Submitted to
manuscript (Please, provide the mansucript number!)

Cost-per-Impression Pricing and Campaign Delivery for Online Display Advertising

Pricing and capacity management represent significant challenges for web publishers that generate revenues by selling advertising space on their websites. Advertisers approach a publisher to book an advertising campaign, requesting a number of impressions to be delivered regularly throughout the campaign duration. Publishers offer multiple advertising plans. They face uncertainty in demand and supply, which generates non-uniformity in the campaign delivery. Based on a stylized model of the publisher's operation, we suggest a capacity allocation mechanism parameterized by a display frequency that allocates viewers to ads in a rotating manner. Through a large-capacity system analysis and under the suggested mechanism, we prove that the fluid price and display frequency are asymptotically optimal. We also obtain correction terms for the fluid solution when used under a regular regime. The pricing and display frequency can be translated into inputs to delivery engines used in practice. We obtain data from a publisher and perform an extensive numerical analysis that reveals the interrelation between prices, traffic load, impressions, and display frequency.

Key words: Online advertising; capacity management; pricing; asymptotic analysis.

1. Introduction

Online advertising has been a fast growing area within the media industry with 26 billion dollars revenues in 2010 (IAB (2011)). The Internet, with its access to an enormous consumer base, remains a very attractive media to advertisers and offers many different display possibilities compared to traditional media. Web publishers providing content and services commonly use advertising as the main revenue source for their businesses instead of charging usage or subscription fees. The two largest areas of online advertising are display advertising and sponsored search advertising. In the latter, pricing is well defined with auctions as the main mechanism. However, pricing in display advertising is often ad-hoc and could benefit from systematic approaches. A very common pricing scheme in display advertising is the pay-per-impression scheme[1] where the advertiser pays for each time his ad is displayed to a visitor.

In display advertising, most web publishers use delivery engines such as Dart by DoubleClick to deliver the contracts made with advertisers. However, the engines require important inputs from the user (such as the frequency of display) that are not easy to determine. Furthermore, the

[1] Other pricing schemes are i) Pay-per-click where the advertiser pays for each time a visitor clicks on his ad. ii) Pay-per-action where the advertiser pays only if the visitor purchases the product being advertised.

capacity management task of the engines is often disconnected from pricing decisions leading to suboptimal results.

Web publishers continuously face uncertain demand from advertisers (or agencies acting on their behalf) and are not (as in, e.g., TV broadcasting), restricted by a specific season or horizon. In this paper, we consider a capacity management problem from a continuous-time infinite horizon point of view. Advertisers want a certain number of viewers to see their ad, which needs to be matched with uncertain supply from viewers visiting the website. We consider contracts between publishers and advertisers that specify the price charged per impression, the number of impressions to be delivered, and the campaign duration. An unwritten rule, rarely included in a contract yet of great concern to the advertiser, requires the impressions to be delivered *regularly* throughout the campaign horizon. Our model explicitly accounts for this rule.

The main contributions of the paper are the following. First, we suggest a stylized, infinite-horizon model based on continuous time dynamics that is relevant for web publishers. The model captures operational challenges related to supply and demand uncertainties, regular delivery, and time constraints. It allows one to study the interconnection between the typical drivers of the online problem; price per impression, capacity allocation, and display frequency, on one hand, and the campaign delivery and the uncertainty generated, on the other. Second, we obtain from an analysis performed on large-capacity systems, a simple and effective policy (price per impression, capacity allocation, and display frequency) that reduces the impact of uncertainty and maximize revenues. The allocation mechanism used to obtain the solution relies on matching ads and viewers in a rotating manner. The solution itself could be the input to delivery engines used in practice. In particular, we show that the suggested pricing policy and the allocation mechanism (induced by the solution to the deterministic/fluid problem) are asymptotically optimal. Third, from a methodology point of view, in large-scale operations, a balanced loading setting is proven to be economically optimal causing little congestion and irregularity effects. This methodology is extended to multiple types of advertising plans. Fourth, in the online advertising media, publishers are able to collect a large amount of data on their advertising operations. The complexity of their operational problems makes it hard to take full advantage of such data sets. However, relying on the model developed in this paper, enriched with two demand models adapted to the online setting, and based on data from the Scandinavian web publisher Aller Internett, we present an extensive numerical analysis that gives insights into the interrelation between design parameters (e.g., display frequency, size of the contract) and the system's performance (e.g., delivery shortage).

The paper is organized as follows. In the next section we review the literature. In Section 3 we introduce the model ingredients and provide a discussion on the main assumptions. In Section 4 we set up the problem, formulate the non-uniformity cost that induces uniform service delivery, and solve the fluid version of the problem, which motivates the allocation mechanism chosen. In Section 5 we analyze the single advertising plan under this suggested allocation mechanism and obtain an asymptotically optimal solution of the optimization problem. This asymptotic analysis implies an approximation of the non-uniformity cost, which we analyze in detail. In Section 6 we extend the single advertising plan case to the multi-plan setting. Section 7 is devoted to a numerical analysis based on data from the Scandinavian web publisher. In the concluding section we list some further research questions.

2. Literature Overview

The work presented in this paper is related to the literature on concurrent capacity and pricing management. It involves many aspects of a revenue management problem where a finite capacity (in terms of viewers per unit of time as well as advertising slots) needs to be allocated among different advertisers' campaigns. A comprehensive reference of revenue management models and applications is the book by Talluri and van Ryzin (2004). In many ways the problem we study is different from the typical revenue management setting. First, the supply is uncertain and provided over a specified horizon. Second, in the online setting where the duration can last for weeks and slots might be shared by multiple advertisers, the notion of a dedicated resource is harder to define, especially in the presence of both demand and supply uncertainty. The setting we consider makes our work closer to the literature on capacity management using queueing systems techniques. Such approach was used previously in the context of a revenue management. The paper of Savin et al. (2005) models the rental car problem as a multi-server queueing system with a continuous stream of customer arrivals having independent and exponentially distributed rental times. Many differences exist with that model starting with the supply dynamics mentioned above. Furthermore, that paper considers two classes of customers with accept-reject type control policy. We consider multiple types of advertising plans and allow campaigns to incur delays. A more recent work using such a continuous time approach in online advertising is that of Radovanovic and Zeevi (2009). They consider the allocation of advertisers' campaigns to a set of products based on specified budget and effectiveness. They suggest an asymptotically optimal allocation driven by an LP solution. The problem specifics are different from ours. In particular, Radovanovic and Zeevi (2009) do not consider any pricing control and are not explicitly constrained by a number of impressions or campaign duration.

Another stream of literature in the queueing context relevant to ours is the one using large-capacity systems in the so-called Halfin-Whitt regime (see Halfin and Whitt (1981)). This type of setting has been used extensively in various applications, in particular in call centers (see Gans et al. (2003) for an overview). In the context of pricing and capacity sharing, both Maglaras and Zeevi (2003) and Maglaras and Zeevi (2005) provide an equilibrium analysis determining the demand rate and among other results obtain approximations for the optimal solution through "large-capacity asymptotics". The asymptotic analysis they undertake is similar in nature to the one we perform in the single plan case (where the "heavy-traffic" regime is shown to be optimal from an "economic optimization" point of view). Nevertheless, both the system we study and its analysis remain different from theirs. In particular, we look at capacity allocation decisions under fixed total capacity, while they consider capacity sizing. We discuss these differences more in detail in Section 5.2.

Some aspects of pricing and capacity management in TV broadcasting are similar to the online advertising case (see Araman and Popescu (2010) and Bollapragada and Mallik (2008)). These papers also consider the supply to be uncertain as well as the number of contracted impressions needed to be met through accumulation of viewers seeing the ad. However, in TV broadcasting, the problem is naturally set as a finite horizon with two channels: the upfront demand, which is contracted at the beginning of the horizon and the scatter market demand that gets realized throughout the horizon. The paper that is possibly the closest to our paper in terms of *context* is Roels and Fridgeirsdottir (2009) who consider dynamic admission control and delivery of advertising contracts over a finite horizon. The problem is formulated as a dynamic program and a Certainty Equivalent Control heuristic is proposed and tested.

Finally, there is a large body of literature on online advertising in general. We refer the reader to Ha (2008) who provides an overview of online advertising research in advertising journals and Evans (2008) summarizes the economics of the online advertising industry. Scheduling of online ads is one of the most popular topics within the operations research literature. The work by Kumar et al. (2006) provides a good overview on that stream of literature.

3. Modeling Framework

We consider a web publisher that generates revenues by posting ads on its website. The publisher offers J different advertising plans that the advertisers can choose from. A plan j $(1 \leq j \leq J)$, is defined by the number of viewers, N_j, that should see the ad (number of impressions) during a period of T_j (campaign duration) with the price p_j charged per impression. The uncertainty present

in the system does not allow the publisher to commit to an exact number of impressions N_j and an exact duration T_j. We observe in practice two types of contracts that we address in the paper.

N-Contracts. The publisher can commit to a certain number of impressions, N_j, and the ad is displayed until that number is met. The publisher then designs the system so that the *expected* campaign duration is T_j. We call this contract the *N-contract*.

T-Contracts. The publisher keeps the ad in the system for *exactly* T_j units of time, after which the ad is removed. The publisher designs the system so that the *expected* number of viewers that will see the ad during T_j is N_j. We call this contract a *T-contract*. In the T-constract, the advertiser will only pay for the number of impressions collected at the end of the campaign, which is N_j in expected value.

Both types of contracts can be implemented in practice in delivery engines such as Dart by DoubleClick. The publisher offers either of the two types. Note that we are slightly abusing the notation by letting N_j and T_j represent either the expected value or the exact value. However, the context should clarify which representation is being used.

3.1. The Demand

The web publisher is approached continuously by advertisers (or agencies acting on their behalf) requesting advertising campaigns. The advertisers choose one of the plans offered by the web publisher and as a result their ad is then displayed during a certain period of time and shown to a certain number of viewers. One typical form of demand realization is when advertisers place an order for an advertising campaign through an online platform, the same way one can rent a car or reserve a hotel room. We assume the demand to follow a Poisson process, which depicts the number of advertisers requesting a campaign through the website. This is a realistic assumption in the online platform setting (see, e.g., Radovanovic and Zeevi (2009)) and a common assumption in service settings in general (see, e.g., Savin et al. (2005)). More specifically, we let $(v_{i,j} : i \geq 1, 1 \leq j \leq J)$ be an i.i.d. sequence of interarrival times (the times separating two campaign requests for the same plan), which are exponentially distributed with mean $1/\lambda_j$.

3.2. The Supply

The supply for the web publisher's advertising operation consists of the viewers visiting the website. A website often consists of multiple webpages that viewers can either visit directly or through links from the homepage. The overall traffic, to which ads can be displayed to (often referred to as the inventory), is the combined traffic from all webpages belonging to the website. We formulate the aggregated traffic of viewers visiting all the webpages and we assume it to be a Poisson process with

rate μ (see Gong et al. (2005) who argue that traffic of web servers could be well approximated by a Poisson process). Without loss of generality, the reader can think of the website as being made of one single webpage, which viewers visit according to a Poisson process with an aggregated rate μ. We let $(u_i : i \geq 1)$ be the i.i.d. sequence of interarrival times of viewers.

3.3. The Service Procedure

Every page of a website is assumed to have s advertising slots. The advertisers pay for each time a visitor uploads the webpage while their ad is on display. Every visit counts as one impression towards the total count of N_j impressions. Furthermore, every time a viewer uploads the page the number of impressions delivered goes *simultaneously* up by one for *all* ads posted.

3.4. Summary of the Main Modeling Assumptions

Tactical vs. Upfront Contracts. Our investigation shows that publishers face two types of contracts. "Upfront" contracts are those long term bookings placed by few but major advertisers that are negotiated and displayed over a long period of time such as a year. They are in many ways similar to TV broadcasting contracts. For such contracts, the publisher reserves a capacity and is not impacted much by daily uncertainties. The other types of contracts that the publisher face are often known as "tactical" contracts and can be booked at any time. In TV broadcasting, most of the capacity is sold upfront and the remaining is left for the so-called "scatter market". In the online world, the tactical contracts represent a large portion and are more challenging to manage with possible missed sales due to capacity shortages. In this paper we focus on tactical contracts.

Targeting. Advertisers sometimes specify a few characteristics of the viewers they would like their ads to be displayed to (e.g., a male interested in sport). Overall, targeting is becoming a major characteristic of ad-networks/exchanges. It is a complex problem, beyond the scope of this paper. We restrict our attention to advertisers that are targeting the same pool of viewers. This is still a rich and common setting in practice. In the data set we analyze from Aller Internett, the publisher does not offer any targeting as the online magazines it runs attract a well defined audience.

Negotiations. We do not model contract negotiations but capture, at an aggregated level, the uncertainty of demand, some of which could be the result of negotiations.

4. Problem Formulation

We consider a web publisher that maximizes the profit rate in steady-state while meeting the advertisers' contract requirements. Given, the number of impressions, N_j, and the duration of the campaigns, T_j, the publisher decides on the price, p_j, per impression to charge for an advertising plan j, and on a policy, π, of how to allocate the viewers to the campaigns.

In practice, every time a viewer visits the website, the publisher decides which ads to display, depending on how many campaigns are currently unfulfilled, the number of impressions already allocated to each, and the remaining campaign durations. In addition, the publisher aims at delivering a campaign's impressions in a *regular* way (an advertiser is not keen on having, e.g., all his impressions delivered in the beginning of the campaign horizon, leaving only few towards the end). In all its generalities, this is an untractable dynamic and stochastic optimization problem. Delivery engines deal with this problem using some optimization techniques but require many inputs from the advertiser. Our objective is to model the problem at a tactical level and focus on the link between the pricing and campaign delivery components. The output of our model can help advertisers setting inputs for their delivery engine.

Revenues. The web publisher collects revenues for each impression made with total revenues of $p_j N_j$ per campaign requesting plan j. We denote by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_J)$ the vector of demand intensities generated when the advertiser sets the vector of prices $\mathbf{p} = (p_1, p_2, ..., p_J)$ for the J plans available. We let $\mathbf{N} = (N_1, N_2, ..., N_J)$. We assume the demand function, $\boldsymbol{\lambda(p)}$, has an inverse $\boldsymbol{p(\lambda)}$ and that the revenue rate $r_j(\boldsymbol{\lambda}; \boldsymbol{N}) := \lambda_j \, p_j(\boldsymbol{\lambda}, \boldsymbol{N}) N_j$ is concave in $\boldsymbol{\lambda}$, which is a common assumption in the literature (see, e.g., Gallego and van Ryzin (1997)).

Fulfillment Constraint. In the case of a T-contract the publisher keeps the ad in the system for exactly T_j units of time during which, on average N_j viewers see the ad. In the case of an N-contract, the ad is displayed N_j times with the constraint that the expected campaign duration is T_j. We formulate these constraints later in this section.

In addition to the fulfillment constraint, the advertisers expect the impressions to be delivered regularly throughout the campaign horizon. The reasons for this requirement are numerous: *i)* an online campaign is often a part of a campaign across multiple media that the advertiser is investing in and the advertiser expects that these advertising efforts are synchronized. *ii)* There are empirical studies (see Chatterjee et al. (2003)) that show that the impact of a campaign on one viewer is greater when it is spread out than when it is condensed in time. That is even more relevant in our case, as we do not (similarly to many web publishers) track unique viewers. *iii)* A regular campaign is less biased to the specific content of the website and thus the ad has less chance to be associated with a specific event addressed on the web site that occurs at some point during the campaign duration.

For modeling purposes, we introduce a cost of non-uniformity as a way to capture the regularity agreement and to avoid extreme policies. We recognize that some non-uniformity is acceptable, but such cost creates an incentive for the publisher to spread the campaign as much as possible

throughout its duration. We also argue below that this cost measures the uncertainty present in the system and thus allows the publishers to balance between the level of uncertainty and the expected revenues.

Non-Uniformity Cost. During a campaign length T_j, a certain number of viewers, $n(T_j)$, visit the website with an average of μT_j. We number these viewers from 1 to $n(T_j)$. As these viewers arrive, a policy π selects N_j of them to meet the campaign requirement. We denote by A_i^π the position of the i^{th} viewer allocated to the campaign according to policy π. We denote by $\nu_i^\pi = A_{i+1}^\pi - A_i^\pi$, the gap between viewers i and $i+1$ that are allocated to the campaign. If we want to deliver N_j impressions evenly in T_j, then ν_i^π should be exactly equal to $\mu T_j/N_j$. The non-uniformity cost ought to measure how uneven the delivery of viewers to a specific campaign is and we measure it as the deviation from the ideal allocation of $\mu T_j/N_j$. For that purpose we use a normalized mean squared error type cost given by

$$C_j^\pi = \frac{c_j}{\mu T_j/N_j}\Big(\sum_{i=1}^{N_j}\big(\mathbb{E}\nu_i^\pi - \mu T_j/N_j\big)^2\Big)^{1/2}, \tag{1}$$

where c_j is a positive constant. The main results of the paper hold under a general form of the cost function. In the next section we show that the cost of non-uniformity can be avoided in the deterministic setting. However, the presence of uncertainty in both demand and supply results inevitably in an irregular campaign delivery measured by the non-uniformity cost. This cost increases as the uncertainty increases. Therefore, under a regularity requirement, the cost of non-uniformity accounts for the impact of uncertainty in the optimization problem.

We move now to state the generic optimization problem. As mentioned before the publisher's goal is to determine the price, p_j, per impression to charge and the service policy, π, to use that maximizes the profit given the contractual agreement with the advertiser. As we have a one-to-one relationship between $\boldsymbol{\lambda}$ and $\boldsymbol{p}$ we choose to determine the optimal $\boldsymbol{\lambda}$. The optimization problem can be stated as follows for the T-contract:

$$\begin{aligned} &\max_{\lambda,\,\pi} \sum_{j=1}^{J} r_j(\boldsymbol{\lambda}, \boldsymbol{N}) - \lambda_j C_j^\pi && (P_T)\\ &s.t.\ \ \mathbb{E}[\#Impressions_j^\pi] = N_j \qquad 1 \le j \le J. \end{aligned}$$

The optimization problem for the N-contract has the same objective function as the one above with the constraint $\mathbb{E}[Duration_j^\pi] = T_j$. We denote that problem by P_N.

The exact formulation for the non-uniformity costs and the constraints depend on the policy π adopted. We formulate those in Section 5.

