Submitted to *XXXXXXXXXXX*
manuscript as of *December 28, 2010*

An Analysis of "Fixed-Up-To" (FUT) Pricing Using A Stochastic Model of Consumer Behavior

Atanu Lahiri
Michael G. Foster School of Business, University of Washington

Rajiv M. Dewan and Marshall Freimer
William E. Simon Graduate School of Business Administration, University of Rochester

"Fixed-Up-To" (FUT) tariffs, which are commonly used for pricing telecommunications services, are made up of three parts: a monthly subscription fee (fixed fee), a usage-limit (allowance), and an over-limit rate that applies to consumptions over the usage-limit. Typically, usage during each billing cycle is metered and billed separately. Consequently, a consumer, who randomly gets opportunities (or feels a need) to use the service, makes his usage-decisions depending on the allowance at hand as well as the time left in the ongoing billing cycle. We propose a stochastic model to capture this decision-making process of a consumer. We provide a closed-form expression for the utility of a FUT pricing plan under such uncertainty. We also examine a monopolist's tariff design problem in which consumers are offered a menu of FUT plans. We model two types of consumer heterogeneity: that in the rate at which opportunities to use the service arrives, and that in the average willingness-to-pay per opportunity. We characterize optimal FUT price menus and also numerically compare them with two-part tariffs and other simpler nonlinear menus. We prove that, contrary to the findings in the literature on nonlinear pricing, there are situations in which it is optimal to serve both "high" and "low" type consumers regardless of their relative proportions.

Key words: Nonlinear Pricing, Second-degree Price Discrimination, Fixed-Up-To Tariff, Pricing of Telecommunications Services

1. Introduction

Fixed-up-to or FUT pricing is quite common in practice. Many cellular phone voice plans in the US look like: \$50 per month for usage up to 500 "plan" minutes and \$0.40 per minute for usage over 500 minutes. The \$50 fee is commonly called the *fixed fee*, the usage-limit of 500 minutes is also called the *included allowance* or just *allowance*, and the \$0.40 rate at which over-limit usage is priced is commonly referred to as the *over-limit rate*. Consumers often have a menu to choose from (see Table 1). The apparent reason that such menus comprising several FUT plans exist is that they help sellers practice second-degree price discrimination.

There exists some thought-provoking research on FUT pricing. Masuda and Whang (2006) argue that over-limit rates are not necessary to serve a market that has discrete consumer types and in which consumers face no uncertainty regarding their monthly usages. On the other hand, Lambrecht et al. (2007) and Grubb (2009) have argued that consumers' inability to choose right plans for

Fixed Fee	Allowance	Over-limit Rate
$39.99 / month	450 minutes	$0.45/ minute
$59.99 / month	900 minutes	$0.40/ minute
$69.99 / month	Unlimited	—

Table 1 Individual Voice Plans Offered by a National Carrier (as reported on its website on 11-29-10)

themselves is the reason is that some of them often go over their usage-limits while others fail to use their allowances fully. The existing research does not explain whether FUT pricing would be useful if consumers were capable of choosing right plans for themselves. Besides, it does not explain how a seller should design a price schedule comprising FUT plans. We fill these voids.

We model the phenomenon that opportunities to use a telecommunication service (e.g., opportunities to make calls) arrive randomly to a consumer and, for each opportunity, the consumer decides on whether to use the service or to save his allowance for more valuable opportunities. Consuming the allowance at hand increases the risks of going over the usage-limit and incurring over-limit charges. On the other hand, not using the service and saving the allowance means forgoing a valuable opportunity to use the service. In our model, every consumer is a utility-maximizing decision-maker who decides on each opportunity in a way that maximizes his or her expected utility. Our model also considers the fact that not all opportunities received by a consumer may have the same value; for example, the opportunity to have a routine conversation with a friend is unlikely to be as valuable to the consumer as the need to call a doctor during a medical emergency.

By solving the dynamic program that models consumers' utility-maximizing behavior, we derive a closed-form expression for the expected *utility* of a FUT plan to a consumer. We also show that a forward-looking consumer uses a state-dependent "hurdle" for deciding which opportunities to accept. If the value of an arriving opportunity is below this hurdle, he forgoes the opportunity and waits for better opportunities. If the value exceeds the hurdle, he accepts the opportunity. We analyze how this hurdle depends on the consumer's state, i.e., on the inventory of allowance at hand and the time left in the billing cycle.

We provide closed-form expressions for the *expected allowance consumption* per period and *expected over-limit usage* per period. We show that the expected over-limit usage during a cycle is always positive even though the expected consumption of the allowance is always strictly less than the included allowance. This finding is consistent with those of Lambrecht et al. (2007), who empirically find that a fraction of consumers go over limit while the rest does not. However, we argue that the reason is *not* that some consumers sign up for plans that offer less allowance than they need and the others sign up for plans with more allowance than they need. As we show, the reason is the nature of the uncertainty that each consumer faces.

After examining the consumer behavior under uncertainty, we examine the second-degree price discrimination problem of a monopolist facing a heterogeneous consumer market. Second-degree discrimination is the most common form of pricing in the US: all major carriers in the US offer consumers menus to choose from. In our model consumers differ in their arrival rates (i.e., the rate at which opportunities to use the service arrive differs from one consumer segment to another) and in their average willingness-to-pay per opportunity (e.g., consumers who use their phones for business purposes typically gets a higher value per call than those using their phones for personal purposes). We analyze both forms of heterogeneity.

Specifically, we examine a setting in which there are two classes of consumers. When one class gets opportunities at a higher rate, or values each opportunity more on average, or both, it gets a higher expected utility from every plan than does the other class. We call this class the *high type*. We call the other class the *low type*. Among the key findings of our research is that it is always optimal to serve both the high type and low type when the only source of heterogeneity is the rate at which opportunities arrive. This result is in contrast to the well-known result that the low type is served only when the fraction of the high type is not sufficiently high (Mussa and Rosen 1978). The reason our result differs is that, in the case of FUT pricing, regardless of the relative proportions of the two types, the seller is able to control the consumer surplus that it needs to cede to the high type. As we explain later, it does so by optimally adjusting the over-limit rate and allowance of the low type's plan.

Additionally, we find that, regardless of the nature of consumer heterogeneity, when a zero-marginal cost seller serves both consumer types, it is optimal for it to offer an "unlimited" plan (i.e., a plan not having any over-limit charges) to the high type and to offer a "limited" plan to the low type, explaining in part the prevalent tariff structure in the US. We also show that the optimal FUT menu outperforms traditional .

We also examine the other scenario in which one class of consumers gets more opportunities but the other class values each more on average. In this case one class may get a higher utility from some plans and the other class may get a higher utility from the others. Therefore, no consumer class is strictly the high type and no class is strictly the low type. We show that, in this setting as well, one consumer class is always served at optimality: we *label* this class the "high type" and the other class the "low type." We further prove that, in situations in which the "high type" is also the class with the higher arrival rate, it is optimal to serve both types regardless of their proportions, due to very similar reasons.

The rest of the paper is organized as follows. We first discuss the related literature. Next, we analyze the consumer's problem. Then, we briefly discuss the case of a monopolist facing a homogeneous market. Finally, we discuss the two kinds of heterogeneity mentioned earlier: (a) the consumer class who experiences a higher arrival rate values each opportunity more on average, and (b) the consumer class who experiences a higher arrival rate values each opportunity less on average. We also present numerical examples to compare different types of plans that a seller may offer, e.g., unlimited plans, pay-go plans, etc. We conclude by summarizing and explaining major contributions.

2. Literature

Our work primarily belongs to the long stream of work in the area of price discrimination using nonlinear tariffs (Mussa and Rosen 1978, Goldman et al. 1984, Sundararajan 2004). The existing literature explains in detail issues involved in design and implementation of nonlinear tariffs (Mussa and Rosen 1978, Moorthy 1984). In particular, two-part tariffs, which are special cases of nonlinear tariffs, have been extensively studied and have been shown to be optimal in many circumstances (Oi 1971, Rao and Peterson 1998). FUT tariffs, however, have not received adequate attention despite their popularity.

The problem of designing a price menu comprising FUT plans has so far been examined only under limited circumstances, e.g., Masuda and Whang (2006) describe the optimal FUT tariff and prove its optimality assuming that there are discrete consumer types with each type facing a deterministic demand in every period. One major contribution of Masuda and Whang (2006) is that they examine the optimality of FUT pricing in the presence of congestion externalities (Naor 1969, Mendelson 1985). However, their assumption that each consumer faces a deterministic demand every period leads to a setting in which no consumer goes over-limit in the equilibrium. As a result, the seller finds it optimal to use plans that do not have any over-limit component. Also, each consumer fully uses his allowance every period. Our contribution is that we examine the problem in a setting in which consumers face a random demand, explaining the need for an over-limit rate.

Grubb (2009) explains the usefulness of the over-limit rate by arguing that consumers "mis-estimate" their random needs. On the contrary, we show that using an over-limit rate is important even when consumers are fully informed about their demand distributions. We show that the over-limit rate is critical to the seller's menu design problem and that the seller controls the consumer surplus ceded to high type consumers effectively by offering low and high type consumers plans having different over-limit rates.

Lambrecht et al. (2007) examines FUT pricing empirically by assuming a quadratic form for consumers' reservation prices. Their main conclusion is that consumers often chooses sub-optimally from the menu offered by the seller. While quadratic forms are common in empirical research, they do not fully capture the major aspects of consumer behavior under FUT pricing. Our contribution is that we do not assume any functional form for the utility to begin with. We model consumers' utility-maximizing behavior and derive the utility function. Using this derived utility function, we show that consumers tend to use their plans more frequently when they have a greater allowance at hand or have less time left in the billing cycle. Besides, we show also that, on average, each consumer incurs a positive over-limit charge in every billing period, which happens despite the fact that the same consumer, on average, does not use his allowance fully. These conflicting phenomena occur even when consumers choose optimally from the menu offered.

It is worthwhile nothing that the stochastic model of the consumer's utility-maximization problem is non-stationary. This non-stationarity arises because any unused allowance during a billing cycle expires at the end of the cycle. Because of this non-stationarity, it is suboptimal for consumers to use fixed hurdles for deciding which opportunities to accept. We thus differ substantially from stochastic pricing models such as Miller and Buckman (1987). We are somewhat closer to the literature on finite-horizon inventory pricing (Gallego and van Ryzin 1994), which examines inventory-dependent prices that are similar to the state-dependent hurdle in our model.

The most salient novelty of our approach is that we combine elements of finite-horizon non-stationary stochastic optimization (Gallego and van Ryzin 1994) with economic theories on second-degree price discrimination (Mussa and Rosen 1978). As already mentioned, we model the decision-making problem of rational forward-looking consumers using a stochastic dynamic program. We solve this problem to determine the expected utility of a FUT plan to a consumer. This expected utility is what we then use for solving the seller's second-degree price discrimination problem.

3. The Consumer's Problem

Consider a consumer who has uncertain demand for a service over a period T. We structure this demand as:

- Emergence of opportunities to use this service described by a counting process $\{n(t),\ t \leq T\}$ where $n(t)$ is the number of opportunities that have arisen up to time t.

- Each opportunity has a value v that is not known *a-priori* but is revealed to the customer when the opportunity emerges. Let these values be independently drawn from a distribution G that is public knowledge.

