we see that \( \bar{p} \) is the unique root of \[ \gamma - \ln(p(b))e^{-\beta(x-b)} - 1 = 0 \].

(b) From the representations of \( \epsilon(p) \) in the proof of Theorem 1 and \( \epsilon(p) \) in the proof of Theorem 2, we have \( \epsilon(p) = \epsilon(p) + \frac{\beta}{2} \). Hence, \( \epsilon(p) = \epsilon(p) + \frac{\beta}{2} = 0 + \frac{\beta}{2} > 0 \), where the second equality follows from the optimality of \( p_1^* \). Now, since (1) \( \epsilon(p) \) is decreasing in \( p \); and (2) \( \epsilon(p) = 0 \) (due to the optimality of \( p_2^* \)), we have \( p_2^* > p_1^* \).

Next, we would like to extend our results to systems with more than two customers. However, it is not hard to notice that, computing the delay compensation of customer \( n \) involves the convolution of the delay of all the previous customers, which makes the mathematical analysis intractable (as well-known in the scheduling literature). Through extensive numerical experiments, we see that the properties in Theorem 2 and Theorem 3 still hold, for systems with more than two customers.

Figure 3 illustrates these for an example system with \( \mu = 1, \beta = 0.1, b = 2, \gamma = 1.5 \) and number of customers up to four.

### 4 Conclusion

In this paper, we study a special newsvendor type of problem where the demand distribution depends on the shortage cost, and the optimal solution does not monotonically increase with the shortage cost. This problem is observed in two different case studies, the low-cost airline
baggage case, and the by-minute car rental case. In both cases, customers’ booking quantity decreases with the shortage penalty cost for some region of the penalty cost. The behavior explanation behind this interesting and unexpected phenomenon is that, when the shortage penalty cost increases, customers will be more serious on baggage (trip) planning and weight (time) estimation, and therefore, the underlying stochasticity in the actual baggage weight (usage time duration) decreases. As a consequence, the possibility and magnitude of shortage (i.e., overweight or return delay) decrease, and customers do not need to book more for shortage protection.

Through two stochastic decision models, we analytically capture how important quantities (e.g., customers’ booking allowance, customers’ total cost, and firm’s profit) are affected by the shortage penalty cost. We show the quasiconvavity of customers’ booking allowance on penalty cost, which theoretically explains our practical observation. Our results and insights on the optimal penalty cost (with respect to the number of customers in system) can be used as guidelines for the firm to maximize its profit.

For future research, there are two meaningful directions. First, we would like to study the detail impact of the penalty cost on customers’ booking decisions in different practical settings. Notice that, the function forms that we used in our model (e.g., Equations (1) and (2)) are constructed (based on the interview findings) for the ease of deriving analytical results. Although the main results and insights remain valid for other function forms with the same key properties, to help the firm to maximize its profit in real practice, we need to know the exact relationship between the penalty cost and the other quantities. This can be done either by analyzing the firms’ data on booking and penalty, or by conducting behavior laboratory experiments with human subjects. Second, we would like to look at newsvendor type of problems in other practical settings where the “demand” distribution depends on the cost parameters (e.g., shortage cost or overage cost), and explore other classic stochastic operations management models where the

Figure 3: Systems with More than Two Customers
assumed exogenous stochasticity in fact depends on some endogenous variables. The insights from our paper could possibly be extended to explain other counterintuitive phenomena in practice.

References