4.1. Fluid Problem

To gain an insight into the solution structure of our problem, we analyze the case where uncertainty is disregarded, which we denote by the fluid problem. The advertiser sets a price, $\boldsymbol{p}^0$, which determines the rate, $\boldsymbol{\lambda}^0$, at which campaigns are requested in a way that the system remains stable, i.e., $\sum_{j=1}^{J} \lambda_j N_j \leq s\,\mu$, where s is the number of advertising slots. We denote by $\rho := \sum_{j=1}^{J} \lambda_j N_j/(s\,\mu)$ the publisher's utilization. We suggest a policy π^0, which we prove to be optimal in this setting. The policy π^0 is structured such that campaigns requiring the same plan j are grouped together. A fixed proportion of viewers, f_j^0, is then directed to each group of campaigns belonging to plan j with $\sum_{j=1}^{J} f_j^0 \leq 1$. Hence, campaigns requesting plan j are allocated viewers at a rate μf_j^0. In order to satisfy the fulfillment constraint with zero non-uniformity cost, each ad is not shown to every viewer, rather it is shown to every $\kappa_j^0 = \mu f_j^0\,(T_j/N_j)$ viewer (i.e., with a display frequency of $1/\kappa_j^0$). In the absence of uncertainty, this policy ensures uniform delivery as long as there are exactly $s\kappa_j^0$ campaigns requested at any point in time. Note that the number of campaigns requesting plan j at any point in time is exactly equal, by Little's law, to $\lambda_j^0 T$. By letting $f_j^0 = \lambda_j^0 N_j/(s\mu)$, we get that $\lambda_j^0 T = s\mu f_j^0 T/N_j = s\kappa_j^0$.

In conclusion, the policy π^0 guarantees in a fluid setting that the fulfillment constraint is met and avoids any cost of non-uniformity. It remains to solve for $\boldsymbol{\lambda}^0$, which is the solution to the following concave optimization problem

$$\begin{aligned} &\max_{\boldsymbol{\lambda}} \sum_{j=1}^{J} \lambda_j\, p_j(\boldsymbol{\lambda}; \mathbf{N})\, N_j \\ &s.t.\, \rho = \sum_{j=1}^{J} \lambda_j\, N_j/(s\,\mu) \leq 1. \end{aligned} \tag{MP0}$$

The following proposition summarizes the solution to the optimal solution in the fluid setting. We denote by $\bar{\boldsymbol{\lambda}}$ the maximizer of the revenue function, $\sum_{j=1}^{J} r_j(\boldsymbol{\lambda}, \boldsymbol{N}) = \sum_{j=1}^{J} \lambda_j\, p_j(\boldsymbol{\lambda}; \mathbf{N})\, N_j$.

Proposition 1 *We consider the fluid case where exactly every $1/\lambda_j$ time unit an arrival of a campaign requesting plan j occurs and exactly every $1/\mu$ time unit an arrival of a viewer occurs. In this case, the publisher should follow policy π^0 defined above with the parameters:*

$$f_j^0 = \frac{\lambda_j^0 N_j}{s\,\mu} \qquad \textit{and} \qquad \kappa_j^0 = \frac{\mu f_j^0 T_j}{N_j},$$

and the arrival rates such that

i.) if $\sum_{j=1}^{J} \bar{\lambda}_j N_j/(s\mu) \leq 1$, then $\boldsymbol{\lambda}^0 = \bar{\boldsymbol{\lambda}}$,

ii.) otherwise, for all $j \in \mathcal{J}$, we solve for the unique $\boldsymbol{\lambda}^0$ that satisfies $\sum_{j=1}^{J} \partial_{\lambda_i} r_j(\boldsymbol{\lambda}; \boldsymbol{N}) = \frac{m N_i}{s\mu}$, where m is a constant uniquely determined by $\sum_{j=1}^{J} f_j^0 = 1$.

All proofs are provided in Appendix A. Practically, the allocation mechanism π^0, suggested above, implies that multiple ads would share the same advertising slots. This takes advantage of the fact that the total arrival rate of viewers to a website, μ, is much larger than the rate, N_j/T_j, at which a campaign should be delivered in order to satisfy the fulfillment constraint.

In the absence of uncertainty, the previous solution provides a relationship between, the price per impression or demand rate λ_j^0, each plan's share of capacity, f_j^0 and the display frequency, $1/\kappa_j^0$. This solution clearly generates the best possible revenues and incurs zero non-uniformity cost. Therefore, it will be considered as a benchmark for the stochastic setting.

4.2. Partitioned Uniform Allocation Policy

When uncertain demand and supply are present in the system, a non-uniformity cost cannot be avoided no matter which static policy, π, is implemented. Furthermore, determining an optimal policy across all policy classes does not seem tractable. Our goal is to suggest a policy that is simple to model and performs well in terms of trading off revenues and cost of non-uniformity. We first propose a class of policies inspired by the fluid problem discussed above and then determine the optimal policy within that class.

We restrict ourselves to a class of policies π defined through a set of parameters (f_j, κ_j) assigned to each plan j, $1 \leq j \leq J$. As in the fluid setting, the first parameter f_j, is the share of capacity (in terms of viewers) that the web publisher directs to all campaigns requesting plan j. The second parameter κ_j regulates the frequency of display of these campaigns. We call this class of policies *partitioned uniform allocation* (PUA). By following such policy, the dynamics are decoupled in the sense that all campaigns requiring plan j are allocated a Poisson process of viewers with rate μf_j. Every viewer of that process is directed to a set of s ads in a rotating manner so that if viewer i sees one set of s ads, the next viewer, $i+1$, sees another set of s ads and so on; until viewer $i+\kappa_j$, who sees the same set of ads as viewer i, and the cycle starts again. This way, campaigns are uniformly spread receiving one impression for every κ_j viewers directed to plan j.

We call a campaign belonging to plan j *active*, if it is part of the $s\kappa_j$ campaigns that viewers are being directed to in a rotating manner. In the stochastic setting, the number of active campaigns varies through time. If this number drops below $s\kappa_j$, then the web publisher, in order to keep the constant pace, complements the active campaigns by showing viewers filler ads (e.g., Yahoo website showing Yahoo ads or ads for non-profit organizations). If the number of active campaigns is already at $s\kappa_j$, then every *additional* campaign requested has its starting time delayed (set to *passive*). Passive campaigns become active in a first-come-first-served manner when currently

active campaigns are completed. The solution to the optimization problem should guarantee that this delay is acceptable.

In summary, each campaign experiences first a small lag (or delay) W_j (possibly none) during which no viewers are allocated to it, then when one of the $s\kappa_j$ slots is available it becomes active. Active campaigns of plan j are allocated viewers at a rate μf_j. Each viewer is then directed to a specific set of s campaigns at a rate κ_j. Any other mechanism (even dynamic) that one could implement in practice would generate inevitably irregularity in the display. The aggregation of the non-uniformity generated by a mechanism throughout the campaign duration is to be compared to that initial lag that PUA type mechanisms generate. Because of its quadratic structure, the cost of non-uniformity penalizes more PUA type mechanisms than ones with more dispersed non-uniformity. Consequently, the solution (in terms of pricing, display frequency and capacity allocation) obtained under a PUA mechanism is expected to be more conservative than other mechanisms that aim for regular delivery. In addition, we show below that the PUA mechanism is optimal under large-scale systems. Finally, we note that this allocation mechanism has similarities with the so-called "partitioned nesting" policy that is used in yield management where the seller reserves a fraction of its total capacity to each class of customers.

5. Single Plan Under PUA Policy

In this section, we assume that the publisher offers only one type of a plan (i.e. $J=1$ and $f_1=1$). We focus on solving the single plan version of problems P_T and P_N. Let us first formulate the fulfillment constraint, which guarantees that N impressions are collected during time T. We let $\varpi = \mathbb{E}W$ and refer to it as the delay. (A superscript T and N will be added as necessary to differentiate between the two contracts). We denote by $n(t)$ the number of viewers that visit the website in $(0,t)$, which is a Poisson random variable with mean μt. In the T-contract case, the ad will be posted during $T-W$, which is the *effective* campaign duration and will be displayed to the viewers at a frequency of $1/\kappa$. Thus the expected number of viewers that see the ad in $T-W$ is $(\mu/\kappa)(T-\varpi^T)$. Hence, the fulfillment constraint is given by, $N=(\mu/\kappa)(T-\varpi^T)$, equivalently, $\varpi^T = T - N\kappa/\mu$. In the N-contract case, the effective campaign duration is $\sum_{i=1}^{N\kappa} u_i$, where u_i's are the interarrival times of viewers. The contract requires $\mathbb{E}\sum_{i=1}^{N\kappa} u_i + \varpi^N = T$, which leads to $\varpi^N = T - N\kappa/\mu$. Therefore, the form of the fulfillment constraint is the same for both contracts, even though the formulation of the delay is different. By allowing κ to be any real number, both constraints will be proven to admit a unique solution.

The objective function is made of two terms, the revenue function and the cost of non-uniformity discussed earlier. The number of viewers that arrive since a campaign is requested and until it

is allocated its first viewer, is $n(W)+\kappa$, where $n(t)$ is the Poisson counting process of viewers during time t as defined above. The other $N-1$ viewers are uniformly allocated to that campaign. The non-uniformity cost generated by a campaign delivery is the measure of the deviation from a uniform allocation. It can be formulated by recalling Equation (1) and the fact that the fulfillment constraint requires on average, every $\kappa^0=\mu T/N$ viewers to be allocated to an ad (which is exact in the fluid setting).

$$\begin{aligned} C &= c\Big(\big(\frac{\mathbb{E}\,n(W)+\kappa_j-\kappa^0}{\kappa^0}\big)^2+(N-1)\big(\frac{\kappa-\kappa^0}{\kappa^0}\big)^2\Big)^{1/2} \\ &= c\Big(\big(\mu\varpi/\kappa^0-(1-\frac{\kappa}{\kappa^0})\big)^2+(N-1)\big(1-\frac{\kappa}{\kappa^0}\big)^2\Big)^{1/2} = c\frac{(N(N-1))^{1/2}}{T}\,\varpi. \end{aligned}$$

The main results of this paper hold for more general costs of non-uniformity, as they translate into a function of the moments of W. For clarity of exposition, we focus in the rest of this paper on this mean-squared error type cost, linear in ϖ. From now on, it is essential to think of the delay as an aggregated measure of how spread out a campaign delivery is. As the number of impressions, N, is large, we let $C=c\,\frac{N}{T}\,\varpi$. Putting the optimization components together, the publisher solves for:

$$\begin{aligned} &\max_{\lambda,\,\kappa}\Big\{\lambda\,p(\lambda;N)\,N-\frac{c\,N}{T}\lambda\,\varpi(\lambda,\kappa)\Big\} \\ &\quad s.t.\;\; \varpi(\lambda,\kappa)\big)=T-N\kappa/\mu \;\;\text{ and }\;\; \rho=\lambda N/(s\mu)\le 1. \end{aligned} \tag{P}$$

The utilization ρ measures how loaded the system is. It is independent of κ. This is intuitive as κ is a parameter of the allocation mechanism, which primary role is to pace and allocate the total capacity among the different campaigns of the same plan. That said, both the allocation mechanism chosen and the contract offered do affect the utilization level through the optimal pricing policy they generate. We introduce another load factor that turns out to play a critical role in the dynamics of the problem. We define the ratio $\varrho=\lambda T/s\kappa$ and call it the *congestion factor*. It is the ratio between the average number of campaigns that are currently booked and not yet fulfilled and the maximum possible number of active campaigns $s\kappa$. It represents another measure of the load of the system, which depends on both κ and λ and can be smaller or larger than one. Using both load factors, we re-write the equality constraint as $\varpi(\lambda,\kappa)=T(1-\rho/\varrho)$. This relationship implies that meeting the number of impressions during time T imposes a trade-off between the non uniformity of the campaign and the two load factors. In particular, it requires that $\rho\le\varrho$ or equivalently $\kappa\le\kappa^0$.

Next, we formulate and analyze in details the delay. We then solve Problem (P) for large systems and obtain a limiting solution, which corresponds to the fluid solution and hence shows that the uniform allocation mechanism suggested is asymptotically optimal.

5.1. The Non-Uniformity Cost

One of the main contributions of this paper is the analysis of the impact of demand and supply uncertainties on the web publisher. These uncertainties affect the web publisher through two interrelated dimensions: *i.*) the N contracted impressions might not be delivered (in the case of the T-contract - duration T might not be met for the N-contract), *ii.*) the publisher might not be able to maintain a uniform delivery of the impressions. Under the uniform allocation mechanism suggested above, the impact of these two dimensions are captured through the non-uniformity cost, which is proportional to the delay. The delay therefore quantifies the uncertainty in the system and measures its impact. Moreover, it allows one to measure the cost of rejecting advertisers if delays are not accepted in practice (i.e., number of rejected advertisers could be estimated to be $\lambda\varpi$.)

The next result gives an expression of the delay for both T-contracts and N-contracts.

Proposition 2

i.) Under a T-contract, the delay is given by

$$\varpi^T(\lambda,\kappa) = \mathbb{E}\big[T - \sum_{i=1}^{s\kappa} v_i\big]^+,$$

where v_i's are the interarrival times of the advertisers.

ii.) Under an N-contract, the delay is given by

$$\varpi^N(\lambda,\kappa) = \mathbb{E}\max_{n\geq 0}\sum_{i=1}^{n} X_i(\lambda,\kappa),$$

where the sequence $(X_1, X_2, ...)$ is i.i.d. with $X_1 \stackrel{d}{=} \sum_{i=1}^{N\kappa} u_i - \sum_{i=1}^{s\kappa} v_i$, and u_i's and v_i's are the interarrival times of the viewers and the advertisers, respectively.

iii.) No matter which type of contract is used, the delay satisfies the following:

$$\text{for all } \kappa \geq 1, \quad \varpi(\kappa) \leq \varpi(1) \qquad \text{and} \qquad \varpi(\kappa) \to 0, \quad \text{as } \kappa \to \infty.$$

The results stated in Proposition 2 give a precise formulation of the delay a campaign incurs under PUA. The optimization strives to keep the cost of non-uniformity that depends on the delay to a minimum. The main property that enables us to derive this formulation is the fact that campaigns are fulfilled in the order they were booked, which results directly from the allocation policy adopted. It is possible to get a closed form formulation for the delay.

Corollary 1 *Under a PUA mechanism, we have that*

$$\varpi^T(\lambda,\kappa) = Te^{-\lambda T}\sum_{j=s\kappa}^{\infty}\big(1 - \frac{s\kappa}{j+1}\big)\frac{(\lambda T)^j}{j!}. \tag{2}$$

A similar explicit formula as in Corollary 1 can be found for the N-contract (see page 338 of Ross (1996)). These formulations can be useful for instance to obtain monotonicity results but are not of much help when solving the optimization problem.

Next, we move to solving the optimization problem under a special regime whereby we let the demand and supply take large values (asymptotic analysis). We induce from this solution an approximation of the delay that is coherent and consistent with the setting we are in. We then analyze this approximation, which will be instrumental for the multi-plan setting and our numerical analysis.

5.2. Asymptotically Optimal Solution

In this section we present an asymptotic analysis of the optimization problem (P) set in a regime where demand and capacity grow large. Specifically, we define a sequence of problems (P^n) parameterized by an integer $n \geq 1$. We let $\lambda^n(\cdot) = n\lambda(\cdot)$ and $\mu^n = n\mu$, while keeping $T^n = T$ and $N^n = N$. We denote by

$$\Pi^n(p,\kappa) = \lambda^n(p)\, p\, N - c\lambda^n(p)\, \varpi(\lambda^n(p),\kappa)$$

the profit rate achieved by problem (P^n) when price p per impression and the display frequency $1/\kappa$ are selected.

Furthermore, we define the function $\Psi(x) = \phi(x) - x\bar{\Phi}(x)$ on $\mathbb{R}$ where ϕ and Φ are the standard normal pdf and cdf. The function Ψ is increasing and convex where $\Psi'(x) = -\bar{\Phi}(x) \leq 0$ and $\Psi''(x) = \phi(x) \geq 0$. We use hereafter the notation $o(\cdot)$ for two real functions f and g where $g(x) = o(f(x))$ for all x in a neighborhood of x_0 if $g(x)/f(x) \to 0$ as $x \to x_0$. The next result gives an exact solution to problems (P^n) when n grows large. We denote by e^0 the elasticity coefficient of the demand function around λ^0, with $e^0 = \lambda^{0\prime} p^0/\lambda^0$, where $\lambda^{0\prime}$ is the derivative of λ with respect to price taken at $p^0 = \lambda^{-1}(\lambda^0)$. We recall that $\bar{\lambda}$ is the maximizer of the revenue function $r(\lambda; N) = \lambda p(\lambda; N)N$ and we define $\bar{\rho} = \frac{\bar{\lambda}N}{\mu s}$. We assume in the next result that $|e^0| > 1$, i.e., the fluid demand is elastic[2].

Proposition 3 *Suppose that the arrival stream of advertisers requesting a T-contract follows the demand process described in Section 3.1 with both demand and supply rates scaled as suggested above. Assume that $\bar{\rho} > 1$ and $\lambda^{0\prime}$ exists and is finite such that $|e_0| > 1$. Then the solution of the optimization problem (λ^n, κ^n) is such that*

i.) $\lambda^n = \lambda^0\, n - \lambda^0 \frac{\varpi^T - \eta}{T}\sqrt{n} + o(\sqrt{n})$

ii.) $\kappa^n = \kappa^0 n - \kappa^0 \frac{\varpi^T}{T}\sqrt{n} + o(\sqrt{n})$

[2] A similar assumption was also imposed by Maglaras and Zeevi (2003). We refer the reader to that paper for a brief discussion and illustrative examples of price-demand elasticity.

iii.) $\rho^n = 1 - \frac{\varpi^T - \eta}{T\sqrt{n}} + o(1/\sqrt{n})$

iv.) $\varpi^{T,n}(\lambda^n, \kappa^n) = \frac{\varpi^T}{\sqrt{n}} + o(1/\sqrt{n})$,

as $n \to \infty$, *where* $\varpi^T = \sigma\,\Psi(-\eta/\sigma)$ *and* $\sigma = \sqrt{s\,\kappa^0}/\lambda^0$. *Finally, denote by* $\Pi^{0,n} := \lambda^0 p^0 N n$ *the profit obtained in the fluid setting. The optimal profit under the stochastic setting is given by*

v.) $\frac{\Pi^n}{\Pi^{0,n}} = 1 - \xi^*/\sqrt{n} + o(1/\sqrt{n})$, *as* $n \to \infty$, *where* $\xi^* = \frac{c}{p^0}\frac{\varpi^T}{T} + (1 + 1/e^0)\frac{\varpi^T - \eta}{T}$ *and* η *is the unique solution to*

$$\bar{\Phi}(-\eta/\sigma) = \big(1 + \frac{c}{p_0(1 + 1/e^0)}\big)^{-1}.$$

This result shows that, in the single plan case, the fluid solution given in Proposition 1 is asymptotically optimal. It also suggests an improved solution by providing a correction term to the fluid limit. We obtain a similar result for the N-contract. The delay ϖ^N does not have a simple closed form (see the result stated in Appendix A). However, it is easy to show that $\varpi^T \leq \varpi^N$, which implies that campaigns running under N-contracts face more irregularity than those running under T-contracts. We analyze the asymptotic results numerically in Section 7.3 and confirm that the profit generated from a T-contract always upperbounds the profit of an N-contract.

When we un-scale the demand function and the capacity we see that the demand rate approaches λ^0 and the frequency parameter κ approaches κ^0. Hence, under the right regime applying the fluid solution guarantees a close-to-optimal behavior. This result is in line with the asymptotic optimality of a fixed price policy in the context of revenue management (see, e.g., Gallego and van Ryzin (1994)) and in the context of capacity sharing as in Maglaras and Zeevi (2003). The result *v.*) in Proposition 3 emphasizes further this conclusion, where we see that the fluid solution ensures a decreasing gap in profits (in the order of $1/\sqrt{n}$) with respect to the deterministic setting.