Consider a consumer who signs up for a periodic service for a fee of $\$F$ per period that is T long. The service offers an allowance of Z units, and an over-limit rate of $\$r$ per unit that applies to usage beyond the allowance. We will assume that the opportunity to the use the service follows a stochastic process.

Assumption 1. *Opportunities to use the service (e.g., to make calls) arrive according a Poisson process with rate λ.*

3.1. Expected Utility

At a point in the middle of the period of service, let t and z represent the amount of time and allowance, respectively, remaining in the period. Let $u(t,z)$ denote the expected utility of the remaining part of the contract at this point in the period. This utility includes the value of calls made less any applicable over-limit charges, for the expected activity in the t remaining time in the period. The fixed fee is not included in this term. On an ex-ante basis, the expected utility net of fees is $u(T,Z)-F$.

Let a non-negative random variable v represent the value of a randomly arriving opportunity. We assume that the uncertainty regarding the value of an opportunity is resolved at the instant it arrives to a consumer, i.e., the consumer can precisely figure out the realization of v before making the decision to accept or reject the opportunity. If the consumer decides to accept, his unused allowance goes down by one unit; otherwise, the allowance remains the same.

Let y denote the number of opportunities arriving in an arbitrarily small interval of length Δt. Then, for a consumer seeking to maximize his expected utility, the following holds:

$$\begin{aligned} u(t+\Delta t,z) &= P[y=0]u(t,z)+P[y=1]E_v[\max\{u(t,z-1)+v,u(t,z)\}]+o(\Delta t) \\ &= u(t,z)+\lambda\Delta t E_v[\max\{v-(u(t,z)-u(t,z-1)),0\}]+o(\Delta t) \qquad (1) \end{aligned}$$

After dividing by Δt and taking the limit $\Delta t\to 0$, we get $\forall z>0$:

$$u_t(t,z)=\lambda E_v[\max\{v-(u(t,z)-u(t,z-1)),0\}] \qquad (2)$$

From the above we note that a consumer in state (t,z) accepts the opportunity to make the call if and only if the realized value v exceeds the hurdle $u(t,z)-u(t,z-1)$.[1] Henceforth, we denote this hurdle by $h(t,z)$.

We make the following assumption to simplify the problem:

[1] Given the fact that consumers can check their available allowance at any time over phone and Internet, it is not unreasonable to assume that they can find out this information at little or no additional cost, and, thus, are always able to make informed decisions.

Assumption 2. *The value of an opportunity v follows an Exponential distribution with mean μ.*

Using this assumption we can rewrite equation 2 as:

$$u_t(t,z) = \lambda\mu e^{-(u(t,z)-u(t,z-1))/\mu} \tag{3}$$

We next determine the boundary conditions for equation 3. The expected utility of the remaining part of the contract is zero if no time is left. Only calls that have value that exceed the over-limit rate of r will be made if no allowance is left.

$$u(0,z) = 0,\ \forall z > 0 \tag{4}$$

$$u(t,0) = \lambda t \int_r^{\infty} (v-r)\frac{1}{\mu}e^{-v/\mu}dv = \lambda\mu t\, e^{-r/\mu},\ t \geq 0 \tag{5}$$

The differential equation 3 with boundary equations 4 and 5 can easily be solved by repeated integration. The solution is shown in the next lemma.

Lemma 1. *Let $g(t,z) = e^{u(t,z)/\mu}$ and $\rho = e^{-r/\mu}$. Then, the following holds $\forall z \geq 0$ and $\forall t > 0$.*

$$\begin{aligned} g(t,z) &= \sum_{i=0}^{z} \frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty} \frac{(\lambda t)^i (\exp(-r/\mu))^{(i-z)}}{i!} \\ &= \rho^{-z}\left(e^{\rho\lambda t} - \sum_{i=0}^{z} \frac{(\lambda t)^i}{i!}(\rho^i - \rho^z)\right) \end{aligned} \tag{6}$$

Since $u(t,z) = \mu \log[g(t,z)]$, Lemma 1 solves the dynamic program. It is easy to see that the net expected utility of a plan, $u(T,Z) - F$, is increasing in λ and μ. In other words, a consumer who gets opportunities at a higher rate or values each opportunity more on average gets a higher utility from the same plan.

Example 1. Figure 1 shows how the expected utility, $u(t,z)$, behaves, and illustrates Lemma 1. The plot assumes the following parameter values: $\lambda = 2$/day, $r/\mu = 1$. As is seen from the plot, the impact of a higher allowance is very little when there is little time (and opportunity) left to use them fully. Further, at any given instant of time, the expected utility rises with the allowance. However, the rise is initially more rapid and then it becomes less and less rapid as the allowance increases (the spacing between a line and its neighbor decreases as z increases), indicating that

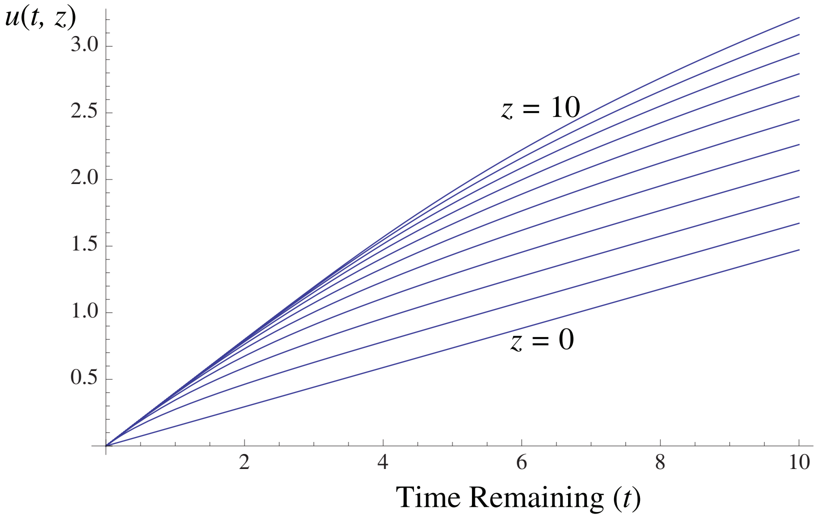

Figure 1 **Expected Utility (Value of Calls Made Minus Over-limit Charges) ($\lambda = 2$ per day, $r = \mu =$ $0.20)**

at any instant of time the marginal utility is diminishing. In other words, it is not optimal for a seller (i.e., seller) facing a positive marginal cost to offer an infinite amount of allowance to any consumer. We discuss the seller's problem in detail in the next section.

With $u(t,z)$ obtained above, we now turn to the hurdle $h(t,z) = u(t,z) - u(t,z-1)$ for the value v that the consumer uses to decide whether or not utilize an opportunity to make a call.

3.2. The Decision Hurdle

We are now going to establish that a rational consumer acts most conservatively (i.e., uses the highest hurdle) after he fully uses up the included allowance Z. At all other occasions he uses a lower hurdle. Further, we are going to prove that the consumer acts more conservatively if he has less inventory of unused allowance at hand or if he has more time left in the billing cycle. The following theorem, which describes these characteristics of the hurdle, i.e., the hurdle function, in effect summarizes all salient aspects of the consumer behavior under FUT pricing.

THEOREM 1. *The decision hurdle of a rational consumer, $h(t,z)$, is equal to the over-limit rate r when the unused allowance $z = 0$. The hurdle is strictly decreasing in z, increasing in the time t left in the billing cycle, and increasing in the over-limit rate r.*

After the consumer exhausts his allowance, he compares values of arriving opportunities to the over-limit rate. If the value of an opportunity is greater, he accepts it. However, as Theorem 1 shows, using the same hurdle is not utility-maximizing if the allowance is not completely exhausted. A lower hurdle is utility-maximizing in case there is unused allowance. Interestingly, even when there is unused allowance, the hurdle is sensitive to the over-limit rate. Because a higher over-limit rate compels the consumer to adopt a greater caution with regard to how he uses his allowance and how he manages the risks of going over the usage-limit. For the same reason, a higher hurdle is also optimal when the allowance is mostly used up. Further, as the theorem shows, the consumer typically uses a higher hurdle earlier in the billing cycle: one way to interpret this behavior is to recognize that using too much allowance early raises the risks of incurring an excessive over-limit charge.

Example 2. Figure 2 shows how the hurdle, $h(t,z)$, varies with respect to the time left, t, and the allowance left, z. It illustrates Theorem 1. The plot assumes the following parameter values: $\lambda = 2/$ day, $r = \mu =$ \$0.20. As is evident from it, the hurdle goes down rapidly with z, the inventory of unused allowance at hand, when t, the time remaining, is very low. The rate of decline is lower at higher values of t. In other words, when there is little time remaining in the billing cycle, the consumer's decision-making process is very sensitive to the inventory of unused plan minutes at hand while it is not so when there is more time remaining. This numerical observation points to the fact that the consumer adjusts the hurdle sharply down near the end of the billing cycle in order to use up any unused allowance. Earlier in the billing cycle the consumer does not make sharp adjustments because reducing the hurdle too low increases the risk exhausting the allowance too soon.

One major implication of Theorem 1 is that the consumer uses a variable or dynamic hurdle that depends on z and t, i.e., the consumer's state. It is natural to ask whether using such a variable hurdle is at all critical. Or, could the consumer do as well using a fixed hurdle? We now numerically answer this question. According to Lemma 1, if $r =$ \$0.30, $u(t = 10$ days, $z = 60$ units$)$ is \$11.25 for $\lambda = 15/$day and $\mu =$ \$0.10. On the other hand, if the consumer uses a fixed hurdle, $\hat{h}$, his expected

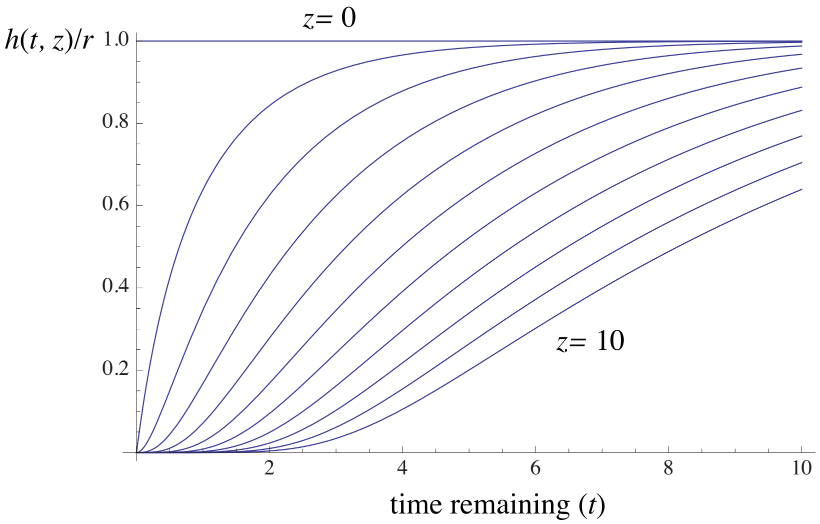

Figure 2 **Decision Hurdle ($\lambda = 2$ per day, $r = \mu =$ $0.20)**

utility from using this fixed hurdle, $\hat{u}$(z, t, $\hat{h}$), is going to be as follows:

$$\hat{u}(z,t,\hat{h}) = \lambda t e^{-\hat{h}/\mu}(\mu + \hat{h}) - r\sum_{i=z+1}^{\infty}(i-z)\exp(-\lambda t e^{-\hat{h}/\mu}))\frac{(\lambda t e^{-\hat{h}/\mu})^i}{i!}$$

The first term above is simply the expected value from use while the second term is the expected over-limit charge. By numerically studying this function at different values of $\hat{h}$ (see Figure 3), we observe that the best fixed hurdle or hurdle for a consumer starting with an allowance of 60 units with 10 days of time left is about $0.10. This hurdle leads to an expected utility of $\hat{u}$(z=60, t=10, $\hat{h}$=$0.10) or $10.68, which is about 5% below what the consumer gets using a variable decision hurdle. In fact, the loss from using a fixed hurdle may even be larger depending on the parameter values assumed. Hence, it is critical that one considers dynamic decision hurdles for analyzing the utility of a FUT plan to a consumer who faces uncertainty with regard to usage.

3.3. Expected Allowance Consumption

We now consider the expected consumption of the allowance (i.e., the expected total usage minus the expected over-limit usage) during $[T-t,T]$. Let us denote this expected value by $m(t,z)$. Then, $\forall z > 0$, the following holds:

$$\begin{aligned} m(t+\Delta t, z) =& \lambda\Delta t P[v > h(t,z)](m(t,z-1)+1) \\ &+ (1-\lambda\Delta t P[v > h(t,z)])m(t,z) + o(\Delta t) \end{aligned} \quad (7)$$

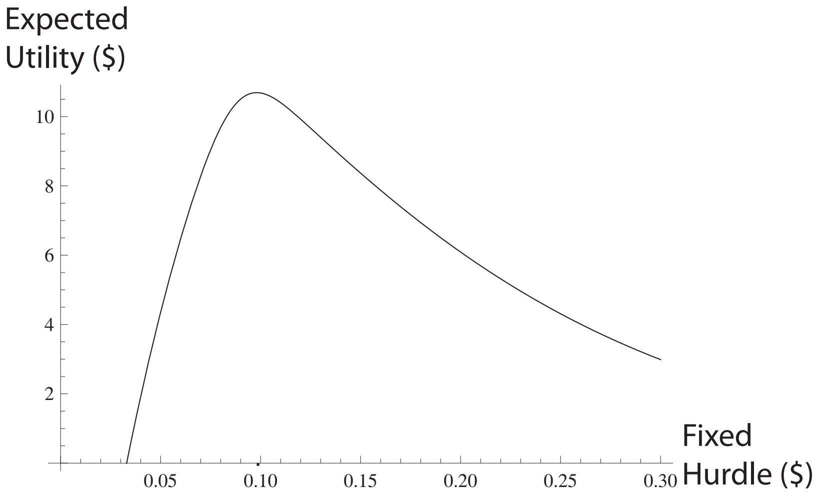

Figure 3 Expected Utility from Using a Fixed Threshold ($\lambda = 15$/ day, μ =$0.10/ call, and r =$0.30/ over-limit call); Starting Allowance: 60 Calls, Time until the End of the Cycle: 10 Days

And, the boundary conditions are:

$$m(0,z) = 0, \forall z > 0 \tag{8}$$

$$m(t,0) = 0, \forall t \geq 0 \tag{9}$$

LEMMA 2. *For all $z \geq 0$ and $t > 0$:*

$$m(t,z)g(t,z) = zg(t,z) - \sum_{i=0}^{z} \frac{(z-i)(\lambda t)^i}{i!} \tag{10}$$

The expected consumption for a consumer starting with an allowance of Z units that expire in T time units is $m(T,Z)$. According to the lemma above, $m(T,Z) < Z$, i.e., the average consumption of the allowance is strictly less than the allowance. This happens because there is always a positive probability that the consumer may not get sufficient number of opportunities to use up his allowance fully. Empirical findings (Lambrecht et al. 2007) also suggest that a fraction of consumers do not use up their allowance fully, which implies that our model is consistent with empirical observations.

3.4. Expected Over-limit Usage

We now consider the over-limit usage in $[T-t,T]$. Let us denote the expected value of this usage by $n(t,z)$. Then, $\forall z > 0$, the following holds:

$$\begin{aligned} n(t+\Delta t, z) =& \lambda \Delta t P[v > h(t,z)] n(t, z-1) \\ &+ (1 - \lambda \Delta t P[v > h(t,z)]) n(t,z) + o(\Delta t) \end{aligned} \tag{11}$$

And, the boundary conditions are:

$$n(0,z) = 0,\ \forall z > 0 \tag{12}$$

$$n(t,0) = \lambda t\rho,\ \forall t \geq 0 \tag{13}$$

Lemma 3. *For all $z \geq 0$ and $t > 0$:*

$$n(t,z)g(t,z) = \sum_{i=z+1}^{\infty} \frac{(i-z)(\lambda t)^i \rho^{(i-z)}}{i!} \tag{14}$$

The expected over-limit usage at the beginning of the billing cycle is $n(T,Z)$. According to the lemma above, this expected over-limit usage is positive. Recall that the expected value of the unused part of the allowance, $Z - m(T,Z)$, is also positive. In other words, each consumer, on average, does not exhaust the allowance but at the same time goes over-limit. We state this important result, which follows immediately from Lemmas 2 and 3, as a theorem.

Theorem 2. *The expected allowance consumption is strictly less than the allowance, and at the same time the expected over-limit usage is strictly positive.*

Theorem 2 solves the puzzle that prior researchers have observed (Lambrecht et al. 2007). Often, this puzzling phenomenon has been attributed to some consumers buying plans with allowances too big for their need and others buying plans with allowances too small. What we show here is that this phenomenon has nothing to do with what consumers buy. It is a byproduct of the random demand that consumers face. The probability that a consumer does not get enough opportunities to use the allowance is positive just as the probability that he gets plenty of opportunities is.

Using Lemma 3, we can also characterize, $T^0(T,Z)$, the expected length of time between the instant the consumer exhausts his free allowance, Z, and the instant the billing period ends. The following corollary provides this characterization.

Corollary 1.

$$T^0(T,Z) = \frac{n(T,Z)}{\lambda\rho}$$

4. The Seller's Problem

We now consider the monopolist's tariff design problem assuming that the monopolist faces a constant marginal cost. We make this constant marginal cost assumption to separate out issues related to congestion (which leads to an increasing marginal cost) or scale-economy (which leads to a decreasing marginal cost). While the issues on the cost side are important and relevant, we do not want them to clutter our analysis of the revenue side. The assumption that the monopolist faces a constant marginal cost essentially allows us to focus on the revenue side of the problem. We first briefly examine the scenario in which the seller faces a homogeneous market. Then, we examine a heterogeneous market with two consumer types.

Assumption 3. *The monopolist faces a constant marginal cost of c dollars per unit.*

Let us introduce a few terms to describe different possible classes of tariff the monopolist may offer besides offering a nonlinear tariff made up of FUT plans. We intend to numerically compare these different classes of tariffs with the optimal FUT tariff to assess the impact of FUT pricing on both the seller and consumers.

Definition 1. "Unlimited" plan: An unlimited plan is one that charges a fixed fee (i.e., $F > 0$) and offers a over-limit rate of zero (i.e., $r = 0$).

Definition 2. "Simple nonlinear" plan: A simple nonlinear plan is one that charges a fixed fee (i.e., $F > 0$), offers a strictly positive allowance (i.e., $Z > 0$), but does not allow any over-limit calls.

Definition 3. "Pay-go" plan: A pay-go plan is one that does not charge a fixed fee (i.e., $F = 0$) and does not offer any allowance (i.e., $Z = 0$); consumers pay as they use the service based on a strictly positive over-limit rate (i.e., $r > 0$).

Definition 4. "Two-part" plan: A two-part plan is one that charges a fixed fee (i.e., $F > 0$) but does not offer any allowance ($Z = 0$); consumers pay as they use the service based on a strictly positive over-limit rate (i.e., $r > 0$).

Note that a consumer using an unlimited plan can use the service as much as he wants without paying additional charges. Equivalently, these plans offer an infinite allowance. The other three types do not offer the same benefit, i.e., they can be termed "limited."

Plan type	Optimal plan	Profit	Consumer Surplus
FUT plans	$F=42.51,\ Z=0,\ r=0.15$	42.51	0.00
Two-part plans	$F=42.51,\ Z=0,\ r=0.15$	42.51	0.00
Simple nonlinear plans	$F=42.05,\ Z=214,\ r=\infty$	42.05	0.00
Pay-go plans	$F=0,\ Z=0,\ r=c+\mu=0.35$	15.64	15.64
Unlimited plans	$F=\lambda\mu T=90,\ Z=\infty,\ r=0$	22.50	0.00

Table 2 Homogeneous market ($c=$\$0.15/unit, $\lambda=15$/day, $T=30$ days, and $\mu=$\$0.20/unit); F, r, Profit, and Consumer Surplus are all in \$

4.1. Homogeneous Market

In this subsection we assume that all consumers get opportunities to use the service at a rate λ. Further, for each consumer, the value distribution is exponential with a mean of μ. F denotes the fixed part of the FUT plan offered, Z the allowance included, and T the length of the billing cycle. Consumers make their decisions to buy (or not use) at the beginning of the cycle, and they do so based on their expected utility $U(T,Z)$ and the fixed fee F.

The Individual Rationality (IR) constraint, that says that a consumer uses the service only if it provides a non-negative net utility, is as follows.

$$u(T,Z)\geq F \qquad \text{(IR)}$$

The monopolist in this case extracts the entire surplus. Therefore, this IR constraint binds. The profit, $\pi(Z,r)$ is consequently given by the following expression. Note that the seller's problem here is to maximize $\pi(Z,r)$ subject to the constraints $Z\geq 0$ and $r\geq 0$.

$$\pi(Z,r)=u(T,Z)-c\,m(T,Z)+(r-c)n(T,Z)$$

Lemma 4. *In the case of homogeneity, it is optimal for the seller to choose $Z=0$ and $r=c$. Hence, if the marginal cost is zero, it is optimal for the seller to offer an unlimited plan.*

Example 3. *Homogeneous Market.* We know from Lemma 4 that the optimal FUT plan in this homogeneous case is the same as the optimal two-part plan. Let us examine how different types of plans compare in this case. Table 2 shows this comparison.

Table 2 shows that the best simple nonlinear plan leads to a profit of \$42.05, which is close to the optimal profit of \$42.51. This plan is however not practical as consumers typically do not want plans with hard usage-limits. It is also apparent from the table that pay-go plans are the worst from the seller's viewpoint while they are also the best from the consumer's perspective — these plans are currently offered by smaller sellers in the US, who lack pricing power. Most big sellers have substantial monopoly-like pricing power (which results from alliances between sellers and handset manufacturers, switching costs, network effects, a lack of portability of handsets, among other reasons) and they do not use such plans. The reason such plans are not attractive from the seller's viewpoint is that they do not have a fixed fee, which makes it impossible for the seller to effectively extract surplus from consumers.

As is also evident from Table 2, unlimited plans are the worst from the social planner's angle: they lead to an excessive amount of consumption as consumers accept all opportunities including those worth less than their marginal cost. Note that unlimited plans are not viable in some cases; for example, if $\mu < c$, all unlimited plans will lead to a negative profit and therefore the seller will not offer such plans.