By scaling linearly the capacity, the requirement, N of one campaign becomes increasingly smaller than the total capacity $\mu^n T$. Hence, the publisher is driven to share this capacity among an increasing number of advertisers and thus increases the demand load and with it the number of active campaigns slots, $s\kappa^n$, at a comparable rate.

The asymptotic solution presented above is made of the fluid component and a correction term characterized by η, which is uniquely defined. A simple asymptotic analysis shows that the congestion factor is

$$\varrho^n = \rho^n(1 - \frac{\varpi^T}{T\sqrt{n}} + o(1/\sqrt{n}))^{-1} = 1 + \frac{\eta}{T\sqrt{n}} + o(1/\sqrt{n}),$$

as $n \to \infty$. In other words, as the system scales, it optimally moves towards a balanced load, and η/T is the rate at which the congestion factor reaches one. Similarly, the utilization gets close to one as well (such that $\sqrt{n}(1 - \rho^n)$ converges to a constant), while the delay approaches zero or

equivalently the campaigns delivery become more regular. This is a key result where the optimization tends to drive the system into high utilization while reducing the impact of uncertainty. This behavior has been highlighted previously in various queueing contexts starting with Whitt (1992) and made explicit through economic considerations in Maglaras and Zeevi (2003) and Maglaras and Zeevi (2005). The analysis in these papers is based on a multi-server queueing system in heavy traffic obtained through the Halfin-Whitt regime (see Halfin and Whitt (1981)), i.e., by holding constant the probability of *delay*. The embedded queueing behavior that is generated in our case is different from the above papers. First, ours is not a typical multi-server queue as campaigns do not have independent service times, instead their effective duration is governed by the viewers' arrival process. Our model is similar to that of Maglaras and Zeevi (2003) and Maglaras and Zeevi (2005) as capacity sharing is a fundamental feature. However, ours is obtained through the uniform delivery policy while theirs is obtained through processor sharing; more importantly, we consider a control on the price and the capacity allocation while keeping the total capacity fixed, and where they focus on pricing and capacity sizing. Moreover, Maglaras and Zeevi (2003) and Maglaras and Zeevi (2005) consider a queueing loss system while the delay in our case is the main representation of non-uniformity. The dynamics of the system in our case are primarily driven by the behavior of the congestion factor ϱ parameterized by η and not as much by the utilization factor ρ. Finally, another difference is that our results are proven to hold (see Section 6) in the multiple plan setting.

By following the proof of Proposition 3, one can observe that the results *i.*) - *iv.*) are valid for any value of η as long as $\sqrt{n}\,(1-\varrho_n) \to \eta/T$. Even the elasticity constraint is not required except for *v.*). Whether ϱ is larger or smaller than one depends on the sign of η. If the elasticity is high (in absolute value) then η depends on the ratio c/p_0, which is the units delay cost divided by the marginal revenue per impression. If this ratio is small (i.e., the penalty cost of delaying a campaign is not significant compared to the revenues) then η is large and positive and $\varrho > 1$. On the other hand, if the ratio is larger than 1 then η is negative and the larger the ratio is the more negative η is and $\varrho < 1$. Finally, if the elasticity $e_0 \approx -1$ then $\bar{\Phi}(-\eta/\sigma)$ is close to zero and η is again negative and $\varrho < 1$.

It is important to stress that the notion of slots sharing multiple ads introduced through the allocation mechanism suggested, allows the publisher to share the increasing capacity among the increasing number of advertisers while meeting the advertisers' requirements. The result of Proposition 3 proves the asymptotic optimality of the uniform allocation mechanism, which implies that the concept used in practice of sharing slots is quite effective in reducing the implied uncertainty and generating maximum revenues.

To summarize, for systems where the capacity μ is large relatively to N, say $\mu = n$, computing the fluid solution together with setting the optimal congestion factor ϱ through the computation of η allows one to solve for the optimal price and the display frequency and be able to measure the system's performance through the utilization ($\rho \approx 1 - \frac{\varpi^T - \eta}{T\sqrt{\mu}}$), the irregularity ($\varpi^T \approx \frac{\varpi}{\sqrt{\mu}}$), and the profit ($\Pi \approx \Pi^0(1 - \xi^*/\sqrt{\mu})$). It is important to highlight that this asymptotic analysis leads to approximations that numerically (see Section 7.3) are valid for reasonable values of n (≈ 5), making these results even more valuable in practice.

5.3. The Cost of Non-Uniformity and Its Approximation

The previous section presents an optimal solution to Problem (P) when the system is scaled. This solution also induces a natural approximation of the cost of non-uniformity. We denote the approximated delay by

$$\varpi_a^T(\lambda, \kappa) = \frac{\sqrt{s\kappa}}{\lambda}\, \Psi\Big(\frac{s\kappa - \lambda T}{\sqrt{s\kappa}}\Big).$$

(See Corollary 2 below). This formulation is interesting as it defines a simple relationship between the publisher's control parameters, price and display frequency, and the resulting level of irregularity in the campaigns delivery. As we have argued above, this formulation can be a valid representation of the cost of non-uniformity for other delivery mechanisms (beyond PUA). Recall that Proposition 2 and Corollary 1 give expressions of the delay. The first one was instrumental for the asymptotic analysis but both are not easy to manipulate beyond such a limiting regime. Moreover, they are hard to use numerically especially in the context of an optimization with an equality constraint. We devote this section to studying the suggested approximation. In Section 7.2 we compare numerically the performance of the approximation to the simulated values of the delay. We define Problem (P_a) similarly to Problem (P) where the cost of non-uniformity is being replaced by its approximation:

$$\begin{aligned} \max_{\lambda,\kappa} \Big\{ \lambda\, p(\lambda; N)\, N - \frac{c\, N}{T} \lambda\, \varpi_a^T(\lambda, \kappa) \Big\} \qquad (P_a) \\ s.t.\ \ \varpi_a(\lambda, \kappa) \big) = T - N\kappa/\mu \qquad \text{and} \qquad \rho \le 1. \end{aligned}$$

By following a similar proof to Proposition 3 one can show that the solutions to Problems (P_a) and (P) are asymptotically the same.

Corollary 2 *Under the same scaling as before, the asymptotic solution to the sequence of optimization problems (P_a^n) is asymptotically the same as that of (P^n) given in Proposition 3. In particular $\sqrt{n}\,|\lambda_a^n - \lambda^n| \to 0$ and $\sqrt{n}\,|\kappa_a^n - \kappa^n| \to 0$ and $\varpi^n/\varpi_a(\lambda^n, \kappa^n) \to 1$, as $n \to \infty$.*

For the rest of this section, we characterize the delay in a normal regime. We start by some monotonicity results, which show that the suggested approximation is not only relatively close to the actual delay but they both behave in a similar way with respect to λ and κ.

Proposition 4

i.) ϖ_a^T *and* ϖ^T *are both increasing in* λ

ii.) $\varpi^T(\lambda,0)=T$ *and* $\varpi_a^T(\lambda,\kappa)\to T$ *as* $\kappa\to 0$ *and* $\partial_\kappa\varpi_a^T(\lambda,\kappa)\to -s/\lambda$ *as* $\kappa\to 0$

iii.) ϖ_a^T *and* ϖ^T *are both decreasing in* κ *and go to zero as* $\kappa\to\infty$.

Parts *i.)* and *iii.)* in the previous result show in particular, how on average the delay behaves with respect to the two main control variables, price and display frequency, independently of any fulfillment constraint. The first result is intuitive, as the demand rate increases, the irregularity increases as well. The second behavior is less intuitive. If we disregard the fulfillment constraint and assume that the price is fixed, then by increasing κ two phenomena compete. On one hand, more campaigns can be active simultaneously, which tends to reduce the lag, but on the other hand the delivery pace decreases, which tends to keep a campaign active longer. Part *iii.)* shows that one lever is stronger, as κ increases the delay decreases. Equivalently, if the publisher runs simultaneously more campaigns (by reducing their display frequency, $1/\kappa$), an economies-of-scale effect is generated, which reduces the impact of uncertainty and makes the delivery on average more uniform. Next, we introduce the fulfillment constraint and discuss its impact on the delay.

Proposition 5

i.) For any fixed value of λ, *both equations* $\varpi^T(\lambda,\kappa)=T-N\kappa/\mu$, *and* $\varpi_a^T(\lambda,\kappa)=T-N\kappa/\mu$, *admit a non-zero solution. We denote by* $\kappa(\lambda)$ *(resp.* $\kappa_a(\lambda)$*) the largest one.*

ii.) Furthermore, $\kappa(\lambda)$ *(resp.* $\kappa_a(\lambda)$*) is decreasing in* λ *all else fixed with* $\kappa(\lambda)\le\kappa_0$ *(resp.* $\kappa_a(\lambda)\le\kappa_0$*). Finally,* $\varpi^T(\lambda,\kappa(\lambda))$ *(resp.* $\varpi_a^T(\lambda,\kappa(\lambda))$*) is increasing in* λ.

Proposition 5 *i.)* guarantees that the equality constraint has a non-empty solution set and reduces the optimization to a single variable. We note that for fixed λ, the largest solution to the constraint $\kappa(\lambda)$ (resp. $\kappa_a(\lambda)$) guarantees the smallest delay among the other solutions (if many exist[3]). In *ii.)*, the monotonicity of κ is intuitive. If the price per impression is lowered, the demand rate increases and with it the number of impressions to be met during T. Therefore, to meet the fulfillment constraint, one needs to direct more viewers to the ads and that is achieved by increasing the display frequency, i.e., by decreasing κ. As λ increases $\kappa(\lambda)$ decreases and the resulting delay also

[3] We show in the proof of Proposition 5 that κ_a, the non-zero solution to the equality constraint, is unique.

increases. Finally, recall that the fluid model solution (λ^0, κ^0) represents an upper bound for both values of λ and κ.

Proposition 6

i.) The approximated cost function, $c_a(\lambda, \kappa) := \frac{cN}{T}\lambda\, \varpi_a^T(\lambda, \kappa)$*, is convex in* λ *for fixed* κ *and is convex in* κ *for fixed* λ*, if and only if,* $\varrho := \lambda T/s\kappa < \varrho_0$ *for some* $\varrho_0 > 1$*. Finally, the cross derivative of the cost function is negative and its hessian is non-positive for all values of* λ *and* κ*.*

ii.) The function $c_a(\lambda, \kappa(\lambda))$ *is concave in* λ *for all* $\lambda \leq \lambda^0$*.*

Proposition 6 *i.)* shows that the objective function in Problem (P_a) does not behave "nicely" as the hessian is non-positive. However, from *ii.*) we can see that once the constraint is embedded into the objective (i.e. $\varpi(\lambda, \kappa(\lambda))$), the objective function becomes concave in λ (on the entire feasible set). In particular, this implies that the optimal solution (λ^*, κ^*) is unique, which is very helpful especially for the numerical analysis. Furthermore, one can easily prove that $\kappa(\cdot)$ is not only decreasing but also concave in λ (for the latter conclusion see the proof of Proposition 4 in Appendix A). The result in *i.*) shows that the cost (respectively, the profit) function is convex (concave) in each variable separately. The condition $\varrho < \varrho_0$ is not that constraining. First, because $\varrho_0 > 1$ does not impose any upper bound on the utilization (ρ can still take any value in $(0, 1)$, and thus no restriction on the price itself except on the gap between λT and κs. Second, if we recall the analysis for large scale systems, we know that the control variables will drive naturally the system towards high utilization and thus both load factors ρ and ϱ close to one, which would not be affected by the constraint on the congestion factor. The concavity results obtained in the previous proposition are quite helpful in solving other variants of our problem where the fulfillment constraint is relaxed (e.g., a system where only a minimum number of impressions is guaranteed.)

6. Multiple Advertising Plans Under PUA policies

6.1. Problem Formulation

In this section we revisit the general formulation (MP) where the publisher offers a number of plans $J > 1$. As we mentioned earlier PUA policies have the advantage of being tractable and simple to implement. By applying a PUA policy, where campaigns requiring the same plan j are grouped together and viewers are uniformly allocated to them, the non-uniformity cost and the fulfillment constraint become tractable. Both are reduced to a linear function of the delay. One property of the PUA policy is that among campaigns of the same plan j, the one that is booked first is always delivered first. This property allows one to formulate the non-uniformity cost and with it the optimization problem.

The PUA policies are similar in nature to partitioned nesting policies often adopted in yield management. The latter are known to perform well in practice yet are clearly suboptimal. Similarly, the PUA policies we suggest, have their inefficiencies, especially in the multi-plan case due to their *non-work-conserving* nature. For instance, if one group of campaigns, requiring plan j, is experiencing a slow booking rate while another plan i is being flooded with advertisers, it is possible that the slots $s\kappa_j$ allocated to plan j are under-utilized while those allocated to plan i are saturated with some campaigns experiencing a positive delay. However, setting optimally the price per impression for each plan, p_j, the display frequency through κ_j, as well as the proportion of viewers, f_j, directed to each plan, strives to eliminate these inefficiencies. Moreover, similarly to the single-plan setting, we show below that a well designed PUA is asymptotically optimal. Therefore, the solution obtained in this setting can well be used under a different mechanism (e.g., as an input to a delivery engine) and it would still behave almost optimally as long as regular delivery is important.

As defined before, we denote by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_J)$ the vector of demand intensities generated when the advertiser sets the vector of prices $\mathbf{p} = (p_1, p_2, ..., p_J)$ for the J plans available. The publisher solves the following problem:

$$\begin{aligned} &\max_{\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{f}} \Big\{ \sum_{j=1}^{J} \lambda_j \, p_j(\boldsymbol{\lambda}\,;\mathbf{N})\, N_j - \frac{c_j N_j}{T_j} \lambda_j \varpi_j^T(\lambda_j, \kappa_j) \Big\} \qquad \text{(MP)} \\ &s.t. \quad \varpi_j^T\big(\lambda_j, \kappa_j\big) = T_j - N_j \kappa_j / (\mu\, f_j), \qquad j = 1, 2, ..., J \\ &\qquad \rho_j := \lambda_j \, N_j / (s\, \mu\, f_j) \le 1, \qquad j = 1, 2, ..., J, \qquad \sum_{j=1}^{J} f_j \le 1. \end{aligned}$$

Two settings could be considered. The first one relates to non-substitutable plans where $\lambda_j(\boldsymbol{p}\,;\mathbf{N}) = \lambda_j(p_j\,;N_j)$. These plans divide the set of advertisers into disjoint classes. Every advertiser from a specific class is associated with one plan and decides whether to buy that plan or not. This setting makes sense when the plans are very different and segment the market naturally. In the case of substitutable plans, we consider one class of advertisers approaching the web publisher who could be interested in any of the J plans offered. The advertiser then selects the plan that fits his campaign best. Hence, the arrival rate for plan j is affected by all the J plans. We suggest in Section 7.1 two price-demand functions covering each case. For clarity purposes, we present the results here for non-substitutable plans. All the analysis extends to substitutable ones.

We end this section by suggesting another formulation of the optimization problem (MP). For that, we take advantage of the separability of the objective function, the necessarily binding constraint on the proportions and the redundant utilization constraint. This formulation requires a

version of the delay that is differentiable in κ. Therefore, we use the approximation ϖ_a instead of ϖ, defined in the previous section, which is asymptotically equal to ϖ, leads to the same fluid analysis, and is also differentiable. We will drop the subscript a.

$$\begin{aligned}
&\frac{\partial}{\partial f_j}\Big[\frac{c_j N_j}{T_j}\lambda_j \varpi_j(\lambda_j, \kappa_j(f_j;\lambda_j))\Big] = m \\
&\lambda_j = \arg\max_{\lambda_j}\Big\{\lambda_j p_j(\lambda_j) - \frac{c_j N_j}{T_j}\lambda_j \varpi_j(\lambda_j, \kappa_j(f_j;\lambda_j))\Big\} \qquad \text{(OC)} \\
&\sum_{j=1}^{J} f_j = 1,
\end{aligned}$$

where m (a Lagrange multiplier) is a constant independent of j.

6.2. Asymptotically Optimal Solution

In this section we present an asymptotic analysis of the multi-plan problem (MP) set in a regime where demand and capacity grow large. We follow the same approach as in Section 5.2 and define a sequence of problems (MP^n) parameterized by an integer n, which will increase towards infinity. We let $\lambda_j^n(\cdot) = n\lambda_j(\cdot)$ and $\mu_n = n\mu$ while keeping $T_j^n = T_j$ and $N_j^n = N_j$. The approach we use to solve the asymptotic regime relies on the general formulation given in the previous section and on the asymptotic solution of the single plan problem. First, based on the second equality condition of (OC), we adjust the single class asymptotic analysis by defining for all n a set of capacity portions f_n^j for each advertising plan j. Second, we use the first equation of (OC) to determine a characterization of f_j, which holds in particular for f_j^n. Finally, we solve for f_j^n recalling the last equation of (OC).

Proposition 7 *Suppose that the arrival stream of advertisers follows the demand process described in Section 3.1 and both demand and supply rates are scaled as suggested above. Assume that $\bar{\rho}_j > 1$ and that the first derivatives of $\lambda_j(\cdot)$ w.r.t. p at λ_j^0, $\lambda_j^{0\prime}$, exist and are finite for all $1 \le j \le J$. The solution of the optimization problem parameterized by n is such that*

i.) $\lambda_j^n = \lambda_j^0 n - \lambda_j^0(\frac{\varpi_j^T - \eta_j}{T} + \varsigma_j)\sqrt{n} + o(\sqrt{n})$

ii.) $\kappa_j^n = \kappa_j^0 n - \kappa_j^0(\varsigma_j + \frac{\varpi_j}{T})\sqrt{n} + o(\sqrt{n})$

iii.) $f_j^n = f_j^0 - f_j^0\varsigma_j/\sqrt{n} + o(1/\sqrt{n})$

iv.) $\rho_j^n = 1 - \frac{\varpi_j^T - \eta_j}{T\sqrt{n}} + o(1/\sqrt{n})$

v.) $\varpi_j^{T,n}(\lambda_j^n, \kappa_j^n) = \frac{\varpi_j^T}{\sqrt{n}} + o(1/\sqrt{n})$,

where, for all $1 \le j \le J$, we have $\varpi_j^T = \sigma_j^0\,\Psi(-\eta_j/\sigma_j^0)$, with $\sigma_j^0 = \sqrt{s\,\kappa_j^0}/\lambda_j^0$. The values of η_j's are given by $\Phi(-\eta_j/\sigma_j^0) = (1 + m/N_j)^{-1}$, for some positive m. Finally, we denote by $\Pi^{0,n} = \sum_j \lambda_j^0 p_j^0 N_j n$ the profit obtained in the fluid setting. The optimal profit in the stochastic setting is of the form

vi.) $\frac{\Pi^n}{\Pi^{0,n}} = 1 - \xi(m)/\sqrt{n} + o(1/\sqrt{n})$, *where* m *is selected to minimize* $\xi(m)$.

The second order terms of the capacity portions, ς_j's, and that of the profit, $\xi(m)$, are formulated in the proof of Proposition 7 in Appendix A. Furthermore, the existence of the η_j's is guaranteed as long as m is positive. This constraint translates to conditions on the demand function, similar to the one on the elasticity (i.e., $|e^0| > 1$) in the single plan case.

The result above shows that the fluid solution remains asymptotically optimal despite the complexity of the multi-plan setting. Recall that the value of Π^0 is independent of the capacity allocation mechanism used. Hence, the result shows that this optimal solution can be reached by implementing our suggested PUA policy, which is then asymptotically optimal. Hence, the ineffectiveness inherited by decoupling the advertising plans is diluted in large-scale systems and is not exaggerated as one could have expected.

The asymptotic solution reveals how the fluid solution gets corrected. Compared to the single plan, the policy components include a new factor, which is the capacity portion, f_j. This proportion has a complicated correction term (see Appendix A) as it is linking all the plans together. It is sensitive to the delay ϖ_j^T and the utilization through the term, $\varpi_j^T - \eta_j$, experienced by all the plans. The numerical analysis identifies some of these adjustments. Note that to our knowledge this kind of multi-product setting has not been analyzed before in the context of Halfin-Whitt regimes, obtained through economic considerations.

7. Numerical Analysis

Our numerical analysis is based on data from a large Scandinavian web publisher, Aller Internett, which runs several online magazines. It does not charge subscription fees rather revenues are generated by posting ads on the websites. There is a sales team that takes down orders for advertising campaigns where some negotiation can take place using a rate card price as a starting point. Even though the actual time the order was placed is not kept track of, the starting time of the campaigns are randomly spread, which reflects the randomness of the ordering time. Furthermore, the randomness on the traffic side is evident with large fluctuations throughout the day as well as across days for the same time of the day. Aller Internett does not offer its advertisers targeted impressions (e.g. young males interested in sport), however, the magazines have targeted audience (e.g., IT and women magazines).

The data is from a particular online magazine that Aller Internett runs with $600,000$ visitors on average per day (it does not track unique visitors). We consider around 250 orders made over a six month period after mid year 2009 to the beginning of 2010. On average there are about

1.3 campaigns[4] starting per day. The average duration of the campaigns (excluding the long-term contracts) is 40 days. The advertisers request different number of impressions to be delivered with the average being around 2 millions. The advertisers are not restricted to select from a menu of campaign lengths or number of impressions. However, we do see their choices clustered around 3 values of the campaign length (30 - 50, 70 and 90 days) and number of impressions (1, 2 and 3.5 million). Aller Internett uses Dart by DoubleClick to deliver the advertising campaigns, which requires multiple inputs such as how many ads can share a slot, for which our display frequency parameter κ could represent a good proxy.

7.1. Price-Demand Models

In order to perform the numerical analysis we explore in details the relationship between the price per impression, the number of impressions offered and the resulting demand rate. We suggest two models to depict this relationship; a utility-based demand function and a budget-based demand function. All our analysis holds for general demand functions. As argued in the literature (see, e.g., Gallego and van Ryzin (1994)), the demand rate $\lambda(p)$ can be considered to be a non-price-dependent demand rate Λ multiplied by the probability that the specific buyer has a reservation price larger than the listed price. The models we suggest generalize this idea to the setting of the online problem, where the probability of booking a campaign depends also on the number of impressions expected to be delivered. The two models are as follow:

Utility-based demand function. For this demand function we assume that advertisers interested in booking a campaign will only do so if their net utility is positive. In the single-plan setting the net utility is formulated as follows:

$$U(p;N) = \theta\, N^{\alpha} - pN$$

where θ is a measure of the sales impact generated by a campaign. (This model can easily be extended to the multi-plan case for both non-substitutable and substitutable plans.) The parameter θ is advertiser dependent and is taken to be uniformly distributed on $[0,\Theta]$. The resulting demand rate $\lambda(p;N)$ is given by

$$\lambda(p;N) = \Lambda\mathbb{P}(U(N)\geq 0) = \Lambda\Big(1-\Theta^{-1}\,p\,N^{1-\alpha}\Big) \qquad \text{or, equivalently,} \qquad p(\lambda;N) = \Theta\Big(1-\frac{\lambda}{\Lambda}\Big)N^{\alpha-1}.$$

Budget-based demand function. In this model, we assume that advertisers approach the website while having a budget constraint β (equivalent to a reservation price for the entire campaign)

[4] There are some long-term contracts but those are becoming less common and we exclude them.

on their spending and a minimum number of viewers ν to reach. For tractability, we assume ν to be a uniform random variable on $[0, M]$ and β a normally distributed random variable with mean $h(\nu) := d\,\nu + g$ and a standard deviation σ_b. We consider the setting where multiple plans, (N_j, p_j), are offered to a single class of advertisers (identified by d, g, M, σ_b). The advertisers book a campaign only if their budget constraint is satisfied and their reach target is fulfilled (i.e., $\nu < N_j$ and $\beta > p_j N_j$). We prove that for $1 \leq j \leq J$,

$$\lambda_j(p; N) = \frac{\Lambda}{dM}(G_{j+1}(N_j) - G_j(N_j)).$$

We denote by $G_j(N) = \mathbb{E}[\epsilon_j \wedge dN]^+$ with $G_{J+1}(N) = dN$, where ϵ_j is a normal random variable with mean $m_j := p_j N_j - g$ and standard deviation σ_b.

Based on the data available to us from Aller Internett, we try to generate reasonable estimates of the models' parameters to use for our numerical analysis.

7.2. Delay Approximation

To verify the quality of the approximation for the delay, which the non-uniformity cost depends on, we compare in Table 1 the simulated value of the delay to its approximation for the T contract through their absolute and relative difference (Δ^T and $\Delta^T(\%)$, resp.). For that, we explore different values of κ for the parameter values chosen: $\mu = 600,000$, $s = 5$, $T = 40$, and $N = 2,000,000$. We consider two values for the price per impression (corresponding to utilizations of 0.8 and 0.95). The

κ	ρ	ϖ^T	ϖ_a^T	Δ^T	$\Delta^T(\%)$	ρ	ϖ^T	ϖ_a^T	Δ^T	$\Delta^T(\%)$
1	0.80	35.83	35.83	0.00	0.00	0.95	36.49	36.49	0.00	0.00
5	0.80	19.16	19.17	0.00	-0.02	0.95	22.46	22.46	0.01	0.02
10	0.80	1.58	1.61	-0.03	-2.11	0.95	5.38	5.33	0.05	0.91
15	0.80	0.00	0.00	0.00	-660.94	0.95	0.02	0.04	-0.02	-93.35

Table 1 Approximations of the delay for the T-contract and the simulated values

approximation for ϖ^T performs in general very well. When κ increases the campaign delay goes to zero and thus the difference goes to zero as well, making the relative difference quite unstable. However, the relative difference is quite low for the values of κ ($\leq \kappa_0 = 12$) of interest to us.

7.3. Asymptotic Solution

We now analyze the asymptotic optimal solution and focus on the single plan case. A similar analysis can be done for the multi-plan case. We use a utility based price-demand function and set s and T as above with $\Lambda = 30$, $c = 0.022$, $\alpha = 0.9$, and $\Theta = 0.09$, in line with Aller Internett's data. We explore two sets of $(\mu; N)$: $(40,000; 400,000)$ and $(200,000; 2,000,000)$. The first case has the asymptotic parameter $\eta = 0.48$ and the second one has $\eta = -0.43$. Both cases have the same

elasticity, $e_0 = -59$. As illustrated in Section 5.2, we scale the capacity and the demand function linearly by introducing a parameter n that we increase from 1 to 50. The values of n below 25 correspond to reasonable settings in practice. The value $n = 50$ is an extreme case. In Table 2 we compare the values of the optimal solution obtained using the asymptotic approach (λ^n and κ^n) to the one obtained using the approximation of the non-uniformity cost (λ^a and κ^a). The differences ($\Delta^{\mathrm{a,n}}(\lambda)$ and $\Delta^{\mathrm{a,n}}(\kappa)$) are the relative differences measured in %. We also list the fluid values, λ^0 and κ^0. The last columns are dedicated to the two utility measures and the comparison of the delay for the T and N-contracts.

Table 2 confirms the theoretical result of Proposition 3 and that Problem (P_a) is asymptotically equivalent to Problem (P). Furthermore, it illustrates how the uncertainty can be absorbed by increasing κ and/or by decreasing λ away from their fluid values. We start with the case of $\mu = 40,000$ and $N = 400,000$ (the first four lines in the table). We observe that the demand rate remains very close to its fluid counterpart. The value of κ is more impacted by the presence of uncertainty; it is still asymptotically close to the fluid solution. In the lower part of the table the values of μ and N are $200,000$ and $2,000,000$. In this case η is negative and thus ϱ^n converges to one from below, which imposes a constraint on the utilization ($\rho \leq \varrho$). This forces the demand rate a bit further away from the fluid limit. Hence, we do see a slightly larger gap between λ^n and λ^0 and an opposite behavior for κ with respect to κ^0.

On the delay side, which captures the campaign irregularity and the non-uniformity cost, the values are reasonable with less than 1.70 day delay (out of a 40-days campaign) for $n \geq 5$. Again, the importance of this delay is that it measures the uncertainty in the system reflected by the non-uniformity cost. If in practice, the publisher cannot afford a delay beyond a certain time length, then an additional constraint can be imposed, which will dictate a higher price and probably a lower display frequency. Equivalently, this indicates that the publisher has underestimated the importance of the non-uniformity cost and should use a higher cost parameter c.

Table 3 considers the difference in profit measured in relative error (%). The difference in the optimal profit based on the approximation of the delay compared to the asymptotic optimal profit is less than 1.04% for all values of n. The second column shows that the asymptotic optimal profit converges to the fluid one. The third column makes the comparison to the profit values based on the simulated delay and the difference is negligible. The last column compares the two contracts through their profits. We observe that the T-contract's profit is always higher as shown in Section 5.2 but their relative difference seems to converge reasonably fast.

n	$\lambda^0 n$	λ^{a}	λ^{n}	$\Delta^{\mathrm{a,n}}(\lambda)$	$\kappa^0 n$	κ^{a}	κ^{n}	$\Delta^{\mathrm{a,n}}(\kappa)$	ρ^{n}	ϱ^n	$\varpi^{T,n}$
1	0.50	0.46	0.46	0.38	4	3.57	3.62	-1.34	0.92	1.01	3.81
5	2.50	2.41	2.41	0.06	20	19.10	19.15	-0.24	0.96	1.01	1.70
10	5.00	4.87	4.87	0.03	40	38.75	38.79	-0.11	0.97	1.00	1.21
25	12.50	12.29	12.29	0.01	100	98.05	98.09	-0.04	0.98	1.00	0.76
50	25.00	24.71	24.71	0.01	200	197.26	197.31	-0.02	0.99	1.00	0.54
1	0.50	0.45	0.45	0.32	4	3.62	3.66	-1.27	0.91	0.99	3.36
5	2.50	2.40	2.39	0.05	20	19.21	19.25	-0.22	0.96	1.00	1.50
10	5.00	4.85	4.85	0.02	40	38.90	38.94	-0.11	0.97	1.00	1.06
25	12.50	12.26	12.26	0.01	100	98.28	98.32	-0.04	0.98	1.00	0.67
50	25.00	24.67	24.67	0.00	200	197.58	197.62	-0.02	0.99	1.00	0.48

Table 2 Asymptotic optimal values for the T-contract with comparison to the approximate optimal values. The first set of lines have $\mu = 40,000$ and $N = 400,000$ and the next have $\mu = 200,000$ and $N = 2,000,000$. The differences are presented in %.

n	$\Delta^{\mathrm{a,n}}(\Pi)$	$\Delta^{0,\mathrm{n}}(\Pi)$	$\Delta^{\mathrm{sim,n}}(\Pi)$	$\Delta^{\mathrm{T,N}}(\Pi^{\mathrm{n}})$
1	-0.81	16.10	-0.39	13.68
5	-0.15	7.37	-0.07	5.95
10	-0.07	5.24	-0.04	4.18
25	-0.03	3.33	-0.01	2.63
50	-0.01	2.36	-0.01	1.85
1	-1.04	17.38	-0.47	14.89
5	-0.19	7.98	-0.10	6.46
10	-0.09	5.68	-0.04	4.53
25	-0.04	3.61	-0.02	2.85
50	-0.02	2.56	0.00	2.01

Table 3 Asymptotic optimal profit values for the T-contract compared to the approximate profit, the fluid profit, the simulated profit, and the asymptotic optimal profit for the N-contract. The first set of lines have $\mu = 40,000$ and $N = 400,000$ and the next have $\mu = 200,000$ and $N = 2,000,000$. The numbers are presented in %.

7.4. Numerical Analysis for The Single Plan

We analyze numerically the single-plan case and extract insights beyond the analytical results derived so far. We consider the budget price-demand function. The numerical results for the optimal solution are based on the approximation of the delay for the T-contract. We choose the parameter values extracted from the Scandinavian web publisher's data and set $\mu = 600,000$, $s = 5$, $T = 40$, $M = 3,000,000$, $\Lambda = 30$, $c = 0.022$, $\alpha = 0.98$, $\Theta = 0.09$, $g = 6000$, $d = 0.07$, and $\sigma_b = 15,000$.

We consider the effect of increasing the number of impressions on the optimal solution, see Table 4. In the fluid setting, as the number of impressions increases, the demand rate first increases before hitting the upper bound for which $\rho^0 = 1$ and then it decreases. The display frequency, $1/\kappa$, increases as the number of impressions increases. In the stochastic setting, both the arrival rate and the display frequency absorb the uncertainty by deviating away from their fluid values. However, the value of κ decreases with N more aggressively. This can be illustrated through the

N	$\lambda mbda^0$	λ^*	$\Delta(\lambda)$	κ^0	κ^*	$\Delta(\kappa)$	ρ^0	ρ^*	ϱ	Shortage
400,000	0.05	0.05	0.00	60.00	60.00	0.00	0.01	0.01	0.01	0.00
600,000	3.12	3.12	0.00	40.00	40.00	0.00	0.62	0.62	0.62	0.00
800,000	3.75	3.57	4.77	30.00	29.34	2.21	1.00	0.95	0.97	3.26
1,000,000	3.00	2.93	2.50	24.00	22.76	5.17	1.00	0.98	1.03	3.64
2,000,000	1.50	1.47	1.89	12.00	10.72	10.66	1.00	0.98	1.10	5.15
3,500,000	0.86	0.84	2.17	6.86	5.83	15.00	1.00	0.98	1.15	6.81

Table 4 Fluid and optimal values for different number of impressions based on the budget price-demand function. The differences are presented in %.

congestion factor $\varrho = \frac{\lambda T}{s\kappa}$. This ratio increases with N (see Table 4) implying a larger gap between the two decision variables, i.e., κ decreases faster than λ allowing the profit to remain close to the fluid benchmark. These observations confirm the following. First, from a pricing point of view, one should expect, in a loaded system, a larger price per impressions for larger contracts[5]. Second, the allocation mechanism relying on regular delivery, turns out to be critical in managing the uncertainty by creating effective economies-of-scale (recognized and discussed previously in the asymptotic analysis).

Note that $\Delta(\kappa) = 1 - \kappa^*/\kappa^0 = \varpi/T$. Hence, $\Delta(\kappa)$ measures the delay proportionally to the campaign length. For most contract sizes the delay should be acceptable (less than 5.17% of the duration). For large ones ($N \geq 2,000,000$), the uncertainty is amplified and the non-uniformity is harder to manage. If that is a serious issue for the web publisher, then the load ought to be reduced by, for instance, increasing the price per impression.

Finally in Table 4 we measure the shortage, which is the number of impressions under delivered by the fluid solution. A web publisher that disregards uncertainty will consistently miss on the expected number of impressions targeted. The number of impressions under delivered increases with N, starting at around 3% for low values of N and exceeding easily 6% for large ones.

A few other parameters can be explored such as the number of slots, the traffic, etc. We will conclude by exploring the impact of demand uncertainty by considering a variant of the Poisson process modeling the arrivals of campaigns. For that, we let as before $(v_i : i \geq 1)$ be an i.i.d. sequence of interarrival times (times separating two campaign requests), with mean $1/\lambda$ and standard deviation γ/λ, (all other moments are unchanged). In the Poisson case, v_1 is exponential and $\gamma = 1$. The number of campaigns booked in a period of time t has an average of λt and for large t, its standard deviation can be approximated by $\gamma\sqrt{\lambda t}$ (see, e.g., Ross (1996)). If we look at weekly demand,

[5] This is not surprising from an operations point of view as more impressions mean more workload. For example, Yahoo! recognizes the uncertainty caused by the supply scarcity and takes it into account when pricing by offering a higher price if the contract uses up a big portion of the impression inventory.

i.e., $t = 7$, for a demand rate of 1 advertiser/day, a value of $\gamma = 3$ leaves us with a coefficient of variation of 1.13, a quite reasonable value. We vary γ between 1 and 4. The behavior of the system with respect to the uncertainty parameter, γ, is of the same nature as the uncertainty impact we explored earlier through N, see Table 5. The demand rate is affected in a nonlinear way but the κ absorbs here more of the uncertainty and decreases quickly with γ. This behavior stresses the critical role of display frequency in handling uncertainty.

γ	λ^0	λ^*	$\Delta(\lambda)$	κ^0	κ^*	$\Delta(\kappa)$	Shortage
1	3.00	2.93	2.56	24.00	22.76	5.46	3.64
2	3.00	2.86	4.95	24.00	21.37	12.30	7.28
3	3.00	2.80	7.11	24.00	19.84	20.97	10.93
4	3.00	2.75	8.96	24.00	18.17	32.09	14.57

Table 5 Fluid and optimal values for different values of the uncertainty parameter γ based on the budget price-demand function with $N = 1,000,000$

7.5. Numerical Analysis for Multiple Plans

We now move to the multi-plan setting where we perform a numerical analysis for two plans with the same campaign length but different number of impressions. We determine the optimal price to charge, the optimal display frequency, and the optimal capacity proportion for Plan 1 with $N_1 = 1,000,000$ and for Plan 2 with $N_2 = z \cdot N_1$, $1 \leq z \leq 5$. We choose the parameters similarly as in the single plan case with $\mu = 600,000$, $s_i = 5$, $T_i = 40$, $\Lambda_i = 30/2$, $c_i = 0.022$, and $\gamma_i = 1$. Purposely, we picked the utility price-demand function that represents non-substitutable plans in order to keep the interaction of the two plans at an operational level. Each class of advertisers has a different utility function with $\alpha_1 = 0.99$, $\alpha_2 = 1.01$, and $\Theta_1 = 0.07$, $\Theta_2 = 0.05$. The magnitude of the values chosen were inspired by our data set.