4.2. Heterogeneous Market: Arrival Rates and Average Valuations Identically Ordered

Let us now consider two types of consumers, namely 1 and 2, comprising fractions $(1-\alpha)$ and α of the overall population. The rates at which opportunities arrive are assumed to be λ_1 for type 1 and λ_2 for type 2, respectively. Similarly, the average valuations associated with these opportunities are assumed to be μ_1 for type 1 and μ_2 for type 2, respectively. Note that the parameters that do not depend on the consumer type are the length of the billing period, T, and the marginal cost, c.

In this subsection we consider the case in which $\lambda_2 \geq \lambda_1$ and $\mu_2 \geq \mu_1$ with at least one of the two inequalities being strict. Such ordering means that not only type 2 gets an equal or higher value on average from every opportunity than does type 1 but it also gets on average at least as many opportunities to use the service as does type 1. Hence consumer type 2 gets a strictly higher utility from every plan, i.e., type 2 can be regarded as the "high type."

The seller's problem is to optimally design two FUT plans, namely (F_1, Z_1, r_1) and (F_2, Z_2, r_2), one aimed at each consumer type. Henceforth, *we use shortcuts for the expected utility functions and the expected usage functions:* we denote $u(T,Z)$ computed using an over-limit rate of r, an arrival rate of λ_i, and an average reservation price per opportunity of μ_i by $u_i(Z,r)$, $i \in \{1,2\}$; and, we express the corresponding usages by $m_i(Z,r)$ and $n_i(Z,r)$, respectively. We use similar shorthands for all other functions as well, for example, we use $g_i(Z,r)$ to denote $g(T,Z)$ computed using parameter values of r, λ_i, and μ_i.

The Individual Rationality constraints (IR) for the two consumer types are:

$$u_1(Z_1,r_1) \geq F_1 \tag{IR1}$$

$$u_2(Z_2,r_2) \geq F_2 \tag{IR2}$$

The Incentive Compatibility constraints (IC) for the two consumer types are:

$$u_1(Z_1,r_1) - F_1 \geq u_1(Z_2,r_2) - F_2 \tag{IC1}$$

$$u_2(Z_2,r_2) - F_2 \geq u_2(Z_1,r_1) - F_1 \tag{IC2}$$

Since $u_2(T,Z) \geq u_1(T,Z)$ at all Z and r, the optimal policy for seller occurs when IR1 and IC2 binds (Mussa and Rosen 1978). When IR1 and IC2 bind, the monopolist's profit is as follows.

$$\begin{aligned}
&\pi(Z_1,r_1,Z_2,r_2)\\
&\qquad = (1-\alpha)(F_1 - cm_1(Z_1,r_1) + (r_1-c)n_1(Z_1,r_1))\\
&\qquad + \alpha(F_2 - cm_2(Z_2,r_2) + (r_2-c)n_2(Z_2,r_2))\\
&\qquad = (1-\alpha)(u_1(Z_1,r_1) - cm_1(Z_1,r_1) + (r_1-c)n_1(Z_1,r_1))\\
&\qquad + \alpha(u_1(Z_1,r_1) + u_2(Z_2,r_2) - u_2(Z_1,r_1) - cm_2(Z_2,r_2) + (r_2-c)n_2(Z_2,r_2))
\end{aligned}$$

Theorem 3. *In the presence of consumer heterogeneity, it is always optimal for the seller to choose $Z_2 = 0$ and $r_2 = c$. Further, if it is optimal to serve the low type, then $r_1 > c$.*

The marginal cost is zero for many forms of telecommunication. In this case, we show that an unlimited plan is optimal to offer.

Corollary 2. *If the marginal cost is zero, it is optimal for a seller serving both types to offer an unlimited plan to the high type (type 2) and to offer a limited plan to the low type (type 1).*

The implication of Theorem 3 is that it is optimal to charge the low type a higher over-limit rate, which is consistent with what we observe in practice. In some sense, this finding is similar to

Consumer Type	Fixed Fee	Allowance	Over-limit Rate
Type 2, the high type	$F_2 =$\$ 109.00	$Z_2 = \infty$	$r_2 =$\$0.00
Type 1, the low type	$F_1 =$\$ 89.16	$Z_1 = 404$	$r_1 =$\$0.22

Table 3 **Optimal FUT Price Menu ($c = 0$, $\lambda_1 = 15$/day, $\lambda_2 = 25$/day, $T = 30$ days, $\alpha = 0.2$, and $\mu_1 = \mu_2 =$\$0.20/unit**

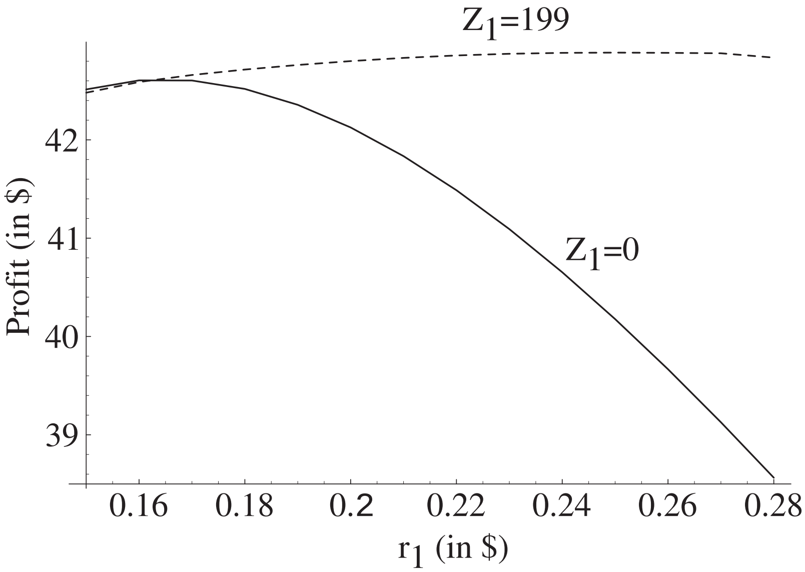

Figure 4 **Maximum value of $\pi(Z_1, r_1, Z_2, r_2)$ at $c =$\$0.15/unit, $\lambda_1 = 15$/day, $\lambda_2 = 25$/day, $T = 30$ days, $\alpha = 0.1$, and $\mu_1 = \mu_2 =$\$0.20/unit**

existing findings (Mussa and Rosen 1978): the high type gets the same service as what it would get when there are no low type consumers; but the same is not true for the low type. However, as is described by following theorems, the conditions under which the low type is served differ significantly in the case of FUT pricing.

Example 4. *The optimal price menu for a zero-marginal cost seller.* Let us illustrate Corollary 2 using a numerical example. Table 3 shows plans that the high type and low type will get if $c = 0$, $\lambda_1 = 15$/day, $\lambda_2 = 25$/day, $T = 30$ days, $\alpha = 0.2$, and $\mu_1 = \mu_2 =$\$0.20/unit. These optimal plans look similar to the ones shown in Table 1: the high type gets an unlimited plan with a higher fixed fee, and the low type gets a limited plan with lower fixed fee and a steep over-limit rate. Further, since both the allowance and over-limit components of the low type's plan is positive, we can infer that *two-part pricing is not optimal in the presence of usage-uncertainty.*

Example 5. *Heterogeneity in the Arrival Rate.* We now examine a case in which the marginal cost is not zero. Figure 4 assumes that $c =$\$0.15/unit, $\lambda_1 = 15$/day, $\lambda_2 = 25$/day, $T = 30$ days,

Plan type	Optimal Menu	Profit	Consumer Surplus
FUT plans	$F_1 = 71.82,\ Z_1 = 199,\ r_1 = 0.25$ $F_2 = 49.97,\ Z_2 = 0,\ r_2 = 0.15$	42.89	2.09
Two-part plans	$F_1 = 40.44,\ Z_1 = 0,\ r_1 = 0.16$ $F_2 = 43.90,\ Z_2 = 0,\ r_2 = 0.15$	42.60	1.72
Simple nonlinear plans	$F_1 = 72.16,\ Z_1 = 202,\ r_1 = \infty$ $F_2 = 103.03,\ Z_2 = 356,\ r_2 = \infty$	42.76	2.06
Pay-go plans	$F_1 = 0,\ Z_1 = 0,\ r_1 = 0.35$ $F_2 = 0,\ Z_2 = 0,\ r_2 = 0.35$	16.68	16.68
Unlimited plans	$F_1 = 90,\ Z_1 = \infty,\ r_1 = 0$ $F_2 = 90,\ Z_2 = \infty,\ r_2 = 0$	24.00	6.00

Table 4 Heterogeneous market (c =$0.15/unit, $\lambda_1 = 15$/day, $\lambda_2 = 25$/day, $T = 30$ days, $\alpha = 0.1$, and $\mu_1 = \mu_2$ =$0.20/unit); F_1, r_1, F_2, r_2, Profit, and Consumer Surplus are all in $

$\alpha = 0.1$, and $\mu_1 = \mu_2$ =\$0.20/unit, and that r_2 and Z_2 have been chosen optimally, i.e., $r_2 = c =$ \$0.15 and that $Z_2 = 0$. The two plots show how the seller's profit varies vis-à-vis r_1 at two different levels of Z_1. Evidently, a Z_1 of 199 leads to a higher profit than does a Z_1 of 0. Further, note that the plot for $Z_1 = 199$ reaches its peak when r_1 is \$0.25. By examining the profit function over wider ranges of values of Z_1 and r_1, it can be confirmed that the optimal choices for the example shown are in fact $Z_1 = 199$ and r_1 =\$0.25. As is predicted by Theorem 3, the optimal r_1 of \$0.25 is bigger than the marginal cost of \$0.15.

Let us now examine how different classes of plans compare. Table 4, which assumes the same arrival rates, average reservation prices, and marginal cost as does Figure 4, shows this comparison. First of all, it numerically illustrates the fact that two-part pricing is not optimal in the presence of heterogeneity. However, the optimal set of two-part plans still leads to a profit that is very close to what does the optimal set of FUT plans.

Note that, in this case, both consumer types get the same plan when the seller restricts itself to just pay-go plans or to just unlimited plans — because the seller cannot implement self-selection with either of these two classes of tariffs. Table 4 shows that these two classes of plans lead to substantially lower profits when compared to the optimal FUT menu, echoing findings from our earlier example. Besides, we again find that unlimited plans can substantially harm social welfare.