Figure 1 shows the profit for the two plans as well as the total profit of the system. The behavior of the profit for Plan 2 depends highly on the parameters (α_2, Θ_2) relatively to (α_1, Θ_1). Clearly, the prices are set to take advantage of the more profitable plan. What is quite consistent among different parameter values is the decreasing nature of the profit of Plan 1. As the two plans become more distant, i.e., N_2/N_1 is large, more viewers are inevitably directed to Plan 2 (see Figure 5), which hurts Plan 1. The total profit is first increasing and will decrease eventually, here at the value of $N = 3.5$ million.

Next, we observe in Figure 2 the intuitive operational fact that as N_2 increases not only does the demand for Plan 2 decrease but also the demand for Plan 1. Many web publishers do not consider this direct impact of different plans on each other. In Figure 3 we see that the display frequency increases with N_2 as the same capacity is shared across larger workload (higher number

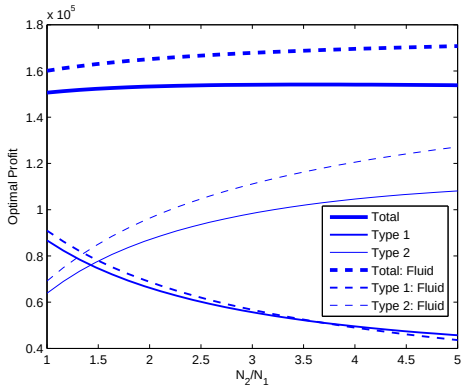

Figure 1 **Two plans: Optimal profit**

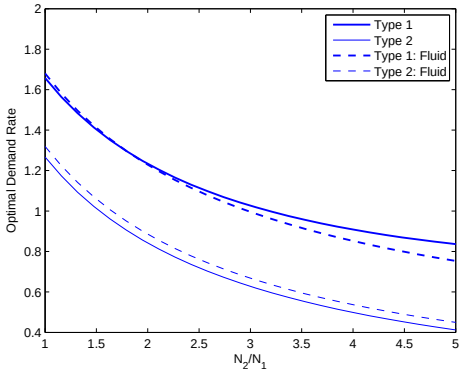

Figure 2 **Two plans: Optimal demand rate**

of impressions). The display frequency parameters help with absorbing the impact of uncertainty and stop the demand rates from dropping aggressively by allowing them to remain close to their fluid values. Note that the utilization for both plans is decreasing in N_2 (see Figure 4).

The impact of uncertainty in this two-plan model is more pronounced than in the single plan setting and the relative difference between the overall profit of the stochastic model and the profits under a deterministic setting is on average at around 8.2%. This reasonable performance of the stochastic model has been achieved through a complex interaction between the different pricing and operational variables.

We end this section by studying the impact of increasing the number of advertising plans offered. We increase the number of plans from 2 to 9 in the following way. We uniformly select the number of impressions for each plan between the smallest value of 1 million impressions and the maximum value of 10 millions with all of them having the same duration. Figure 6 shows that the PUA mechanism, despite the embedded ineffectiveness we previously discussed, does perform well compared to the fluid model. We note that the worst performance of the stochastic model under a uniform capacity allocation is when the number of plans is moderate (around 3 or 4 plans).

8. Conclusions

This paper develops a novel modeling framework for an operation facing uncertainties from both supply and demand while capturing specific delivery requirements. It is inspired by an online advertising problem whereby a web publisher needs to decide on the optimal price to charge per impression and the parameters governing the campaign delivery. The web publisher is constrained by the number of impressions and the campaign duration selected by the advertiser who expects a regular delivery throughout the campaign duration. The irregularity in the delivery is captured by a cost of non-uniformity. We suggest an allocation mechanism, the partitioned uniform allocation,

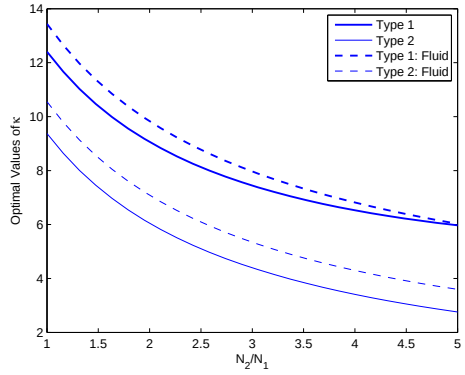

Figure 3 **Two plans: Optimal values of κ**

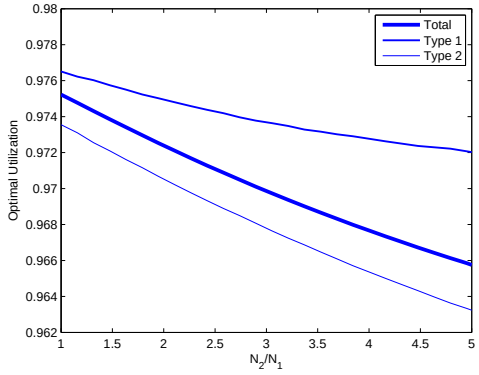

Figure 4 **Two plans: Optimal utilization**

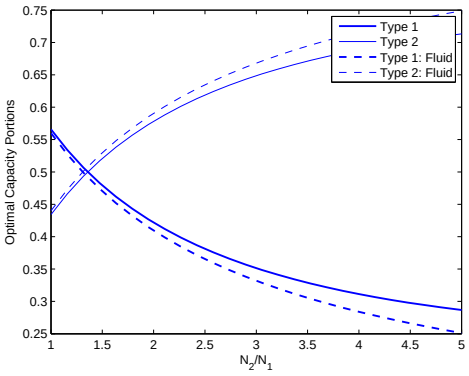

Figure 5 **Two plans: Optimal capacity portions**

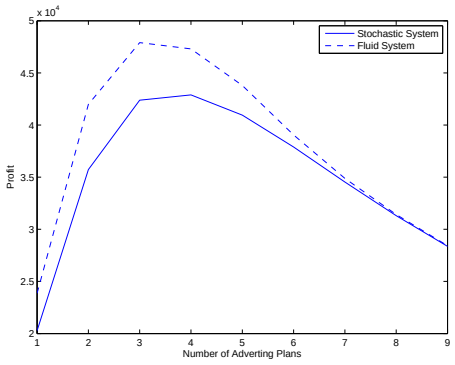

Figure 6 **Profit for multiple advertising plans**

that guarantees a uniform display of an ad through most of the duration of the campaign with the possibility of a delayed starting time. This mechanism relies on having the advertising slots sharing multiple ads in a rotating manner. Through a large-capacity system analysis, we obtain the optimal values for the price per impression and the campaign delivery control parameters. These values correspond to the solution of the fluid/deterministic problem corrected by square root terms, proving that the fluid values together with the allocation mechanism are asymptotically optimal.

The allocation we suggest has many advantages. Not only is it simple and asymptotically optimal, it is a means that allows one to link the control parameters to the system's uncertainty through the non-uniformity cost, making the optimization problem tractable. It is interesting to stress that the optimization problem under a large capacity drives the system into high utilization (under a Halfin-Whitt regime) while keeping the delivery irregularity small (even converging to zero). Undeniably, such behavior is the result of economies-of-scales induced by the allocation mechanism itself.

In practice, publishers use delivery engines relying on some optimization to allocate viewers

regularly to campaigns. They need to set a price per impression and divide their capacity among the different campaign types and fix the display frequency of an ad. Having in mind our asymptotic analysis, we believe that the simple formulation of the pricing, the capacity portion and the ad's display frequency, solution of our optimization problem, could represent reasonable inputs to use in practice.

The framework proposed in this paper considers several complexities of the online problem. However, some simplifying assumptions were made that could generate relevant extensions. Advertisers are in reality very keen on targeted advertising where they specify the attributes of the viewers that see their ads. We have assumed the viewers to be homogeneous as the Scandinavian publisher, (on which we base our numerical analysis) does for each of their online magazines. However, extending our results to non-homogeneous viewers is both relevant and interesting. We have assumed the traffic follows a Poisson process which is a valid assumption at an aggregated level. But, exploring time non-homogenous processes would certainly add value at an operational level. Some publishers take orders through an online system and can easily change their prices. Extending our analysis to dynamic pricing and dynamic display frequency are both challenging and relevant extensions.

Acknowledgments

The authors would like to thank René Caldentey and Gustavo Vulcano for insightful comments and suggestions. Several industry experts are gratefully acknowledged for insightful exchanges about practical aspects of online advertising and revenue management, including Jimmy Yang and Preston McAfee at Yahoo, Dimitri Metaxas at OMD Digital and Terje Johnansen at Aller Internett.

References

Araman, V. F., I. Popescu. 2010. Media revenue management with audience uncertainty: Balancing upfront and spot market sales. *Manufacturing Service Oper. Management* **12**(2) 190–212.

Asmussen, S. 2003. *Applied Probability and Queues*. Springer-Verlag (Second Edition), New York.

Bollapragada, S., H. Mallik. 2008. Managing on-air ad inventory in broadcast tv. *IIE Transactions* **40**(12) 1107–1123.

Chatterjee, P., D. Hoffman, T.P. Novak. 2003. Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Sci.* **22**(4) 520–541.

Evans, D. S. 2008. The economics of the online advertising industry. *Review of Network Economics* **7**(3) 359–391.

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

Gallego, G., G. van Ryzin. 1997. A multiproduct dynamic pricing problem and its applications to network yield management. *Oper. Res.* **45**(1) 24–41.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Gong, W., Y. Liu, V. Misra, D. Towsley. 2005. Self-similarity and long range dependence on the internet: a second look at the evidence, origins and implications. *Computer Networks* **48** 377–399.

Ha, L. 2008. Online advertising research in advertising journals: A review. *Journal of Current Issues and Research in Advertising* **30**(1) 31–48.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–588.

IAB. 2011. Internet advertising revenue report: 2010 full-year results www.iab.net.

Kingman, J.F.C. 1965. The heavy traffic approximation in the theory of queues. *Proc. Symp. on Congestion Theort* .

Kumar, S., V. S. Jacob, C. Sriskandarajah. 2006. Scheduling advertisements on a web page to maximize revenue. *Eur. J. Oper. Res.* **173**(3) 1067–1089.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49**(8) 1018–1038.

Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* **53** 242–262.

Radovanovic, A., A. Zeevi. 2009. Dynamic budget allocation mechanism for reservation based advertising. *The 15th INFORMS Applied Probability Society Conference* .

Roels, G., K. Fridgeirsdottir. 2009. Dynamic revenue management for online display advertising. *Journal of Revenue and Pricing Management* **8** 452–466.

Ross, S. 1996. *Stochastic Processes*. Wiley, New York, NY.

Savin, S. V., M. A. Cohen, N. Gans, Z. Katalan. 2005. Capacity management in rental businesses with two customer bases. *Oper. Res.* **53** 617–631.

Shaked, M., J.G. Shanthikumar. 1994. *Stochastic Orders and Their Applications, Academic Press*. Academic Press, St Louis.

Talluri, K., G. van Ryzin. 2004. *The theory and practice of revenue management*. Kluwer Academic Press.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Sci.* **38**(5) 708–723.

Zhao, H. 2000. Raising awareness and signaling quality t uninformed consumers: A price-advertising model. *Marketing Sci.* **19**(4) 390–396.

APPENDIX A: Main Proofs

A1. Proof of Proposition 1. i.) is straightforward. ii.) The constraint in MP0 is necessarily binding. This is a separable concave optimization problem. The formulation of the solution results directly from the KKT conditions. □

A2. Proof of Proposition 2. The key feature of this setting is the constant number of impressions, N, required by all advertisers. Despite the uncertainty in the arrival of viewers, such uncertainty does not alter the order of the advertisers leaving the system (after having their campaign fulfilled). This order is the same than the one they had when they initially approached the publisher. We rank the slots from 1 to $s\kappa$ and, when multiple slots are free, we assign advertisers to the lowest ranking slot. We can then tell, at arrival, on which slot (among the $s\kappa$ available) the ad will be displayed. Therefore, the slots dynamics can be decoupled each having its arrival process.

Let U_i be the time the i^{th} campaign takes to be completed once it had started to be displayed. The sequence $U = (U_i : i \geq 1)$ is stationary. Every κ viewers is directed to the same campaign, and every campaign needs N viewers. Thus, $U_1 \stackrel{D}{=} \sum_{j=1}^{N\kappa} u_j$ where the u_j's are the interarrival times between viewers. Similarly, let $V_{j+1} = \sum_{l=j+1}^{j+s\kappa} v_l \stackrel{D}{=} \sum_{l=1}^{s\kappa} v_l$, where, v_l's are the interarrival times between campaigns. Similarly to the dynamics of a single server queue, we can track the delay of each advertiser. Assume that the n^{th} was assigned a certain slot (among the $s\kappa$) then the next campaign that will be assigned the same slot is the $(n+s\kappa)^{th}$ campaign received. The arrival time between two consecutive campaigns sharing the same slot is $\sum_{l=n+1}^{n+s\kappa} v_l = V_{n+1}$. The formulation of the delay a campaign suffers follows a Lindley's type recursion $W_{n+s\kappa} = [W_n + U_n - V_{n+1}]^+$. Notice here that W_n is independent of U_n and V_{n+1}. Unfolding this recurrent equation leads to $W_n \stackrel{D}{=} \max_{0\leq m\leq n} S_m(\kappa)$ with $S_m(\kappa) = \sum_{j=1}^{m} X_j$ and $X_j = U_{j-s\kappa} - V_{j-s\kappa+1}$. Observe that X_1 is the difference between two gamma distributed random variable (and not the difference between two exponentially distributed r.v.). This Lindley relationship implies that the stationary distribution of the delay exists and is finite almost surely. Furthermore, it is equal in distribution to an infinite horizon maximum of a random walk $W_n \Rightarrow M(\kappa) = \max_{n\geq 0} S_n(\kappa)$, as $n \to \infty$. Of course, $W_{sn+1}, W_{sn+2}, ..W_{sn+s}$ are dependent random variables as their associated campaign is fulfilled with (at least partially) the same viewers. However, all these variables converge weakly to the same random variable M and hence, W_n as well. This single server queue type-relationship implies that when both u_i's and v_i's are exponentially distributed, then the delay function is equal in distribution to the waiting of a single server queue with interarrival times and service times distributed respectively as gamma random variables.

We move now to i). In the case of a T-contract, the n^{th} campaign is satisfied always before the $(n+1)^{st}$: ($A_n + T > A_k + T$ for all n and $k < n$ where A_j is the arrival time of the j^{th} campaign) and hence the order is preserved. Is it possible that the delay of a campaign reaches T and so leaves the system before being displayed at all? This is not possible. We prove that by contradiction. Assume the n^{th} campaign is the first one that the system has dropped and was not displayed. By definition of the n^{th} campaign, the previous campaign: $(n-1)^{st}$ was served and must have departed before the n^{th} was dropped without being displayed. This is not possible. Now, from the Lindley relationship, we have in the T-contract case (i.e. $T = W_n + V_n$), that $W_{n+s\kappa} = [T - V_{n+1}]^+$. By letting n go to infinity we obtain the result.

Finally, we prove iii.). In the T-contract case, $[T - \sum_{j=1}^{s\kappa} v_i]^+ \leq [T - \sum_{j=1}^{s} v_i]^+$ *a.s.*. Moreover, by the Strong Law of Large Numbers (SLLN), $\sum_{j=1}^{s\kappa} v_i \to +\infty$ *a.s.*, which implies that $W^T(\kappa) \to 0$ *a.s.* As for the N-contract, $W^N(\kappa) = \max_{m\geq 0} S_m(\kappa) \stackrel{d}{=} \max_{m\kappa \geq 0} S_m(1) \leq \max_{m\geq 0} S_m(1) \stackrel{d}{=} W^N(1)$. Again by the SLLN, $S_m(\kappa) \to -\infty$ *a.s.* and hence, $W^N(\kappa) \to 0$ *a.s.*. □

A3. Proof of Corollary 1. We let $N(T) = \max\{j : A_j \leq T\}$, where A_j is the time of the j^{th} arrival of an advertiser. Observe that $N(T)$ is a Poisson random variable with rate λT. Hence, for the values of $N(T)$ below $s\kappa$, $[T - A_{s\kappa}]^+ = 0$ and so,

$$\varpi^T(\kappa) = \sum_{j=0}^{\infty} \mathbb{E}\Big[[T - A_{s\kappa}]^+ | N(T) = j\Big] \mathbb{P}(N(T) = j) = \sum_{j=s\kappa}^{\infty} \mathbb{E}\big[(T - A_{s\kappa}) | N(T) = j\big] \mathbb{P}(N(T) = j).$$

Furthermore, we recall that conditioned on $N(T) = j$, the random variables, $\{A_1, A_2, ..., A_j\}$ are distributed as j i.i.d. uniformly distributed random variables on $(0, T)$ and so $A_{s\kappa}$ is the $s\kappa$ order statistics which is known to be beta distributed with parameters $(s\kappa, j + 1 - s\kappa)$. Hence, $\mathbb{E}A_{s\kappa} = T/(j+1) \cdot s\kappa$, which proves the result. For the N-contract, the result is based on Spitzer's formula see page 338 of Ross (1996). □

A4. Proof of Proposition 3.

LEMMA 1. *Let X be a normal random variable with mean η and standard deviation σ. The expected value of the truncated normal is given by $\mathbb{E}[X]^+ = \sigma\Psi(-\eta/\sigma)$, where the function Ψ is defined for all x, as $\Psi(x) = \phi(x) - x\bar{\Phi}(x)$ (see Section 5.2). Furthermore, Ψ is decreasing on $\mathbb{R}$ and for all $x \in \mathbb{R}$, $\Psi(-x) > x$ with $\Psi(-x)/x \to 1$, as $x \to +\infty$.*

Proof. By definition $\Psi(x) = \phi(x) - x\bar{\Phi}(x) \geq -x$. $\Psi(x)/x = \phi(x)/x - \bar{\Phi}(x) \to -1$ as $x \to -\infty$.

We move to the proof of Proposition 3. Let $\mu^n = n\mu$, while $T^n = T$, $N^n = N$ and $\lambda^n = n\,\lambda(\cdot)$. Notice that any solution to the delay constraint requires that $T - \frac{N\kappa^n}{n\mu} \geq 0$ and hence that $\frac{\kappa^n}{n} \leq \frac{\mu T}{N}$.

Furthermore, the utilization being bounded by one requires that $\frac{\lambda^n}{n} \leq \frac{s\mu}{N}$. We consider the log-moment generating function of the quantity $\sum_{i=1}^{s\kappa^n} v_i - \frac{s\kappa^n}{\lambda^n}$. As long as λ^n/n is bounded away from zero, we have that

$$\log \mathbb{E} \exp \theta(\sum_{i=1}^{s\kappa^n} v_i - \frac{s\kappa^n}{\lambda^n}) = \theta^2 \frac{s\kappa^n}{2{\lambda^n}^2} + O(n^{-2}). \tag{a1}$$

Consider any converging subsequence of κ^n and another converging subsequence of λ^n. Form the bounded sequence $n\kappa^n/{\lambda^n}^2$. Let m^n a common subsequence and denote by l the finite limit of $m_n {\kappa^{m_n}}/{\lambda^{m_n}}^2$. We recall here a result that will be used throughout the proof and that is, if any converging bounded subsequence converge to the same finite limit then the entire sequence converges to that same limit. For clarity of exposition we index the subsequence by m instead of m_n. We have that $\sqrt{m}\,(\sum_1^{s\kappa^m} v_i - \frac{s\kappa^m}{\lambda^m}) \Rightarrow Y \stackrel{d}{=} \sigma^0 Z$ as $m \to \infty$, where Z is a standard normal random variable and $\sigma^0 = \sqrt{s\,l}$. The fulfillment constraint can be written as follows $\sqrt{m}\,(T - \frac{N\kappa^m}{m\mu}) = \mathbb{E}\big[\sqrt{m}\,(T - \frac{s\kappa^m}{\lambda^m}) + Y + \varepsilon_m\big]^+$, where $\varepsilon_m \to 0$ as $m \to \infty$. Equivalently, we have, $0 = \mathbb{E} \max\big\{\sqrt{m}(\frac{N\kappa^m}{m\mu} - \frac{s\kappa^m}{\lambda^m}) + Y + \varepsilon_m, -\sqrt{m}(T - \frac{N\kappa^m}{m\mu})\big\}$. Consider a first regime made of subsequences of m (we use now the index j) for which $\sqrt{j}(T - \frac{N\kappa^j}{j\mu}) \to +\infty$. For such subsequences, the first term in the maximum ought to go to zero in expected value as j gets large and thus $\sqrt{j}(\frac{N\kappa^j}{j\mu} - \frac{s\kappa^j}{\lambda^j}) \to 0$ as $j \to \infty$. In particular, $\frac{N\kappa^j}{j\mu} - \frac{s\kappa^j}{\lambda^j} = \frac{s\kappa^j}{\lambda^j}(\rho^j - 1) \to 0$ as $j \to \infty$. This convergence implies that $\rho^j \to 1$ i.e. $\lambda^j/j \to \lambda^0 = \frac{s\mu}{N}$ and in turn $\kappa_j/j \to {\lambda^0}^2 l$. From the RHS of the fulfillment constraint we conclude that in such regime, the limiting system has a finite delay $\varpi^j \to T - \frac{N{\lambda^0}^2 l}{\mu} \geq 0$

The other possible regime is made of all subsequences for which $\sqrt{m}(T - \frac{N\kappa^m}{\mu^m})$ are bounded. Consider in such regime any converging subsequence, such that $\sqrt{j}(T - \frac{N\kappa^j}{j\mu}) \to \varpi$, for some non negative finite ϖ. For that to occur, we must have $\sqrt{j}\,(T - \frac{s\kappa^j}{\lambda^j}) \to \eta$ for some finite η. From these two limits, we conclude again that $\kappa_j/j \to \kappa^0 = \mu T/N$ and $\lambda^j/j \to s\kappa^0/T = \frac{s\mu}{N}$. From both possible regimes, we conclude that all converging subsequences of κ^n/n and λ^n/n converge respectively to κ^0 and λ^0 as $n \to \infty$. In this context, and based on the above Lemma 1 we have that $\sqrt{n}\,\varpi^n = \mathbb{E}\big[\sqrt{n}\,(T - \frac{s\kappa^n}{\lambda^n}) + Y + \varepsilon_n\big]^+ \to \sigma_0 \Psi(-\eta/\sigma_0)$ where ${\sigma^0}^2 = s\kappa^0/{\lambda^0}^2$. But again, in theory, η depends on the subsequence n_j. The equality constraint at the limit insures that $\sigma^0 \Psi(-\eta/\sigma^0) = \varpi$ and hence, $\sqrt{j}\,\varpi^j = \varpi + o(1)$, as $j \to \infty$.

In both regimes, the revenue side of the profit is maximized at the limit (with $\lambda^n \to \lambda^0$ or equivalently $\rho^n \to 1$ as $n \to \infty$). However, the second regime reaches at the limit a zero delay as opposed to a non-zero delay for the first regime. Thus the second regime always outperforms the first i.e. no matter the subsequence, as j gets larger the solution to the profit maximization ought to follow the second regime. We inject in the optimization constraint, the formulation of ϖ^j and solve for κ^j. We obtain that $\kappa^j = (T - \frac{\varpi^T}{\sqrt{j}} + o(1/\sqrt{j}))\frac{\lambda^j}{s\rho^j} = \kappa^0 j - \frac{\mu\,\varpi^T}{N}\sqrt{j} + o(\sqrt{j})$.

In turns, we inject the expression of κ^j in the term $(\sqrt{j}(T - s\kappa^j/\lambda^j)$ and get that

$$\sqrt{j}\,(T - s\kappa^j/\lambda^j) = \sqrt{j}\,T(1 - (\rho^j)^{-1}) + (\rho^j)^{-1}\,\varpi^T + o(1) \tag{a2}$$

which implies that $\sqrt{j}\,(1-\rho^j) \to d \geq 0$. By writing $\lambda^j = \lambda^0 j - l^j$, we first have that $(l^j\,N)/(\sqrt{j}s\mu) \to d$ and then replace it in Equation (a2), we obtain $\lambda^j = \lambda^0\,j - \frac{s\mu d}{N}\sqrt{j} + o(\sqrt{j})$. We let $j \to \infty$ in Equation (a2) and conclude that $-dT + \varpi^T = \eta$. Note that if $d = 0$, then $\eta = \varpi = \sigma\,\Psi(-\eta/\sigma)$ and that equation does not have any solution (Lemma 1). Hence, $d = (\varpi^T - \eta)/T > 0$, and, $\lambda^j = \lambda^0\,j - \lambda^0/T\,(\varpi^T - \eta)\,\sqrt{j} + o(\sqrt{j})$. The pricing policy that guarantees this arrival can be implied from a Taylor expansion of $\lambda^j(\cdot)$ in the neighborhood of $p^0 := \lambda^{-1}(\lambda^0)$. We write $\lambda^j(p^j) = \lambda^0 j + (p^j - p^0)\lambda'^0 j + o((p^j - p^0)j)$, where $\lambda^{0\prime} = \lambda'(p^0)$, the first derivative of λ at p^0. By comparing the two expressions of λ^j as j is large, we conclude that $p^j = p^0 - \frac{\lambda^0\,(\varpi^T - \eta)}{\lambda^{0\prime}\,T}\frac{1}{\sqrt{j}} + o(1/\sqrt{j})$.

The entire policy is constructed at this point. We still have a free parameter η to determine (which could eventually depend on the subsequence indexed by j). We recall that the profit obtained in the deterministic setting is $\Pi^{0,j} = \lambda^0\,p^0\,N\,j$ which is an upper bound of the the profit rate in the stochastic case. The parameter η will be selected in order to maximize that ratio for large j.

$$\begin{aligned}
\Pi^j(\lambda^j, \kappa^j) &= \lambda^j\,p^j\,N - cN/T\,\lambda^j \cdot \varpi^{T,j}(\lambda^j, \kappa^j) \\
&= \big(\lambda^0 j - \lambda^0/T\,(\varpi^T - \eta)\sqrt{j} + o(\sqrt{j})\big) \cdot \big(p^0 - \frac{\lambda^0\,(\varpi^T - \eta)}{\lambda^{0\prime}\,T}\frac{1}{\sqrt{j}} + o(1/\sqrt{j})\big)\,N \\
&\qquad - cN/T\big(\lambda^0 j - \lambda^0/T\,(\varpi^T - \eta)\sqrt{j} + o(\sqrt{n})\big)(\varpi^T/\sqrt{j} + o(1/\sqrt{j}) \\
&= \lambda^0 p^0 N\,j - p^0\lambda^0 N\,(\varpi^T - \eta)/T\,\sqrt{j} - \frac{\lambda^{0^2}\,(\varpi^T - \eta)N}{\lambda^{0\prime}\,T}\sqrt{j} - cN/T\lambda^0\varpi^T\sqrt{j} + O(1) \\
&= \lambda^0 p^0 N\,j - \lambda^0 p^0 N/T\big[\varpi^T\,(1 + \lambda^0/(\lambda^{0\prime}p^0) + (cN/T)T/(p^0 N)) - \eta(1 + \lambda^0/(\lambda^{0\prime}p^0))\big]\,\sqrt{j} + O(1) \\
&= \lambda^0 p^0 N\,j - \lambda^0 p^0 N/T\big[\varpi^T\,(1 + 1/e^0 + c/(p^0)) - \eta(1 + 1/e^0)\big]\,\sqrt{j} + O(1).
\end{aligned}$$

Hence, $\frac{\Pi^j}{\Pi^{0,j}} = 1 - \xi(\eta)/\sqrt{j} + o(1/\sqrt{j})$, where $\xi(\eta) = 1/T\big[\varpi^T\,(1 + 1/e^0 + c/p^0) - \eta(1 + 1/e^0)\big]$. We pick η, so as to minimize $\xi(\eta)$. We take the derivative of ξ with respect to η and recall that $\varpi'(\eta) = \bar{\Phi}(-\eta/\sigma)$ and so $\eta^* = -\sigma\bar{\Phi}^{-1}\Big(\big(1 + \frac{c}{p^0(1+1/e^0)}\big)^{-1}\Big)$ as long as $e^0 < -1$. This also proves that the constant η is unique independent of the subsequence, which also means that all the subsequences of λ^n and κ^n are asymptotically the same which proves the result.

Proposition 8 *Consider the case of N-contracts. Suppose that the input stream of advertisers follows a Poisson process and both demand and supply are scaled as suggested in Section 5.2. Assume that $\lambda^0 \leq \bar{\lambda}$ and $\lambda^{0\prime}$ exists and is finite such that $e^0 > 1$. Then, the solution of the optimization problem (λ^n, κ^n) is such that*

i.) $\lambda^n = \lambda^0\,n - \lambda^0\frac{\eta^N}{T}\sqrt{n} + o(\sqrt{n})$

ii.) $\kappa^n = \kappa^0\, n - \kappa^0 \frac{\varpi^N}{T}\sqrt{n} + o(\sqrt{n})$

iii.) $\rho^n = 1 - \eta^N/(T\sqrt{n}) + o(1/\sqrt{n})$

iv.) $\varpi^{N,n}(\lambda^n, \kappa^n) = \varpi^N/\sqrt{n} + o(1/\sqrt{n})$

v.) If the profit obtained in the deterministic setting is $\Pi^{0,n} = \lambda^0\, p^0\, N\, n$ *then, the ratio* $\Pi^n/\Pi^{0,n}$ *is of the form* $\frac{\Pi^n}{\Pi^{0,n}} = 1 - \beta(\eta)/\sqrt{n} + o(1/\sqrt{n})$,

where, $\varpi^N = \mathbb{E}\max_{r\geq 0} S_r$, *and* $(S_r : r \geq 0)$ *is a random walk with normally distributed increments with mean* η^N *and standard deviation* $\sigma = (\frac{\kappa^0 N}{\mu^{0^2}} + \frac{\kappa^0 s}{\lambda^{0^2}})^{1/2}$; η^N *is selected so that* $\beta(\eta)$ *is minimized.*

We do have approximations of ϖ^N. One of them, $\varpi^N \approx \frac{\sigma^2}{2\eta^N}$ is given by Kingman (1965). If we replace ϖ^N by this approximation, the optimal value of β is given by $\beta^* = \eta^N/T(1 + 1/e^0) + cN/T/(p^0 N)\sigma^2/(2\eta^N)$, and $\eta^N = \sigma\sqrt{\frac{c}{2p^0(1+1/e^0)}}$, when again $e^0 < -1$. The proof will be skipped. It follows the same approach as for the T-contract.

A5. Proof of Proposition 4. We start by showing monotonicity of ϖ^T in κ and λ. We recall (see Shaked and Shanthikumar (1994)) that for any renewal process $(S^n : n \geq 0)$ $S^n \leq_{st} S_{n+1}$ where $\leq_{st}$ denotes a stochastic ordering (i.e. for any increasing function ϕ, $\mathbb{E}\phi(S^n) \leq \mathbb{E}\phi(S_{n+1})$. In particular, we apply this to the decreasing function $\phi(x) = [T - x]^+$ and conclude that for $\kappa_1 < \kappa_2$,

$$\varpi(\kappa_1) = \mathbb{E}\big[T - \sum_{k=1}^{N\kappa_1} v_k\big]^+ \geq \mathbb{E}\big[T - \sum_{k=1}^{N\kappa_2} v_k\big]^+ = \varpi(\kappa_2). \tag{a3}$$

The same proof applies to show monotonicity in λ as long as v is stochastically decreasing in λ.

We move to ii.). By definition, $\varpi(\lambda, 0) = T$. We denote by $\alpha_\kappa = \frac{s\kappa - \lambda T}{\sqrt{s\kappa}}$ and $\sigma_1 = \sqrt{s}/\lambda$. As $\kappa \to 0$, $\alpha_k \to -\infty$. Based on Lemma 1 (stated above), $\Psi(x)/x \to -1$ as $x \to -\infty$. Hence, $\sigma_1\sqrt{\kappa}\,\Psi(\alpha_\kappa) \sim -\sigma_1\sqrt{\kappa}\,(s\kappa/\lambda - T)/(\sigma_1\sqrt{\kappa}) \to T$ as $\kappa \to 0$. We consider the derivative with respect to κ. We denote by α'_κ the derivative of α w.r.t. κ. Similar calculations show that as $\kappa \to 0$,

$$\begin{aligned}\partial_\kappa \varpi_a(\kappa, \lambda) &= \sigma_1/(2\sqrt{\kappa})\,\Psi(\alpha_\kappa) - \sigma_1\sqrt{\kappa}\,\bar{\Phi}(\alpha_\kappa)\alpha'_\kappa \\ &\sim \sigma_1/(2\sqrt{\kappa})(-(s/\lambda)/\sigma_1\sqrt{\kappa} + T/(\sigma_1\sqrt{\kappa}) - \sigma_1\sqrt{\kappa}((s/\lambda)/(2\sigma_1\sqrt{\kappa}) + T/(2\sigma_1\kappa^{3/2})) \\ &\sim -(s/\lambda)/2 + T/(2\kappa) - ((s/\lambda)/2 + T/(2\kappa)) = -s/\lambda.\end{aligned} \tag{a4}$$

We leave the proof of the monotonicity of ϖ_a with respect to κ and λ to Proposition 6. □

A6. Proof of Proposition 5.

i.) For clarity purposes, we drop the T and the λ in $\varpi^T(\lambda, \kappa)$ and write it as $\varpi(\kappa)$, a function of κ. We do the same with ϖ_a^T. We fix $\lambda \leq \lambda^0$. We start with Problem P_a. We denote by $l(\kappa) = T - N\kappa/\mu$. Note that $l(0) = \varpi(0) = \varpi_a(0)$. From a previous proposition, the derivative $\varpi'_a(0) = -s/\lambda$ while $l'(0) = -N/\mu$. The upper bound on the utilization, $\rho \leq 1$, translates in $|l'(0)| \leq |\varpi'_a(0)|$. The two functions l and ϖ_a have the same positive starting point, and the latter is steeper at zero and

decreases to zero ($\varpi(\kappa) \to 0$ as $\kappa \to \infty$); while the former goes to $-\infty$ ($l(\kappa) \to -\infty$ as $\kappa \to \infty$). Therefore, the two functions must intersect. We denote by $\kappa_a(\lambda)$ the largest value (in case many exist) at which the two functions intersect. Note that the revenue function is independent of κ while the delay function is decreasing in κ and hence, as long as the constraint is satisfied a largest possible κ is optimal. We will see below that ϖ is either concave or convex first and then concave and it is not hard to see that the intersection with $T - N\kappa/\mu$ cannot occur except in the concave region which make this intersection unique.

As for Problem P, we first recall that the constraint needs to be adjusted for the fact that κ is defined on $\mathcal{K} = \{\kappa : \kappa\, s \in \mathbb{N}\}$). For that we analyze P_l where ϖ is replaced by ϖ_l. $\varpi_l(\kappa)$ is equal to $\varpi(\kappa)$ on $\mathcal{K}$ and in between, it is defined through linear interpolation. We look at

$$\begin{aligned} \varpi(1) - l(1) &= \mathbb{E}[T - \sum_{i=1}^{s} v_i]^+ - (T - N/\mu) = \mathbb{E}\max\{N/\mu - \sum_{i=1}^{s} v_i, -(T - N/\mu)\} \\ &\le \max\{N/\mu - s/\lambda, -(T - N\mu)\} \le 0, \end{aligned} \tag{a5}$$

where the first inequality is obtained by Jensen's inequality and the convexity of the max function and the second inequality results from the utilization $\rho \le 1$. Hence, ϖ_l intersects with l at $\kappa \ge 1$ and there exists, as we discussed above, a unique κ that guarantees the smallest delay cost.

ii.) Finally, the monotonicity of $\kappa(\lambda)$ and $\kappa_a(\lambda)$ in λ is the result of ϖ and ϖ_a being both increasing in λ. The delays are always non-negative which implies that both κ_a and κ are upper bounded by κ^0. □

A7. Proof of Proposition 6. We skip the proof of this proposition which is based on straightforward yet tedious derivative calculations.

A8. Proof of Proposition 7. We disregard the index identifying each plan when there is no confusion from doing so. We also disregard the upper index T. We denote by $\lambda^{0,j}, \kappa^{0,j}, f^{0,j}$ the solution of the fluid model. For clarity of exposition and without loss of generality we assume that these sequences converge for each plan j to some finite limits. In principle, we should work throughout the proof with subsequences of κ^n/n and λ^n/n and prove that these subsequences converge to the same limit as we did in the single plan proof. Following the same steps as in the proof of the single plan case, we have that κ^n resp. λ^n is given by $\kappa^n = (T - \varpi^j)\mu n f^n/N = (T\mu/N) f^n\, n - (\mu\varpi/N) f^n \sqrt{n} + o(\sqrt{n})$ resp., $\lambda^n = \lambda^0 f^n\, n - \frac{\lambda^0 f^n}{T}(\varpi - \eta)\sqrt{n} + o(\sqrt{n})$. Similarly to the single plan case, a free parameter η is introduced (for each plan), such that $\sqrt{n}\,(T - s\kappa^n/\lambda^n) \to \eta$, and $\sqrt{n}\,(T - \sum_{i=1}^{s\kappa^n} v_i) \Rightarrow \mathcal{N}(\eta, \sigma)$, with $\sigma^{0^2} = s\kappa^0/\lambda^{0^2}$. Again, $\sqrt{n}\varpi^n \to \varpi := \sigma\Psi(-\eta/\sigma)$. From the fluid solution of the multi-plan problem and the fact that f^n is bounded, we conclude that any subsequence of f^n must converge to f^0. We then write $f^n = f^0(1 - \varsigma^n)$.

$$\begin{aligned} \sqrt{n}\,(1 - \rho^n) &= \sqrt{n}\,(s\mu f^n - \lambda^0 N + l^n N/n)/(s\mu f^n) \\ &= \sqrt{n}\,(-s\mu f^0 \varsigma^n + l^n N/n)/(s\mu f^n). \end{aligned} \tag{a6}$$

We denote by ς the limit of $\sqrt{n}\varsigma^n$ as $n \to \infty$. Hence, $l^n = ((d+\varsigma)s\mu f^0/N)\sqrt{n} + o(\sqrt{n})$, as $n \to \infty$.