Interestingly, in this case of heterogeneity, the optimal set of FUT plans leads to both a higher profit and a higher consumer surplus vis-à-vis the two-part and simple nonlinear plans. A salient

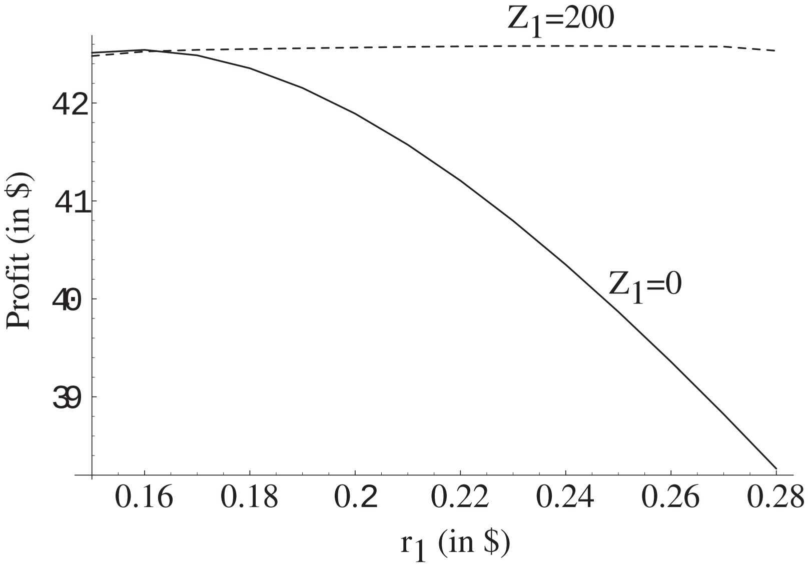

Figure 5 **Maximum value of** $\pi(Z_1, r_1, Z_2, r_2)$ **at** $c =$**$0.15/unit,** $\lambda_1 = \lambda_2 = 15$**/day,** $T = 30$ **days,** $\alpha = 0.1$**,** $\mu_1 =$**$0.20/unit, and** $\mu_2 =$**$0.33/unit**

implication of this finding is that, in the presence of usage-uncertainty, FUT pricing may dominate traditional forms of pricing such as the two-part pricing or simple nonlinear pricing from both the seller's and consumers' viewpoints. The common perception, that consumers always lose when a seller uses a more complex form of tariff to increase its profits, is therefore not true in the case of FUT pricing.

Example 6. *Heterogeneity in the Average Willingness-to-pay Per Opportunity.* Figure 5 assumes that $\lambda_1 = \lambda_2 = 15$/day, $T = 30$ days, $\alpha = 0.1$, $\mu_1 =$\$0.20/unit, and $\mu_2 =$\$0.33/unit, and that r_2 and Z_2 have been chosen optimally, i.e., $r_2 = c =$ \$0.15 and that $Z_2 = 0$. The two plots show how the seller's profit varies vis-*à*-vis r_1 at two different levels of Z_1. Evidently, a Z_1 of 200 is better from the seller's viewpoint than is a Z_1 of 0. Using numerical techniques, it can even be confirmed that the optimal choices for this example are $Z_1 = 200$ and $r_1 =$\$0.24. Thus, we again observe that it is important to use a large Z_1 in conjunction with a r_1 that is larger than the marginal cost.

Theorem 4. *If the only source of consumer heterogeneity is in the rates at which opportunities arrive, i.e., if* $\lambda_2 > \lambda_1$ *but* $\mu_2 = \mu_1$*, it is always optimal for the seller to serve both consumer types regardless of their proportions.*

If the sources of heterogeneity include how consumers value opportunities on average, i.e., if $\mu_2 > \mu_1$*, it is not always optimal for the seller to serve both consumer types. A sufficient condition for the seller to serve both types under optimality is that the fraction of the high type,* α*, is less than* $\frac{\lambda_1\mu_1}{\lambda_2\mu_2}\exp(-(1/\mu_1 - 1/\mu_2)c)$*.*

According to the Theorem 4, when consumers are heterogeneous only with regard to their arrival rates, it is optimal for the seller to serve both consumer classes regardless of their relative sizes. The implication is that the consumer surplus is always positive; the high type (type 2) always gets a consumer surplus that equals $u_2(Z_1, r_1) - u_1(Z_1, r_1)$ per high type consumer. Theorem 4 is thus in conflict with the existing result, which says that it makes sense for the seller to serve the low type and concede a surplus to the high type only when the fraction of the high type is below a certain critical threshold (Mussa and Rosen 1978).

The intuition behind this surprising result is as follows. When the only source of heterogeneity is in the arrival rate of opportunities, the seller can reduce the consumer surplus, which is $u_2(Z_1, r_1)$ $- u_1(Z_1, r_1)$ per high type consumer, as much as it wants by increasing r_1 alone (see also the proof in the Appendix). When $\mu_1 = \mu_2$, increasing r_1 essentially reduces the utility that type 2 gets from type 1's plan more than it reduces that type 1 gets (see Figure 6 for an illustration). In addition, doing so keeps type 2 away from type 1's plan. The downside is that reducing type 1's utility of the plan makes the plan unattractive to type 1. However, with FUT pricing, the seller can overcome this pitfall by compensating type 1 with an adequate allowance, Z_1. This is also the reason that the seller often chooses a positive Z_1 at optimality (see Example 5).

When consumers are heterogeneous in their valuations of opportunities, raising the over-limit rate of the low type's plan is not effective in reducing the consumer surplus — because $u_2(Z_1, r_1)$ $- u_1(Z_1, r_1)$ cannot be made arbitrarily small by raising r_1 alone when $\mu_2 > \mu_1$. As a result, the low type may not served at optimality when its fraction is not sufficiently large.

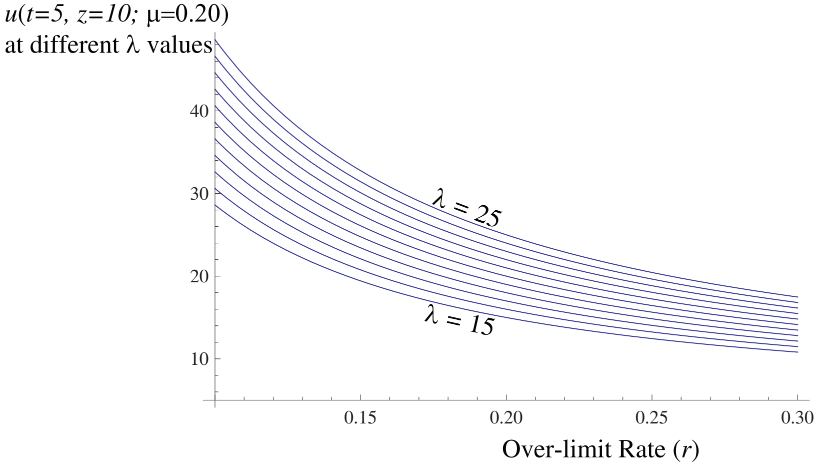

Figure 6 $u(t=5, z=10; r, \mu=0.20)$ **for different values of** $\lambda \in \{15, 16, ..., 25\}$

4.3. Heterogeneous Market: Arrival Rates and Average Valuations Inversely Ordered

We now examine the scenario in which either (a) $\lambda_2 > \lambda_1$ and $\mu_2 < \mu_1$ or (b) $\lambda_2 < \lambda_1$ and $\mu_2 > \mu_1$. Such ordering means that one consumer type gets a higher value on average per opportunity while the other type gets more opportunities on average. In this situation either consumer type can get a higher utility from a plan than does the other type — which type will get a higher utility will depend on the allowance as well as the over-limit rate. In other words, no consumer type can be strictly regarded as the "high type." Therefore, it is not necessary that at optimality the IR1 and IC2 constraints mentioned in the previous subsection bind. It is also not necessary that either type 2 or both types are served at optimality, i.e., it may be optimal for the seller to serve type 1 only. The following theorem provides us with insights into the solution to this problem.

Theorem 5. *In the setting in which one consumer class experiences a higher arrival rate but the other class has a higher reservation price per opportunity, if* $\lambda_2\mu_2 T\exp(-c/\mu_2) > \lambda_1\mu_1 T\exp(-c/\mu_1)$, *type* 2 *is the "high type" in the sense that it is optimal to serve either type* 2 *or both types. Similarly, type* 1 *is the "high type" if the direction of the inequality gets reversed. Additionally, if the "high type" also has a higher arrival rate, e.g., if* $\lambda_2 > \lambda_1$, $\mu_2 < \mu_1$, *and type* 2 *is the high type, it is optimal for the seller to serve both types regardless of their relative sizes.*

Theorem 5 essentially suggests that it is not profitable to exclude consumer type 2 if $\lambda_2\mu_2 T\exp(-c/\mu_2) > \lambda_1\mu_1 T\exp(-c/\mu_1)$ (and the same is true for type 1 if the direction of the inequality reverses). In fact, in this circumstance, type 2 also gets served the same plan as what it would get if there were no type 1 consumers. Recall that, in the case of second-degree price discrimination, the high type is always served the efficient quantity, i.e., the quantity it would be served if there were no low type consumers. In this *limited sense* we can label type 2 the "high type." Also, note that, in the rare degenerate case in which $\lambda_2\mu_2 T\exp(-c/\mu_2) = \lambda_1\mu_1 T\exp(-c/\mu_1)$, the seller can maximize its profit by selling the same plan, which charges a fixed fee of $\lambda_2\mu_2 T\exp(-c/\mu_2)$, offers no allowance, and charges an over-limit rate of c, to both types.

Further, as Theorem 5 establishes, if the primary reason the seller finds it optimal to serve a consumer type is that it gets more opportunities, then it is also optimal to serve both types: this part of Theorem 5 is similar to Theorem 4 and so is the underlying intuition. For example, if $\lambda_2 > \lambda_1$ but $\mu_2 < \mu_1$, it becomes possible for the seller to control the consumer surplus by offering type 1 a plan that has a substantially higher over-limit rate — raising this over-limit rate reduces the utility type 2 can get from type 1's plan more than it reduces what type 1 can, cutting in the process the "information rent" or the consumer surplus ceded to type 2. The seller can still lure type 1 into purchasing its plan by offering an adequate Z_1.

5. Summary and Conclusion

This work is the first attempt to analytically investigate the problem of designing FUT tariffs in a setting in which consumers face stochastic demands for the service being priced. We analyze this problem bottom-up starting with consumers. Contrary to the models used in the literature, we do not assume any specific form for the utility. Instead we assume that consumers are utility-maximizing. We model their utility-maximizing behavior. We hypothesize that opportunities to use the service arrive according to a Poisson process and that the value of each opportunity is drawn from an Exponential distribution. These hypotheses allow us to formulate the utility-maximization problem as a stochastic dynamic program. We derive the utility of a FUT plan to consumer by solving the dynamic program.

The solution to the consumer's problem revealed that consumers use a variable state-dependent hurdle for deciding which opportunities should be accepted. This variable hurdle is always less than or equal to the over-limit rate. Further, we show that consumers use a higher hurdle if they have less allowance at hand or have more time left in the billing cycle. In other words, when faced with a lower amount of unused allowance, consumers get more conservative and forgo more opportunities. Similarly, earlier in the billing cycle, they use conservatively in order to minimize the chances of exhausting their allowances too early.

Prior research has observed that consumers either go over the limit or fail to utilize their allowance. It has attributed such behavior to consumers' inability to pick suitable plans. We show that, even when consumers are able to choose rationally, the expected allowance consumption is strictly less than the allowance while the expected over-limit usage is strictly positive at the same time. Our analysis points to the fact that the reason behind the observed consumer behavior is essentially the stochastic demand that they face.

Turning to the seller's problem, we examine a market with two consumer classes. A major theoretical contribution here is that we develop a framework for analyzing second-degree price discrimination for a setting in which consumers have stochastic demand.

We examine two forms of consumer heterogeneity. Consumers differ either in the arrival rate or the average willingness-to-pay or both. The existing research shows that, in markets characterized by a high type and a low type, the low type is not always served (Mussa and Rosen 1978). We, however, show that the heterogeneity in the arrival rate is very different in this respect. We prove that, regardless of sizes of the two classes, it is optimal to serve both when they differ only in their arrival rates. We explain the intuition behind this counter-intuitive result. The seller's strategy in this case essentially involves increasing the over-limit rate of the low type. A higher over-limit rate for the low type keeps the high type away from the low type's plan while still allows the seller to lure the low type with a sufficiently large allowance. We also prove that not serving the low type can be optimal if consumer classes differ in their willingness-to-pay per opportunity.

The seller's problem is particularly complex when one class of consumers has a higher arrival rate but the other class has a higher willingness-to-pay per opportunity. No consumer group in this case gets a higher utility from every plan than the other. However, we show that it is still possible to label one class the high type and the other the low type. Moreover, we prove that, in this situation as well, it is optimal to serve both types regardless of their relative sizes so long as the high type is the consumer class experiencing the higher arrival rate and the low type is the consumer class willing to pay more per opportunity.

The analysis in this paper also explains the structure of the optimal FUT pricing menu. We prove that it is optimal to charge the high type an over-limit rate that equals the marginal cost. Further, we show that the low type, if served, needs to be charged an over-limit rate that exceeds the marginal cost. Interestingly, if the seller's marginal cost is zero, this structure implies that at optimality the high type is served an unlimited plan while the low type is served a limited plan, which is very similar to the pricing structure we observe today in the US.

Using numerical analysis, we also find that tariffs such as pay-go plans or unlimited plans, which either use a zero fixed fee or a zero over-limit rate, are significantly worse than the optimal FUT tariff, i.e., they lead to significantly lower profits. The lesson is that a positive over-limit rate is critical to ensuring that consumers do not engage in excessive consumption while a positive fixed fee is critical to extracting surplus from them. Besides, we find that two-part tariffs, which have been shown to be optimal in many circumstances (Oi 1971, Rao and Peterson 1998), are also not optimal under uncertainty; this finding explains in part the popularity of FUT pricing.

Finally, it is worth noting that this research is also not without limitations. Most notable is that it does not take into account congestion issues. Congestion is increasingly becoming a critical issue in the US because of the growing demand and popularity of broadband wireless data services. We intend to address this issue in our future research projects. This research, despite this limitation, contributes to the literature on pricing of information goods and services by providing a novel model of consumer behavior in a setting in which consumers face an uncertain demand for an information service, and by examining the problem of FUT tariff design under such uncertainty.

References

Gallego, G., G. van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science* **40**(8) 999–1020.

Goldman, M. B., H. E. Leland, D. S. Sibley. 1984. Optimal nonuniform prices. *Review of Economic Studies* **51**(2) 305–319.

Grubb, M. D. 2009. Selling to overconfident consumers. *American Economic Review* **99**(5) 17701807.

Lambrecht, A., K. Seim, B. Skiera. 2007. Does uncertainty matter? consumer behavior under three-part tariffs. *Marketing Science* **26**(5) 698–710.

Masuda, Y., S. Whang. 2006. On the optimality of fixed-up-to tariff for telecommunication service. *Information Systems Research* **17**(3) 247–253.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Communications of the ACM* **28** 312–321.

Miller, B. L., A. G. Buckman. 1987. Cost allocation and opportunity costs. *Management Science* **33**(5) 626–639.

Moorthy, S. 1984. Market segmentation, self-selection, and product line design. *Marketing Science* **31** 288–307.

Mussa, M., S. Rosen. 1978. Monopoly and product quality. *Journal of Economic Theory* **18** 301–317.

Naor, P. 1969. On the regulation of queue size by levying tolls. *Econometrica* **37** 15–24.

Oi, W. Y. 1971. A Disneyland dilemma: Two-part tariffs for a Mickey mouse monopoly. *Quarterly Journal of Economics* **86** 77–90.

Rao, S., E. R. Peterson. 1998. Optimal pricing of priority services. *Operations Research* **46**(1) 46–56.

Sundararajan, A. 2004. Nonlinear pricing of information goods. *Management Science* **50**(12) 1660–1673.

Appendix. Additional Lemmas and Proofs

[**Proof of Lemma 1**] In order to solve equations 3, we first note that this system of differential equations can be expressed as:

$$\frac{dg(t,z)}{dt} = \lambda g(t,z-1) \tag{15}$$

Using this equation and the second boundary condition (equation 5) as the starting point, we solve equation 15 by repeatedly integrating and determining the constants of integration using the first boundary condition (equation 4).

[**Proof of Lemma 2**] Note that $P[v > h(t,z)] = \exp(-h(t,z)/\mu) = \frac{g(t,z-1)}{g(t,z)}$. After dividing both sides of equation 7 by Δt and letting $\Delta t \to 0$, we get

$$\begin{aligned} &\frac{dm(t,z)}{dt} + \lambda\frac{g(t,z-1)}{g(t,z)}m(t,z) = \lambda\frac{g(t,z-1)}{g(t,z)}(m(t,z-1)+1) \\ \Rightarrow &\frac{dm(t,z)}{dt}g(t,z) + \lambda g(t,z-1)m(t,z) = \lambda g(t,z-1)(m(t,z-1)+1) \\ \Rightarrow &\frac{d}{dt}(m(t,z)g(t,z)) = \lambda g(t,z-1)(m(t,z-1)+1) \end{aligned} \tag{16}$$

The last step follows from equation 15. Using the second boundary condition (equation 9) as the starting point, we solve equation 16 by repeatedly integrating and obtaining the constants of integration using the other boundary condition (equation 8).

[**Proof of Lemma 3**] Note that $P[v > h(t,z)] = \exp(-h(t,z)/\mu) = \frac{g(t,z-1)}{g(t,z)}$. After dividing both sides of equation 11 by Δt and letting $\Delta t \to 0$, we get

$$\begin{aligned} &\frac{dn(t,z)}{dt} + \lambda\frac{g(t,z-1)}{g(t,z)}n(t,z) = \lambda\frac{g(t,z-1)}{g(t,z)}n(t,z-1) \\ \Rightarrow &\frac{dn(t,z)}{dt}g(t,z) + \lambda g(t,z-1)n(t,z) = \lambda g(t,z-1)n(t,z-1) \\ \Rightarrow &\frac{d}{dt}(n(t,z)g(t,z)) = \lambda g(t,z-1)n(t,z-1) \end{aligned} \tag{17}$$

The last step follows from equation 15. Using the second boundary condition (equation 13) as the starting point, we solve equation 17 by repeatedly integrating and obtaining the constants of integration using the other boundary condition (equation 12).

[**Proof of Theorem 2**] The proof follows immediately from Lemmas 2 and 3.

[**Proof of Corollary 1**] Let X be a random variable that denotes the length of the time between the point the consumer uses up his allowance and the point the billing period ends. If the period ends before the consumer uses up his allowance, X would be 0. The proof follows from the following relationship:

$$\begin{aligned} n(T,Z) &= E_X[n(T,Z)|X] \\ &= E_X[\lambda X \exp(-r/\mu)] \\ &= \lambda E[X]\exp(-r/\mu) \end{aligned}$$

Lemma 5.
$$\frac{\partial n(t,z)}{\partial r} < 0,\ \frac{\partial n(t,z)}{\partial \mu} > 0,\ \frac{\partial n(t,z)}{\partial \lambda} > 0,\ \frac{\partial n(t,z)}{\partial t} > 0.$$

And, $n(t,z) \le n(t,z-1)$, $\forall z > 0$.

[**Proof of Lemma 5**] This proof requires using equation 14.

$$\begin{aligned} &\frac{\partial n(t,z)}{\partial r} \\ &= \frac{-(1/\mu)\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)\sum_{j=z+1}^{\infty}\frac{(j-z)^2(\lambda t)^j(\exp(-r/\mu))^{(j-z)}}{j!}}{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)^2} \\ &+ \frac{(1/\mu)\sum_{i=z+1}^{\infty}(i-z)\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\sum_{j=z+1}^{\infty}(j-z)\frac{(\lambda t)^j(\exp(-r/\mu))^{(j-z)}}{j!}}{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)^2} \\ \\ &= \frac{-(1/\mu)\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!}\sum_{j=z+1}^{\infty}\frac{(j-z)^2(\lambda t)^j(\exp(-r/\mu))^{(j-z)}}{j!}}{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)^2} \\ &+ \frac{(1/\mu)\sum_{i=z+1}^{\infty}\sum_{j=z+1}^{\infty}((i-z)(j-z)-(j-z)^2)\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\frac{(\lambda t)^j(\exp(-r/\mu))^{(j-z)}}{j!}}{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)^2} \\ \\ &= \frac{-(1/\mu)\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!}\sum_{j=z+1}^{\infty}\frac{(j-z)^2(\lambda t)^j(\exp(-r/\mu))^{(j-z)}}{j!}}{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)^2} \\ &+ \frac{-(1/\mu)\sum_{i=z+1}^{\infty}\sum_{j=i+1}^{\infty}(i-j)^2\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\frac{(\lambda t)^j(\exp(-r/\mu))^{(j-z)}}{j!}}{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)^2} \\ &< 0 \end{aligned}$$

In a similar fashion we can show that $\partial n(t,z)/\partial \mu > 0$.
To prove $\partial n(t,z)/\partial \lambda > 0$, we need to use equation 14 again.

$$\begin{aligned} &\frac{\partial n(t,z)}{\partial \lambda} \\ &= \frac{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)\sum_{j=z+1}^{\infty}jt(j-z)\frac{(\lambda t)^{(j-1)}(\exp(-r/\mu))^{(j-z)}}{j!}}{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)^2} \\ &- \frac{\big(\sum_{i=0}^{z}it\frac{(\lambda t)^{(i-1)}}{i!} + \sum_{i=z+1}^{\infty}it\frac{(\lambda t)^{(i-1)}(\exp(-r/\mu))^{(i-z)}}{i!}\big)\sum_{j=z+1}^{\infty}(j-z)\frac{(\lambda t)^j(\exp(-r/\mu))^{(j-z)}}{j!}}{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)^2} \\ \\ &= \frac{\sum_{i=0}^{z}\sum_{j=z+1}^{\infty}(j-z)(j-i)t\frac{(\lambda t)^{(i+j-1)}(\exp(-r/\mu))^{(j-z)}}{i!j!}}{\big(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!} + \sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\big)^2} \end{aligned}$$

$$
\begin{aligned}
&+\frac{\sum_{i=z+1}^{\infty}\sum_{j=z+1}^{\infty}(j-i)(j-z)t\frac{(\lambda t)^{(i+j-1)}(\exp(-r/\mu))^{(i+j-2z)}}{i!j!}}{\left(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!}+\sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\right)^2}\\
&=\frac{\sum_{i=0}^{z}\sum_{j=z+1}^{\infty}(j-z)(j-i)t\frac{(\lambda t)^{(i+j-1)}(\exp(-r/\mu))^{(j-z)}}{i!j!}}{\left(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!}+\sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\right)^2}\\
&+\frac{\sum_{i=z+1}^{\infty}\sum_{j=i+1}^{\infty}(i-j)^2t\frac{(\lambda t)^{(i+j-1)}(\exp(-r/\mu))^{(i+j-2z)}}{i!j!}}{\left(\sum_{i=0}^{z}\frac{(\lambda t)^i}{i!}+\sum_{i=z+1}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}\right)^2}\\
&>0
\end{aligned}
$$

In a similar fashion we can show that $\partial n(t,z)/\partial t > 0$.