Without loss of generality, we assume a uniform $c := 1$. Given the solution $\boldsymbol{\lambda}^n$ of MP^n, the cost optimization problem defined by $G(\boldsymbol{\lambda})$ allows to obtain the corresponding $\boldsymbol{\kappa}^n$ and $\boldsymbol{f}^n$. The latter are parametrized by $\boldsymbol{\eta}^n$, hence the solution to the minimization problem $G(\boldsymbol{\lambda})$ has the η's as its solution. The limiting cost in (c7) is $\sum_j N_j/T_j\lambda_j^0\varpi_j$ and it is a function of the free variables η_j's defined above. For every plan j the fulfillment constraint at the limit is given by $-d_j T + \varpi_j = \eta_j$. Finally, the proportion constraint can be reduced to $\sum_j \varsigma_j f_j^0 = 0$. Given $\boldsymbol{\lambda}^n$, the quantity l_j^n is also given for all n, hence the sum $\sum_j l_j^n N_j/(s\mu\sqrt{n})$ is a constant. Furthermore, $\sum_j l_j^n N_j/(s\mu\sqrt{n}) = \sum_j d_j f_j^0 = \sum_j (\varpi_j - \eta_j) f_j^0 = constant$. We can characterize the η's by minimizing the limiting cost.

$$\begin{aligned} &\min_{\eta_j} \sum_j N_j/T_j\lambda_j^0\varpi_j \\ &\text{s.t.} \sum_j (\varpi_j - \eta_j) f_j^0 = constant. \end{aligned} \tag{a7}$$

Recall that $\lambda_j^0 = s\mu/N_j f_j^0$. The optimality conditions are given by the following set of equations parametrized by a constant m. For all $1 \le j \le J$, $\frac{f_j^0}{T_j}\partial_{\eta_j}\varpi_j - m f_j^0 \partial_{\eta_j}\varpi_j + m f_j^0 = 0$. Recalling again that $\Psi'(x) = -\bar{\Phi}(x)$, we obtain that for all j, $\frac{\bar{\Phi}(-\eta_j/\sigma_j)}{\Phi(-\eta_j/\sigma_j)} = mT_j$. This uniquely defines all the η_j's as a function of m. The value of m will be characterized later by the profit maximization.

What we still need to do is to obtain an asymptotic approximation of f^n. In other words get a second order approximation of ς^n. In order to obtain the second order approximation for the capacity portion f_j, it will be necessary to obtain a more accurate approximation of ϖ^n involving ς^n. To do that, we use the approximation ϖ_a developed for the single plan. Recall that for a fixed set of proportions f_j's, $\sqrt{n}\,|\varpi^n - \varpi_a^n| \to 0$ as $n \to \infty$ and for all plans j. This convergence is uniform in f_j. We use the condition given by Equation (c8) which allows a characterization of f through the derivative of ϖ_a with respect to f. Note that $\partial_f \lambda\varpi_a(\kappa(f);\lambda) = \lambda\kappa'(f)\partial_\kappa\varpi_a(\kappa(f)) = m_A$. We take the derivative of the fulfillment constraint equation and we obtain that $\partial_f\varpi_a(\kappa(f);\lambda) = -N\kappa'(f)/(\mu f) + N\kappa(f)/(\mu f^2)$. Putting together the previous equations, we obtain a formulation of $\kappa'(f)$, when re-injected in the first equation gives

$$\frac{\lambda N\kappa(f)}{\mu f^2} \cdot \frac{\partial_\kappa\varpi_a}{\partial_\kappa\varpi_a + N/(\mu f)} = m_A T/N. \tag{a8}$$

We recall that the derivative of ϖ_a with respect to κ, is $\partial_\kappa\varpi_a = \sqrt{s/\kappa}/(2\lambda)\,\phi(\alpha_\kappa) - s/\lambda\bar{\Phi}(\alpha_\kappa)$, where $\alpha_\kappa = \sqrt{s\kappa} - \lambda T/\sqrt{s\kappa}$. If we introduce the scaling by n, it is easy to see that the first term in this derivative is of an order smaller than the second one. Hence, injecting the derivative in Equation (a8), we get that

$$m_A^n T/N = \frac{N\kappa^n}{\mu n {f^n}^2} \cdot \frac{\lambda^n \partial_\kappa\varpi_a(\kappa^n,\lambda^n)}{\partial_\kappa\varpi_a(\kappa^n,\lambda^n) + N/(\mu n f^n)}$$

$$= -\frac{N\kappa^n/f^n}{f^n} \cdot \frac{\bar{\Phi}^n\lambda^n}{-\bar{\Phi}^n+\rho^n} = s(T-\varpi^n)/N\,\mu^2n^2/(\mu n)\frac{\rho^n\bar{\Phi}^n}{\rho^n-\bar{\Phi}^n},$$

where $\Phi^n = \Phi(\alpha_{\kappa^n})$. We introduce at this point another notation, $\Phi_0 = \Phi(-\eta/\sigma)$. A similar notation will also be used for $\bar{\Phi}$ and ϕ. We turn now to study the ratio $\frac{\rho^n\bar{\Phi}^n}{\rho^n-\bar{\Phi}^n}$. For that we start by getting an approximation of α_{κ^n} when n is large.

$$\begin{aligned}
\alpha_{\kappa^n} &= -\frac{\lambda^n}{\sqrt{s\kappa^n}}(T-\frac{s\kappa^n}{\lambda^n}) = \sqrt{s\kappa^n} - \frac{\lambda^n T}{\sqrt{s\kappa^n}} \\
&= \sqrt{s\mu n f^n/N(T-\varpi^n)} - \frac{s\mu n f^n \rho^n T}{N\sqrt{s\mu n f^n/N(T-\varpi^n)}} \\
&= \sqrt{\frac{s\mu T}{N}}(nf^n)^{1/2}(1-\varpi^n/T)^{1/2} - \sqrt{\frac{s\mu T}{N}}(nf^n)^{1/2}\frac{\rho^n}{(1-\varpi^n/T)^{1/2}} \\
&= \sqrt{\frac{s\mu T}{N}}(nf^n)^{1/2}\big(1-\frac{\varpi}{2T\sqrt{n}} - (1-\frac{d}{\sqrt{n}})(1+\frac{\varpi}{2T\sqrt{n}})\big) \\
&= \sqrt{\frac{s\mu T}{N}}(nf^n)^{1/2}\big(-\frac{\varpi}{T\sqrt{n}} + \frac{d}{\sqrt{n}}\big) = -(1-\frac{\varsigma}{2\sqrt{n}})\frac{\eta}{\sigma} + o(1/\sqrt{n}).
\end{aligned} \tag{a9}$$

A Taylor expansion around $-\eta/\sigma$ gives the following $\bar{\Phi}^n - \bar{\Phi}^0 = \frac{\eta}{\sigma}\frac{\varsigma^n}{2}\,\phi^0 + o(1/\sqrt{n}) = -\frac{\eta}{\sigma}\frac{\varsigma^n}{2}\,\phi^0 + o(1/\sqrt{n})$. So, $\frac{\bar{\Phi}^n}{\Phi^n} = \frac{\bar{\Phi}^0}{\Phi^0}(1-\frac{\eta}{\sigma}\frac{\phi^0}{2\Phi^0\bar{\Phi}^0}\varsigma^n) + o(1/\sqrt{n})$.

Hence, $m_A^nT = (1-\frac{\varpi^T}{T}\frac{1}{\sqrt{n}})\frac{\bar{\Phi}^0}{\Phi^0}(1-\frac{\eta}{\sigma}\frac{\phi^0}{2\Phi^0\bar{\Phi}^0}\varsigma^n)(1+d\frac{\bar{\Phi}^0}{\Phi^0\sqrt{n}}) = \frac{\bar{\Phi}^0}{\Phi^0}(1-(\frac{\varpi^T}{T}-d\frac{\bar{\Phi}^0}{\Phi^0})\frac{1}{\sqrt{n}} - \zeta\eta\varsigma_j^n)$, where $\zeta = \frac{1}{\sigma}\frac{\phi^0}{2\bar{\Phi}^0\Phi^0}$. We also have that $\frac{m_A^n}{m} = 1-(\frac{\varpi_j}{T_j} - \frac{\varpi_j-\eta_j}{T_j}mT_j)/\sqrt{n} - \zeta_j\eta_j\varsigma_j/\sqrt{n}$. We multiply by $f_j^0/(\zeta_j\eta_j)$ and sum on all the classes while recalling that $\sum_j f_j^0\varsigma_j^n = 0$. We get $\frac{m_A^n}{m} = 1 - \frac{\sum_j \frac{f_j^0}{\eta_j\zeta_jT_j}(\varpi_j - mT_j(\varpi_j-\eta_j))}{\sum_j \frac{f_j^0}{\eta_j\zeta_j}\sqrt{n}}$. By replacing the formulation of m_A^n/m in the expressions of ς_j, we get the expression of ς_j as a function of m. $\zeta_j\eta_j\varsigma_j = \frac{\sum_j \frac{f_j^0}{\eta_j\zeta_jT_j}(\varpi_j-(\varpi_j-\eta_j)mT_j)}{\sum_j \frac{f_j^0}{\eta_j\zeta_j}} - (\frac{\varpi_j}{T_j} - \frac{\varpi_j-\eta_j}{T_j}mT_j)$. Similarly to the single plan case, the value of m is selected to maximize the profit ratio of the stochastic model with the fluid one. By taking advantage of the calculations in the single plan case, we have that the profit for each plan j is: $\Pi_j^n(\lambda_j^n,\kappa_j^n,f_j^n) = \lambda_j^0p_j^0N_j\,n - \lambda_j^0p_j^0N_j/T\big[\varpi_j^T\,(1+1/e_j^0+c/p_j^0) - (\eta_j - \varsigma_j)(1+1/e_j^0)\big]\,\sqrt{n} + O(1)$. We sum over the profits for all the plans and divide by the total fluid profit, $\Pi^{0,n} = \sum_j \lambda_j^0\,p_j^0\,N_j\,n$, we conclude that $\frac{\Pi^n}{\Pi^{0,n}} = 1-\xi(m)/\sqrt{n} + o(1/\sqrt{n})$, where m is selected to minimize $\xi(m)$. Note that η_i and ς_i are both functions of m. $\square$

SUPPLEMENT MATERIAL

In the following, we included material that we are aware will not be published nor reviewed by the referees. We believe this material could be helpful for the refereeing process. In the below, we included material related to B1. Aggregation procedure for the supply, B2. Price demand functions, B3. Data estimation. We also included material related to proofs that were skipped from the main document. These proofs do not bring any additional contribution. They are either tedious calculations or similar to other proofs already detailed in Appendix A. These proofs relate to C1. Proof of Proposition 1, C2. Proof of Proposition 6, C3. Proof of Proposition 8 and C4. General solution of the multi-plan optimization problem.

B1. The Aggregation Procedure for the Supply

We describe how the traffic coming to all webpages within the website can be aggregated and modeled as one source. We denote by $(u_i : i \geq 1)$ the aggregated sequence of interarrival times of all viewers to all the webpages belonging to the website, which is the time between two uploads of any webpages on the website. We assume that any page requested and uploaded by a viewer contains s identical slots and that the viewer sees all the s ads displayed at that time.

Let us now illustrate how the traffic of viewers to the different webpages belonging to the same website can be aggregated. First, we consider a website made of a single page. Then the u_i's are the times between two viewers visiting the webpage. Suppose now that the website is made of a homepage and two other pages where viewers arrive (according to a Poisson process) first at the homepage at a rate μ_0, spend an exponential time and then leave the homepage at the same rate. They might then access each page $i = 1, 2$ with probability q_i or leave the website with probability $1 - (q_1 + q_2)$. The aggregated process of viewers accessing any of the three pages on this website is Poisson with rate $\mu = (1 + q_1 + q_2)\,\mu_0$. If viewers can go directly to pages 1 and 2 with rate μ_1 and μ_2 then the aggregated rate is $\mu = (1 + q_1 + q_2)\,\mu_0 + \mu_1 + \mu_2$. The u_i's are then the interarrival times of this aggregated process. Hence, we are counting viewers in terms of number of pages uploaded independently of the actual page (as long as it has advertising slots).

B2. Price-Demand Functions

We derived two price-demand functions suitable in the online advertising setting. Note that even though the price and the demand rate are the key variables, the number of impressions, N, also plays an important role.

Utility-based demand function For this demand function we assume that advertisers interested

in booking a campaign will only do so if their net utility is positive. In the single-plan setting the net utility is formulated as follows:

$$U(p;N) = \theta\, N^{\alpha} - pN$$

where θ is a measure of the sales impact generated by a campaign. (This model can easily be extended to the multi-plan case for both non-substitutable and substitutable plans.) The first term of the net utility is the benefit provided by the N impressions and the second one is the amount paid. The parameter θ is advertiser dependent and is taken to be uniformly distributed on $[0,\Theta]$. When the parameter $\alpha<1$, it depicts a repetition wear-out, i.e., a diminishing marginal benefit of repeatedly reaching the same individuals. When $\alpha>1$, it depicts an increasing marginal benefit whereby the number impacted is larger than those that saw the ad. In the case where $\alpha=1$, the price demand function is independent of the number of impressions contracted. The nominal demand rate $\lambda(p;N)$ is given by

$$\lambda(p;N) = \Lambda\mathbb{P}(U(N)\geq 0) = \Lambda\Big(1-\Theta^{-1}\,p\,N^{1-\alpha}\Big) \qquad \text{or, equivalently,} \qquad p(\lambda;N) = \Theta\Big(1-\frac{\lambda}{\Lambda}\Big)N^{\alpha-1}.$$

Furthermore, the revenue rate achieved is given by $r(\lambda;N) := \lambda\, p(\lambda;N)\, N$ and we denote its maximizer $\bar{\lambda} = \arg\max\{r(\lambda;N) : \lambda\geq 0\}$. We have,

$$\bar{\lambda} = \frac{\Lambda}{2} \quad ; \quad \bar{p} = \frac{\Theta}{2\Lambda}N^{\alpha-1} \quad ; \quad r(\bar{\lambda};N) = \frac{\Lambda\,\Theta}{4}N^{\alpha}.$$

This is a linear price demand function where the demand rate decreases as the price increases (all else constant) and thus is consistent with the general price demand functions available in the literature. In the case where $\alpha=1$, the price demand function is independent of the number of impressions contracted. When $\alpha>1$, the demand rate increases when the number of impressions offered increase (all else constant). This depicts some economies of scale behavior, whereby the price per impression decreases with larger N. In the case where $\alpha<1$, the demand rate decreases as the number of impressions increase. By contracting more impressions the advertiser's additional cost is increasing marginally more than the corresponding revenues generated. Hence, the advertisers ready to book an order decreases. Finally, the price and demand rate maximizers are constant independent of the number of impressions.

This price-demand function can easily be extended to the multi-plan case for both non-substitutable and substitutable plans.

Budget-based demand function

In this model, we assume that advertisers approach the website while having a budget constraint β (equivalent to a reservation price for the entire campaign) to spend and a minimum number of viewers ν to reach. These two factors (budget and reach) can be the output of an optimization problem that the advertiser runs before approaching the web publisher (see Zhao (2000)). From the web publisher side, these two thresholds are considered to be random across the advertisers. For tractability, we assume ν to be a uniform random variable on $[0, M]$ and β a normally distributed random variable with mean $h(\nu) := d\,\nu + g$ and a standard deviation σ_b. (The function h could be obtained through a linear regression on available data and models the correlation between the two variables.) We write $\beta = d\,\nu + g + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma_b)$. We consider the setting where multiple plans, (N_j, p_j), are offered to a single class of advertisers (identified by d, g, M, σ_b). The advertisers book a campaign only if their budget constraint is satisfied and their reach target is fulfilled (i.e., $\nu < N_j$ and $\beta > p_j N_j$). Among those plans that satisfy both constraints, we *assume* that they pick the plan that delivers the highest number of impressions.[6] The number of impressions in this model impacts the demand differently than in the utility based model. We introduce the following notations: $\epsilon_j := \epsilon + m_j$ is a normally distributed random variable with mean $m_j := p_j N_j - g$ and standard deviation σ_b.

We have that for $1 \le j \le J$,

$$\begin{aligned}
\lambda_j(p; N) &= \Lambda\,\mathbb{P}(0 \le \nu \le N_j,\, p_j N_j \le \beta \le p_{j+1} N_{j+1}) \\
&= \Lambda \mathbb{P}(0 \le d\nu \le dN_j,\, p_j N_j - g + \epsilon \le d\nu \le p_{j+1} N_{j+1} - \xi + \epsilon) \\
&= \Lambda \mathbb{E}_\epsilon \mathbb{P}\big([\epsilon]_j^+ \le d\nu \le \epsilon_{j+1} \wedge d\epsilon_j \le dN_j\big) \\
&= \Lambda \mathbb{E}_\epsilon \mathbb{P}\big([\epsilon_j \wedge dN_j]^+ \le d\nu \le [\epsilon_{j+1} \wedge N_j]^+\big) \\
&= \frac{\Lambda}{dM}\big(\mathbb{E}[\epsilon_{j+1} \wedge dN_j]^+ - \mathbb{E}[\epsilon_j \wedge dN_j]^+\big) \\
&= \frac{\Lambda}{dM}(G_{j+1}(N_j) - G_j(N_j)).
\end{aligned}$$

We denote by $G_j(N) = \mathbb{E}[\epsilon_j \wedge dN]^+$ with $G_{J+1}(N) = dN$. Note that the second to last equality above is simply the difference between two normally distributed random variables with respective means m_j and m_{j+1} truncated at both 0 and N_j. Simple calculations related to the truncated normal distribution show that

$$\begin{aligned}
G_j(N) &= \big[m_j \Phi(m_j) - (m_j - dN)\Phi(m_j - dN) - \big(\phi(m_j) - \phi(m_j - dN)\big)\sigma_b\big] \\
&= dN + \sigma_b(\Psi(m_j/\sigma_b) - \Psi((m_j - dN)/\sigma_b)),
\end{aligned}$$

[6] Simple modifications can be made to model an advertiser that would select the cheapest plan among all plans that satisfy both constraints.

where $\Psi(x) = \phi(x) - x\bar{\Phi}(x)$ as defined in 5.2. Finally, we note that simple modifications need to be made to model an advertiser that would pick the cheapest product (instead of the most expensive one) that meet his constraints.

B3. Data Estimation

We next estimate the price-demand relationships. We focus on the two models defined above and based on the data available to us from Aller Internett, we try to generate reasonable estimates of the models' parameters. We pick a particular ad (of size 468x400) as a representative ad of our aggregated analysis and notice price variations that often result from negotiations and could reflect how much the publisher needed to lower the price to get the contract. For the utility price-demand function we set the parameter Θ to correspond to the largest price recorded, $\Theta = 0.09$. To estimate the budget function properly we would need to have access to the budget of the individual advertisers. However, the price times the impressions requested give us a sense of the budget and we use it as a proxy. We then perform a regression and estimate $g = 6,000$, $d = 0.07$ and $\sigma = 15,000$ with an adjusted R^2 of 92%. Based on the orders considered we set $M = 3,000,000$ and based on the amount of orders during the horizon of the data and by scaling it by their market share we set $\Lambda = 30$.