To prove that $n(t,z)$ is decreasing in z, we use the fact that the probability that the consumer, who has z units of allowance at time t before the end of the billing cycle, does not make any calls until the end of the cycle, i.e., during $[T-t,T]$, is at least $\exp(-\lambda t)$ (because the probability that no opportunities arrive during this period is $\exp(-\lambda t)$ and only a fraction of the arriving opportunities are used). Therefore, with a probability of $\exp(-\lambda t)$ or more the consumer does not make any over-limit calls. Further, in $[T-t,T]$, if the consumer makes his first call at time $T-t'$, where $T-t' \in [T-t,T]$, the expected number of over-limit calls made until the end of the cycle is $n(t', z-1)$, which is less than $n(t, z-1)$ because $n(t,z-1)$ is increasing in t. Therefore:

$$
\begin{aligned}
n(t,z) &< \exp(-\lambda t).0 + (1-\exp(-\lambda t))n(t,z-1)\\
&< n(t,z-1)
\end{aligned}
$$

Lemma 6.

$$
\begin{aligned}
\frac{\partial u(t,z)}{\partial r} &= -n(t,z)\\
\frac{\partial g(t,z)}{\partial r} &= -g(t,z)n(t,z)/\mu\\
\frac{\partial m(t,z)}{\partial r} &= \frac{(-1/\mu)n(t,z)}{g(t,z)}\sum_{i=0}^{z}\frac{(z-i)(\lambda t)^i}{i!}
\end{aligned}
$$

[**Proof of Lemma 6**] It follows from equations 14 and 6 that:

$$\frac{\partial g(t,z)}{\partial r} = -n(t,z)g(t,z)/\mu$$

It then follows from the definition of $g(t,z)$ that:

$$
\begin{aligned}
\frac{\partial u(t,z)}{\partial r} &= \frac{\mu}{g(t,z)}\frac{\partial g(t,z)}{\partial r}\\
&= -n(t,z)
\end{aligned}
$$

Further, using equation 10, we get:

$$
\begin{aligned}
\frac{\partial m(t,z)}{\partial r} &= \frac{\partial g(t,z)}{\partial r}\frac{1}{g(t,z)^2}\sum_{i=0}^{z}\frac{(z-i)(\lambda t)^i}{i!}\\
&= \frac{-n(t,z)}{\mu g(t,z)}\sum_{i=0}^{z}\frac{(z-i)(\lambda t)^i}{i!}
\end{aligned}
$$

[**Proof of Theorem 1**] By definition, $h(t,z) = u(t,z) - u(t,z-1)$, $\forall z > 0$. Therefore:

$$
\begin{aligned}
\exp(h(t,z)/\mu) &= \frac{g(t,z)}{g(t,z-1)}\\
&= \frac{\sum_{i=0}^{z-1}\frac{(\lambda t)^i}{i!}+\sum_{i=z}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}}{\sum_{i=0}^{z-1}\frac{(\lambda t)^i}{i!}+\exp(-r/\mu)\sum_{i=z}^{\infty}\frac{(\lambda t)^i(\exp(-r/\mu))^{(i-z)}}{i!}} \qquad (18)\\
&< \exp(r/\mu)
\end{aligned}
$$

Also, equation 18 implies that $\exp(h(t,z)/\mu)$ decreases in z and increases in t. To prove this part, let us define the following:

$$\begin{aligned} A(t,z) &= \sum_{i=0}^{z-1} \frac{(\lambda t)^i}{i!} \\ B(t,z) &= \sum_{i=z}^{\infty} \frac{(\lambda t)^i (\exp(-r/\mu))^{(i-z)}}{i!} \end{aligned}$$

Therefore:

$$\begin{aligned} \exp(h(t,z)/\mu) &= \frac{A(t,z)+B(t,z)}{A(t,z)+\exp(-r/\mu)B(t,z)} \\ &= \frac{A(t,z)/B(t,z)+1}{A(t,z)/B(t,z)+\exp(-r/\mu)} \end{aligned} \tag{19}$$

According to equation 19, $\exp(h(t,z)/\mu)$ is a decreasing function of $A(t,z)/B(t,z)$. Let us therefore investigate how this ratio changes with z and t. Let us first consider t and t' where $t' > t$.

$$\begin{aligned} \frac{A(t',z)}{A(t,z)} &\le (\frac{t'}{t})^{z-1} \\ &< (\frac{t'}{t})^{z} \\ &\le \frac{B(t',z)}{B(t,z)} \end{aligned}$$

Re-arranging the inequality above, we get $A(t,z)/B(t,z) > A(t',z)/B(t',z)$, which implies that $\exp(h(t,z)/\mu) < \exp(h(t',z)/\mu)$ or that $h(t,z)$ increases with t. Next, we analyze what happens when z changes.

$$\begin{aligned} A(t,z+1)B(t,z) &= \sum_{i=0}^{z} \frac{(\lambda t)^i}{i!} \sum_{j=z}^{\infty} \frac{(\lambda t)^j (\exp(-r/\mu))^{(j-z)}}{j!} \\ &> \sum_{i=1}^{z} \frac{(\lambda t)^i}{i!} \sum_{j=z}^{\infty} \frac{(\lambda t)^j (\exp(-r/\mu))^{(j-z)}}{j!} \\ &= \sum_{i=0}^{z-1} \frac{(\lambda t)^{(i+1)}}{(i+1)!} \sum_{j=z}^{\infty} \frac{(\lambda t)^j (\exp(-r/\mu))^{(j-z)}}{j!} \\ &> \sum_{i=0}^{z-1} \frac{(\lambda t)^{(i+1)}}{i!} \sum_{j=z}^{\infty} \frac{(\lambda t)^j (\exp(-r/\mu))^{(j-z)}}{(j+1)!} \\ &= \sum_{i=0}^{z-1} \frac{(\lambda t)^{i}}{i!} \sum_{j=z}^{\infty} \frac{(\lambda t)^{(j+1)} (\exp(-r/\mu))^{(j-z)}}{(j+1)!} \\ &= \sum_{i=0}^{z-1} \frac{(\lambda t)^{i}}{i!} \sum_{j=z+1}^{\infty} \frac{(\lambda t)^{j} (\exp(-r/\mu))^{(j-z-1)}}{j!} \\ &= A(t,z)B(t,z+1) \end{aligned}$$

Re-arranging the inequality above, we get $A(t,z)/B(t,z) < A(t,z+1)/B(t,z+1)$, which implies that $\exp(h(t,z)/\mu) > \exp(h(t,z+1)/\mu)$ or that $h(t,z)$ decreases with z.

Recall that $h(t,z) = u(t,z) - u(t,z-1)$, $\forall z > 0$. Therefore, $\partial h(t,z)/\partial r = n(t,z-1) - n(t,z)$ (follows from Lemma 6). Since $n(t,z)$ is decreasing in z (per Lemma 5), $h(t,z)$ must be increasing in r.

[**Proof of Lemma 4**] When $Z = 0$ and $r = c$, the hurdle equals $h(T,0) = c$. In other words, all calls that provide a value of c or more are made. And, no call that provides a lower value is made. In other words, $Z = 0$ and $r = c$ is socially optimal. Since the seller gets the entire surplus in this case, it is also optimal for the seller. Note that the optimal profit equals $F = \lambda\mu T \exp(-c/\mu)$.

When $c = 0$, it is optimal to have $Z = 0$ and $r = 0$, i.e., it is optimal to offer an unlimited plan.

[**Proof of Theorem 3**]

$$\begin{aligned}&\pi(Z_1,r_1,Z_2,r_2)\\&=(1-\alpha)(u_1(Z_1,r_1)-cm_1(Z_1,r_1)+(r_1-c)n_1(Z_1,r_1))\\&-\alpha(u_2(Z_1,r_1)-u_1(Z_1,r_1))\\&+\alpha(u_2(Z_2,r_2)-cm_2(Z_2,r_2)+(r_2-c)n_2(Z_2,r_2))\end{aligned}$$

Evidently, the consumer surplus that the seller concedes to the high type is $\alpha(u_2(Z_1,r_1)-u_1(Z_1,r_1))$. This surplus does not depend on Z_2 or r_2. Consequently, the seller chooses Z_2 and r_2 to maximize the following:

$$u_2(Z_2,r_2)-cm_2(Z_2,r_2)+(r_2-c)n_2(Z_2,r_2)$$

By Lemma 4, it is optimal to have $Z_2=0$ and $r_2=c$. Moreover, if $c=0$, it is optimal to offer an unlimited plan to type 2.

Maximizing $\pi(Z_1,r_1,Z_2,r_2)$ also requires maximizing the following:

$$(1-\alpha)(u_1(Z_1,r_1)-cm_1(Z_1,r_1)+(r_1-c)n_1(Z_1,r_1))-\alpha(u_2(Z_1,r_1)-u_1(Z_1,r_1))$$

We now compute the partial derivative of the last expression with respect to r. We simplify the partial derivative using Lemma 6.

$$\begin{aligned}&\frac{\partial\pi(Z_1,r_1,Z_2,r_2)}{\partial r_1}\\&=(1-\alpha)(r_1-c)\frac{\partial n_1(Z_1,r_1)}{\partial r_1}+\frac{c(1-\alpha)n_1(T,Z_1,r_1)}{\mu g_1(Z_1,r_1)}\sum_{i=0}^{Z_1}\frac{(Z_1-i)(\lambda_1T)^i}{i!}\\&+\alpha(-n_1(Z_1,r_1)+n_2(Z_1,r_1))\end{aligned}$$

Because $n_2(Z_1,r_1)>n_1(Z_1,r_1)$ (follows from Lemma 5), it is not optimal to choose $r_1\le c$, which also implies that it is not optimal to offer an unlimited plan to type 1.

[**Proof of Theorem 4**] **Proof of the first part that assume** $\mu_2=\mu_1$: To show that it is optimal to always serve both types we need to show that there is at least one combination of Z_1 and r_1 that leads to a positive value for the following expression, which is in fact the incremental profit that the seller can get by serving the low type.

$$(1-\alpha)(u_1(Z_1,r_1)-cm_1(Z_1,r_1)+(r_1-c)n_1(Z_1,r_1))-\alpha(u_2(Z_1,r_1)-u_1(Z_1,r_1))$$

Let us choose Z_1 to be zero. Then the expression above gets reduced to:

$$\begin{aligned}&\lambda_1\mu_1T\exp(-r_1/\mu_1)+(1-\alpha)(r_1-c)\lambda_1T\exp(-r_1/\mu_1)-\alpha\lambda_2\mu_1T\exp(-r_1/\mu_1)\\=&(\lambda_1\mu_1+(1-\alpha)(r_1-c)\lambda_1-\alpha\lambda_2\mu_1)T\exp(-r_1/\mu_1)\end{aligned}$$

The last expression is positive if:

$$r_1\ge c+\mu_1\frac{\alpha\lambda_2-\lambda_1}{\lambda_1(1-\alpha)}$$

Proof of the second part that assume $\mu_2>\mu_1$: We first establish the sufficient condition for both segments to get served at optimality. When the seller serves the high type only, it sets F_2 to $\lambda_2\mu_2T\exp(-c/\mu_2)$; the corresponding profit is αF_2 or $\alpha\lambda_2\mu_2T\exp(-c/\mu_2)$.

If the seller lowers the fixed fee to $\lambda_1\mu_1T\exp(-c/\mu_1)$, both types would purchase the plan; the corresponding profit would be $(1-\alpha)F_1+\alpha F_2$ or just $\lambda_1\mu_1T\exp(-c/\mu_1)$.

It follows from the discussion above that, if $\alpha<\frac{\lambda_1\mu_1}{\lambda_2\mu_2}\exp(-(1/\mu_1-1/\mu_2)c)$, the profit generated from serving both types is at least $\alpha\lambda_2\mu_2T\exp(-c/\mu_2)$, which is the optimal profit obtained by serving the high type alone.

We now show that if α is large enough serving the low type is not optimal. And, let us again consider the incremental profit that the seller can get by serving the low type, which is as follows.

$$(1-\alpha)(u_1(Z_1,r_1)-cm_1(Z_1,r_1)+(r_1-c)n_1(Z_1,r_1))-\alpha(u_2(Z_1,r_1)-u_1(Z_1,r_1))$$

When the seller serves the low type, either $Z_1 > 0$ or r_1 is finite or both. We first show that a sufficiently high α makes serving a plan with $Z_1 > 0$ to the low type suboptimal. By Lemma 4, the maximum value of $u_1(Z_1, r_1) - cm_1(Z_1, r_1) + (r_1 - c)n_1(Z_1, r_1)$ is $\lambda\mu_1 T \exp(-c/\mu_1)$. Therefore:

$$\begin{aligned}
&(1-\alpha)(u_1(Z_1,r_1) - cm_1(Z_1,r_1) + (r_1-c)n_1(Z_1,r_1))\\
&-\alpha(u_2(Z_1,r_1) - u_1(Z_1,r_1))\\
&\le (1-\alpha)\lambda_1\mu_1 T\exp(-c/\mu_1) - \alpha(\mu_2\log(g_2(Z_1,r_1)) - \mu_1\log(g_1(Z_1,r_1)))\\
&\le (1-\alpha)\lambda_1\mu_1 T\exp(-c/\mu_1) - \alpha(\mu_2-\mu_1)\log(g_1(Z_1,r_1))\\
&\le (1-\alpha)\lambda_1\mu_1 T\exp(-c/\mu_1) - \alpha(\mu_2-\mu_1)\log\left(\sum_{i=0}^{z}\frac{(\lambda_1 T)^i}{i!}\right)\\
&\le (1-\alpha)\lambda_1\mu_1 T\exp(-c/\mu_1) - \alpha(\mu_2-\mu_1)\log(1+\lambda_1 T)\\
&< 0,\text{ if } \alpha > \frac{\lambda_1\mu_1 T\exp(-c/\mu_1)}{\lambda_1\mu_1 T\exp(-c/\mu_1) + (\mu_2-\mu_1)\log(1+\lambda_1 T)}
\end{aligned}\tag{20}$$

We now show that even when $Z_1 = 0$ a sufficiently large α leads to a negative incremental profit.

$$\begin{aligned}
&(1-\alpha)(u_1(Z_1,r_1) - cm_1(Z_1,r_1) + (r_1-c)n_1(Z_1,r_1))\\
&-\alpha(u_2(Z_1,r_1) - u_1(Z_1,r_1))\\
&= (1-\alpha)(\lambda_1\mu_1 T\exp(-r_1/\mu_1) + (r_1-c)\lambda_1 T\exp(-r_1/\mu_1))\\
&-\alpha(\lambda_2\mu_2 T\exp(-r_1/\mu_2) - \lambda_1\mu_1 T\exp(-r_1/\mu_1))\\
&= (\lambda_1\mu_1 + (r_1-c)(1-\alpha)\lambda_1 - \alpha\lambda_2\mu_2\exp((1/\mu_1 - 1/\mu_2)r_1))T\exp(-r_1/\mu_1)\\
&\le (\lambda_1\mu_1 + r_1(1-\alpha)\lambda_1 - \alpha\lambda_2\mu_2\exp((1/\mu_1 - 1/\mu_2)r_1))T\exp(-r_1/\mu_1)\\
&\le (\lambda_1\mu_1 + r_1(1-\alpha)\lambda_1 - \alpha\lambda_1\mu_2\exp((1/\mu_1 - 1/\mu_2)r_1))T\exp(-r_1/\mu_1)\\
&= (\mu_1 + r_1(1-\alpha) - \alpha\mu_2\exp((1/\mu_1 - 1/\mu_2)r_1))\lambda_1 T\exp(-r_1\mu_1)\\
&\le (\mu_1 + r_1(1-\alpha) - \alpha\mu_2(1 + (1/\mu_1 - 1/\mu_2)r_1))\lambda_1 T\exp(-r_1/\mu_1)\\
&= (\mu_1 + r_1 - \alpha\mu_2(\mu_1 + r_1)/\mu_1)\lambda_1 T\exp(-r_1/\mu_1)\\
&< 0,\text{ if } \alpha > \frac{\mu_1}{\mu_2}
\end{aligned}\tag{21}$$

If both conditions on α, i.e., the conditions mentioned in equations 20 and 21, are satisfied, it is not optimal to serve the low type.

[**Proof of Theorem 5**] If only type 1 is served, the following would hold at optimality: $F_1 = \lambda_1 T\mu_1\exp(-c/\mu_1)$, $Z_1 = 0$, and $r_1 = c$. However, if $\lambda_2 T\mu_2\exp(-c/\mu_2) > \lambda_1\mu_1 T\exp(-c/\mu_1)$, consumer type 2 would find it profitable to buy the same plan.

Therefore, if $\lambda_2\mu_2 T\exp(-c/\mu_2) > \lambda_1\mu_1 T\exp(-c/\mu_1)$, at optimality either type 2 or both types are served.

We now argue that, if $\lambda_2\mu_2 T\exp(-c/\mu_2) > \lambda_1\mu_1 T\exp(-c/\mu_1)$, at optimality IR2 and IC1 cannot simultaneously bind. If both bind, the profit function is going to be as follows:

$$\begin{aligned}
&\pi(Z_1,r_1,Z_2,r_2)\\
&= (1-\alpha)(u_1(Z_1,r_1) - cm_1(Z_1,r_1) + (r_1-c)n_1(Z_1,r_1))\\
&-(1-\alpha)(u_1(Z_2,r_2) - u_2(Z_2,r_2))\\
&+\alpha(u_2(Z_2,r_2) - cm_2(Z_2,r_2) + (r_2-c)n_2(Z_2,r_2))
\end{aligned}$$

It follows from Lemma 4 that $F_1 = \lambda_1\mu_1 T\exp(-c/\mu_1)$, $Z_1 = 0$, and $r_1 = c$ at optimality (because optimizing the profit function requires maximizing $u_1(Z_1,r_1) - cm_1(Z_1,r_1) + (r_1-c)n_1(Z_1,r_1)$). However, IR2 cannot bind in this case because type 2 can get a positive surplus by buying type 1's plan.

Moreover, at optimality at least one of the two IR conditions bind (because the seller can keep raising the fixed fees simultaneously, and by the same amount, until one of them binds). Similarly, either IR1 binds or IC1 binds (because the seller can raise F_1 until one of them binds); and, either IR2 binds or IC2 binds (because the seller can raise F_2 until one of them binds). Therefore, because IR2 and IC1 cannot simultaneously bind at optimality, either, IR1 and IR2 must simultaneously bind, or IR1 and IC2 must simultaneously bind. If the IR1 and IR2 conditions bind, the profit function is going to look as follows:

$$\begin{aligned}
&\pi(Z_1,r_1,Z_2,r_2)\\
&= (1-\alpha)(u_1(Z_1,r_1) - cm_1(Z_1,r_1) + (r_1-c)n_1(Z_1,r_1))\\
&+\alpha(u_2(Z_2,r_2) - cm_2(Z_2,r_2) + (r_2-c)n_2(Z_2,r_2))
\end{aligned}$$

By Lemma 4, the maximum value of $(u_2(Z_2,r_2) - cm_2(Z_2,r_2) + (r_2 - c)n_2(Z_2,r_2))$ is $\lambda_2\mu_2 T\exp(-c/\mu_2)$. Similarly, the maximum value of $(u_1(Z_1,r_1) - cm_1(Z_1,r_1) + (r_1 - c)n_1(Z_1,r_1))$ is $\lambda_1 T\mu_1\exp(-c/\mu_1)$. Therefore, the optimal F_2 is $\lambda_2 T\mu_2\exp(-c/\mu_2)$, and the optimal F_1 is $\lambda_1\mu_1 T\exp(-c/\mu_1)$. However, if $\lambda_2\mu_2 T\exp(-c/\mu_2) > \lambda_1\mu_1 T\exp(-c/\mu_1)$, it becomes incentive compatible for type 2 to buy type 1's plan. In other words, IR1 and IR2 cannot simultaneously bind.

If the IR1 and IC2 conditions bind, the profit function is going to look as follows:

$$\begin{aligned}
&\pi(Z_1,r_1,Z_2,r_2)\\
&= (1-\alpha)(u_1(Z_1,r_1) - cm_1(Z_1,r_1) + (r_1 - c)n_1(Z_1,r_1))\\
&- \alpha(u_2(Z_1,r_1) - u_1(Z_1,r_1))\\
&+ \alpha(u_2(Z_2,r_2) - cm_2(Z_2,r_2) + (r_2 - c)n_2(Z_2,r_2))
\end{aligned}$$

Recall that the maximum value of $(u_2(Z_2,r_2) - cm_2(Z_2,r_2) + (r_2 - c)n_2(Z_2,r_2))$ is $\lambda_2\mu_2 T\exp(-c/\mu_2)$. To prove that serving both types is optimal we therefore need to show that there is at least one combination of Z_1 and r_1 that leads to a positive value for:

$$(1-\alpha)(u_1(Z_1,r_1) - cm_1(Z_1,r_1) + (r_1 - c)n_1(Z_1,r_1)) - \alpha(u_2(Z_1,r_1) - u_1(Z_1,r_1))$$

Let us choose Z_1 to be zero. When $\lambda_2 > \lambda_1$ and $\mu_2 < \mu_1$, the profit from serving type 1 is at least:

$$\begin{aligned}
&\lambda_1\mu_1 T\exp(-r_1/\mu_1) + (1-\alpha)(r_1 - c)\lambda_1 T\exp(-r_1/\mu_1) - \alpha\lambda_2\mu_2 T\exp(-r_1/\mu_2)\\
\geq&\lambda_1\mu_1 T\exp(-r_1/\mu_1) + (1-\alpha)(r_1 - c)\lambda_1 T\exp(-r_1/\mu_1) - \alpha\lambda_2\mu_1 T\exp(-r_1/\mu_1)\\
=&(\lambda_1\mu_1 + (1-\alpha)(r_1 - c)\lambda_1 - \alpha\lambda_2\mu_1)T\exp(-r_1/\mu_1)
\end{aligned}$$

The last expression is positive if $r_1 \geq c+\ \mu_1\frac{\alpha\lambda_2-\lambda_1}{\lambda_1(1-\alpha)}$.