C1. Proof of Proposition 1

Proof. The fluid optimization problem can be states as follows:

$$\begin{aligned}
&\max_{\boldsymbol{\lambda},\boldsymbol{\kappa},\mathbf{f}} \sum_{j=1}^{J} \lambda_j \, p_j(\boldsymbol{\lambda};\mathbf{N}) \, N_j \\
&s.t. \quad \rho_j := \lambda_j \, N_j/(s\,\mu\, f_j) \leq 1, \qquad j = 1,2,...,J \\
&\qquad \sum_{j=1}^{J} f_j \leq 1.
\end{aligned}$$

The KKT conditions for this problem are:

$$\begin{aligned}
-\sum_{j\in\mathcal{J}} \partial_{\lambda_i} r_j(\boldsymbol{\lambda};\boldsymbol{N}) + \frac{z_i N_i}{s\mu f_i} = 0, \qquad & i = 1,2,...,J \\
-\frac{z_i \lambda_i N_i}{s\mu f_i^2} + m = 0, \qquad & i = 1,2,...,J \\
z_i(\frac{\lambda_i N_i}{s\mu f_i} - 1) = 0, \qquad & i = 1,2,...,J \\
m(\sum_{j\in\mathcal{J}} f_j - 1) = 0 & \\
\frac{\lambda_i N_i}{s\mu f_i} - 1 \leq 0, \qquad f_i,\ z_i \geq 0, \qquad & i = 1,2,...,J \\
\sum_{j\in\mathcal{J}} f_j \leq 1, \qquad m \geq 0, &
\end{aligned}$$

where the z_i's are the Lagrange multipliers. We divide the solution space into two parts:

i.) If $\sum_{j\in\mathcal{J}}\bar{\lambda}_j N_j/(s\mu)\leq 1$ then $\bar{\rho}_i:=\bar{\lambda}_i N_i/(s\mu)\leq 1,\ i=1,2,...,J$. If we set $f_i=\bar{\rho}_i$ we have a feasible solution with $\sum_{j\in\mathcal{J}} f_j=\sum_{j\in\mathcal{J}}\bar{\rho}_j\leq 1$ that satisfies the KKT conditions with $m=0$, $z_i=0$, and $\sum_{j\in\mathcal{J}}\partial_{\lambda_i} r_j(\boldsymbol{\lambda};\boldsymbol{N})=0$, $i=1,2,...,J$. Hence, $\lambda_i^0=\bar{\lambda}_i$ and $f_i^0=\frac{\lambda_i N_i}{s\mu}$, $i=1,2,...,J$.

ii.) If $\sum_{j\in\mathcal{J}}\bar{\lambda}_j N_j/(s\mu)>1$ we have $\sum_{j\in\mathcal{J}}\bar{\lambda}_j N_j/(s\mu f_j)>1$ because $f_j\leq 1$, $j=1,2,...,J$. Let us now assume that $\bar{\rho}_j\leq 1$, $j=1,2,...,J$. This means using the fifth KKT condition that $\sum_{j\in\mathcal{J}}\bar{\rho}_j\leq\sum_{j\in\mathcal{J}} f_j\leq 1$, which leads to a contradiction. Hence, the revenue maximizing solution, $\bar{\boldsymbol{\lambda}}$, cannot satisfy the KKT conditions and based on the first KKT condition there exists i such that $z_i\neq 0$, which means that $m\neq 0$ (second condition). Furthermore, having $m\neq 0$ means that $z_i\neq 0$, $i=1,2,...,J$ and that we must have $\sum_{j\in\mathcal{J}} f_j=1$ to satisfy the fourth KKT conditions. By setting $f_i=\bar{\rho}_i$, $i=1,2,...,J$, and solving $\sum_{j\in\mathcal{J}}\partial_{\lambda_i} r_j(\boldsymbol{\lambda};\boldsymbol{N})=\frac{z_i N_i}{s\mu f_i}=\frac{m f_i}{\lambda_i}=\frac{m N_i}{s\mu}$ with $\sum_{j\in\mathcal{J}} f_j=1$, we have a solution satisfying all KKT conditions.

□

C2. Proof of Proposition 6

We already know that r is concave in λ and independent of κ. We move now to study the convexity of $c(\lambda,\kappa)$ through a computation of the different derivatives of the cost function with respect to λ and κ. We will prove also through these simple but tedious calculations the monotonicity of ϖ_a stated in the Proposition 4. Without loss of generality we let $c=1$. Let $x=\sqrt{s\kappa}$ and and $y=\lambda$. Recall that $\Psi'\leq 0$ and $\Psi''\geq 0$. We write, $\alpha_\kappa=\sqrt{s\kappa}-\lambda T/\sqrt{s\kappa}=x-T\,y/x:=f(x,y)$. It is easy to that

$$f'_x(x,y)=1+T\,y/x^2\geq 0 \qquad \text{and} \qquad f''_{x,x}(x)=-2T\,y/x^3\leq 0.$$

On the other hand,

$$f'_y(x,y)=-T/x \qquad \text{and} \qquad f''_{y,y}\equiv 0.$$

Finally,

$$f''_{y,x}=1+T/x^2.$$

We move to $c(\lambda,\kappa)=c\,\lambda\,\varpi_a(\lambda,\kappa):=g(x,y)=x\Psi(f(x,y))$. We look at the derivative with respect to λ.

$$g'_y(x,y)=x f'_y\Psi'(f)=-T\Psi'(f)\geq 0 \qquad \text{and} \qquad g''_{y,y}(x,y)=x f'^2_y\Psi''(f)=T^2/x\Psi''(f)\geq 0.$$

We conclude that $c(\cdot,\kappa)$ is increasing concave in λ. We move now to the cross derivative.

$$g''_{y,x}=-T f'_x\Psi''(f)\leq 0. \tag{c1}$$

We conclude that the cross derivative of c is non-positive. Finally, we look at the the derivatives with respect to κ.

$$\begin{aligned} g'_x(x,y) &= \Psi(f) + xf'_x\Psi'(f) = \Psi(f) + x(1+Ty/x^2)\Psi'(f) \\ &= \phi(f) - (x - Ty/x)\bar{\Phi}(f) - x(1+Ty/x^2)\bar{\Phi}(f) \\ &= \phi(f) - 2x\bar{\Phi}(f). \end{aligned} \tag{c2}$$

Finally,

$$\begin{aligned} g''_{x,x}(x,y) &= f'_x\Psi'(f) + f'_x\Psi'(f) + xf''_{x,x}\Psi'(f) + xf'^2_x\Psi''(f) \\ &= \Psi'(f)(2f'_x + xf''_{x,x}) + xf'^2_x\Psi''(f) \\ &= 2\Psi'(f) + xf'^2_x\Psi''(f) \\ &= -2\bar{\Phi}(f) + x(1+Ty/x^2)^2\phi(f). \end{aligned} \tag{c3}$$

We take one further derivative of g

$$\begin{aligned} g'''_{x,x,x}(x,y) &= 2f'_x\phi(f) + (x(1+Ty/x^2)^2)'\phi(f) - x(1+Ty/x^2)^2 f f'_x\phi(f) \\ &= \phi(f)(2f'_x + (x(1+Ty/x^2)^2)' - x(1+Ty/x^2)^2 f f'_x) \\ &= \phi(f)(2(1+Ty/x^2) - x(1+Ty/x^2)^3(x - Ty/x) \\ &\quad + (1+Ty/x^2)^2 + x(2(1+Ty/x^2)(-2Ty/x^3))) \\ &= \phi(f)(2(1+Ty/x^2) - x^2(1+Ty/x^2)^3(1-Ty/x^2) \\ &\quad + (1+Ty/x^2)^2 - 4Ty/x^2((1+Ty/x^2))^2) \end{aligned} \tag{c4}$$

The sign of the previous equation is the same as the sign of the term inside the parenthesis. We look at that term and show

$$\begin{aligned} &(1+Ty/x^2)(2 - x^2(1+Ty/x^2)^2(1-Ty/x^2) + 1 + Ty/x^2 - 4Ty/x^2) \\ &\quad = (1+Ty/x^2)(3 - 3Ty/x^2 - x^2(1+Ty/x^2)(1-Ty/x^2)) \\ &\quad = (1+Ty/x^2)(1-Ty/x^2)(3 - x^2(1+Ty/x^2)^2) \\ &\quad = (1 - T^2y^2/x^4)(3 - x^2(1+Ty/x^2)^2 \\ &\quad = \big(1 - (\lambda T/(s\kappa))^2\big)(3 - s\kappa(1+\lambda T/(s\kappa))^2). \end{aligned} \tag{c5}$$

We divide the positive line in two regions, depending on whether $\varrho < 1$ or $\varrho \geq 1$. Hence, if $\varrho \geq 1$ (i.e. $\kappa < \lambda T/s$), then $g'''_{x,x,x}$ is positive (as long as $s\kappa > 3/4$). Hence, $g''_{x,x}$ is increasing in this domain. Eventually when $\varrho < 1$ and κ is large enough, $g'''_{x,x,x}$ becomes negative and $g''_{x,x}$ decreases. It is easy to see that g''' will change sign once. By noticing that $g''_{x,x} = (0,y) = -2$, that for $\varrho = 1$, $f \equiv 0$ and $g''_{x,x}(x,y) = -1 + 4\,\phi(0)\cdot x \geq -1 + 4\,\phi(0)\cdot\sqrt{3/4} > 0$ and $\lim_{x\to+\infty} g''_{x,x}(x,y) = 0$, we conclude that g'' changes sign only once as well and there exists x_0 in the region $\varrho > 1$ where $g''_{x,x} = 0$. We denote $\varrho'_0(\lambda) > 1$ the value of ϱ at that point. This shows that $c_a(\lambda,\kappa)$ is convex in κ as long as $\varrho < \varrho'_0$. If $\varrho > \varrho'_0$, then g'' is negative. The second derivative of $\partial_{\kappa,\kappa} c_a = \sqrt{s}/2\kappa^{-1/2}(-s/(2x^2)\,g'_x + g''_{x,x})$. The term inside the parenthesis is equal to $-s\phi(f)/(2x^2) + s\bar{\Phi}(f)/x - 2\bar{\Phi}(f) + x(1+Ty/x^2)^2\phi(f) =$

$\phi(f)(-s/(2x^2)+x(1+Ty/x^2)^2)+\bar{\Phi}(f)(s/x-2)$, which is again positive for κ small enough and so there exists $\varrho_0''(\lambda)$ such that for $\varrho<\varrho_0''$ c_a is convex. We denote by $\varrho_0=\varrho_0'\vee\varrho_0''>1$. Depending on the values of λ, ϱ_0' could be equal to ∞ and then c_a is convex in κ for all κ.

Finally, we look at the hessian of g which gives us

$$(2\Psi'(f)+xf_x'^2\Psi''(f))\cdot T^2/x\Psi''(f)-(Tf_x'\Psi''(f))^2=2T^2/x\Psi'(f)\Psi''(f)\leq 0.$$

We turn to ii.) We multiply both sides of the constraint by λ and take the derivative in that constraint equation with respect to λ. We get that

$$\begin{aligned}\partial_\lambda c(\lambda,\kappa(\lambda)&=\partial_1 c(\lambda,\kappa(\lambda))+\kappa'(\lambda)\partial_2 c(\lambda,\kappa(\lambda))\\&=T-N\kappa(\lambda)-\lambda N/\mu\,\kappa'(\lambda).\end{aligned}$$

We take the derivative again with respect to λ and obtain that

$$\begin{aligned}\partial_{\lambda,\lambda}c(\lambda,\kappa(\lambda)&=\partial_{1,1}c_a(\lambda,\kappa(\lambda))+2\kappa'(\lambda)\partial_{1,2}c_a(\lambda,\kappa(\lambda))+\kappa''(\lambda)\partial_2 c_a(\lambda,\kappa(\lambda))+\kappa'(\lambda)^2\partial_{2,2}c_a(\lambda,\kappa(\lambda))\\&=-2N/\mu\kappa'(\lambda)-\lambda N/\mu\kappa''(\lambda).\end{aligned}$$

First, we notice that for fixed λ, the derivative with respect to κ of ϖ_a at $\kappa(\lambda)$ is strictly less (in absolute value) than N/μ. Hence, $\partial_2 c_a(\lambda,\kappa(\lambda)+\lambda N/\mu)\leq 0$. Second, for fixed λ, the delay function starts concave in κ and then becomes convex. We saw that the derivative at zero is steeper than the derivative of $T-N\kappa/\mu$. Therefore, the two functions cannot intersect except in the concave area and so $\partial_{2,2}c_a(\lambda,\kappa(\lambda)>0$ and ϱ at such point is always smaller than $\varrho_0(\lambda)$. Applying i.), we conclude that as long as $\partial_{2,2}c_a>0$, then $\kappa''(\lambda)\leq 0$ (i.e. $\kappa(\lambda)$ is concave in λ). Now we go back to the the second derivative of $c_a(\lambda,\kappa(\lambda))$ with respect to λ and observe that it is always positive. □

C3. Proof of Proposition 8

For the sake of the proof, we drop the index N. We recall that

$$W^n\stackrel{d}{=}\max_{r\geq 0}S_r^n(\kappa^n),$$

where $S_r^n(\kappa^n)=\sum_{i=1}^r Y_i^n$ where, $Y_i^n\stackrel{d}{=}Y_1^n\stackrel{d}{=}\sum_{j=1}^{N\kappa^n}u_j^n-\sum_{j=1}^{s\kappa^n}v_j^n$ with $\mathbb{E}Y_1^n=\kappa^n\left(N/\mu^n-s/\lambda^n\right)\leq 0$.

The sequence (λ^n,κ^n) is formed, for every $n\geq 1$, as the solution to the optimization problem (P^n). From the fulfillment constraint we have that $T-N\kappa^n/\mu^n\geq 0$ and hence the sequence $\kappa^n/n\leq\kappa^0:=\mu T/N$. Moreover, the utilization being smaller than one implies that the sequence $\lambda^n/n\leq\lambda^0:=s\mu/N$. Finally, the sequence κ^n/λ^n is also bounded as λ^n is assumed to be away from zero.

Consider any subsequence κ^m/m that converges to $l<\infty$. The finiteness of such limit l implies that $\kappa^m(N/\mu^{m2}+s/\lambda^{m2})\to 0$ as $m\to\infty$. The inter-arrivals of campaigns and viewers are both exponentially distributed, we conclude that the log-moment generating function of the random variable Y_1^n is given by

$$\log \mathbb{E}\exp\theta Y_1^n=\theta\kappa^n\left(N/\mu^n-s/\lambda^n\right)+\theta^2\kappa^n(N/\mu^{n2}+s/\lambda^{n2})+O(n^{-2}). \tag{c6}$$

The first term $\kappa^m\left(N/\mu^m-s/\lambda^m\right)=\kappa^m s/\lambda^m\left(\rho^m-1\right)\leq 0$ and all other terms go to zero with m. We infer that $\limsup_{m\to\infty}Y^m\leq 0$ a.s. The same holds for S_r^m for all r. Therefore, their maximum, $W^m\Rightarrow 0$ as $m\to\infty$. By bounded convergence, $\mathbb{E}W^m=\varpi^m\to 0$, as $m\to\infty$. From the the equality constraint we conclude that $\varpi^m\to T-Nl/\mu$ as $m\to\infty$. This imposes that $l=\kappa_0:=\mu T/N$ and hence the entire sequence κ^n/n converges to κ_0 as $n\to\infty$.

Similarly, consider a subsequence λ^m/m that converges to some finite limit l' as $m\to\infty$. Consider the log-moment generating function with θ replaced by $\theta\sqrt{m}$. The quantity, $m\,\kappa^m(N/\mu^{m2}+s/\lambda^{m2})\to N\kappa_0/\mu^2+s\kappa_0/l'^2$ as $m\to\infty$; While $\limsup_{m\to\infty}\sqrt{m}\,\kappa^m\left(N/\mu^m-s/\lambda^m\right)=\eta\leq 0$ and possibly infinite. Assume that $\eta<\infty$, in this case $\kappa^m\left(N/\mu^m-s/\lambda^m\right)\to 0$ and thus $N\kappa_0/\mu-s\kappa_0/l'=0$, equivalently $l'=\lambda_0$ and so all subsequences, that lead to some η finite have that $\lambda^m/m\to\lambda_0$. Any subsequence that lead to an η infinite will still have to satisfy $\lambda^m/m\to 0$; otherwise, it will generate lower profits at the limit. In the finite case, $\lim_{m\to\infty}\sqrt{m}\,Y^m=Y$ where Y is a normal random variable with mean η and standard deviation $\sigma_0=(N\kappa_0/\mu^2+s\kappa_0/\lambda_0^2)^{1/2}$. As for the delay, we claim that $\sqrt{m}\,W^{N,m}\Rightarrow\max_{r\geq 0}S_r$, where $S_r=\sum_{i=1}^r Y_i$ with Y_i's i.i.d. with $Y_1\stackrel{d}{=}Y$. To prove it, we rely on Theorem 6.1 on page 285 of Asmussen (2003) which only require uniform integrability of $\sqrt{m}\,Y_i^m$, which is guarantied by the fact that $\mathbb{E}m\,Y_1^{m2}\to\sigma_0^2$ as $m\to\infty$. We denote by $\varpi^N=\mathbb{E}\max_{r\geq 0}S_r$ and $\varpi^m=\varpi^N/\sqrt{m}+o(1/\sqrt{m})$ as $m\to\infty$. The rest of the proof follows the exact same steps as in the T-contract case. The parameter η is uniquely selected by maximizing the ratio of the profit in the stochastic setting with that in the fluid setting. If the subsequence indexed by m was selected so that η is infinite, in this case, $\varpi^N=0$ and $\sqrt{m}(T-N\kappa^m/\mu^m)\to 0$ as $m\to\infty$ and thus $\kappa^m=\kappa_0+o(1/\sqrt{m})$, which implies by injecting κ^m in $\sqrt{m}\,\kappa^m\left(N/\mu^m-s/\lambda^m\right)$ and recalling that the latter converge to $-\infty$ that $\lambda^m/m=\lambda_0+l^m$ where $\sqrt{m}\,l^m\to-\infty$. Hence, the demand rate grows at a slower rate than the subsequences corresponding to a finite η. □

C4. General Solution of the Multi-Plan Optimization Problem

We discuss here a formulation of the general case of the multi-plan optimization problem which is helpful in performing the asymptotic analysis. We first formulate the problem as follows

$$\max_{\lambda_j\,;\,j\in\mathcal{J}}\sum_{j\in\mathcal{J}}r_j(\lambda_j;N_j)\,N_j-G(\boldsymbol{\lambda})$$

where,

$$G(\boldsymbol{\lambda}) = \min_{\kappa_j, f_j; j\in\mathcal{J}} \sum_{j\in\mathcal{J}} c_j \lambda_j \varpi_j(\lambda_j, \kappa_j) \tag{c7}$$
$$s.t. \quad \varpi_j(\kappa_j; \lambda_j) = T_j - \frac{N_j \kappa_j}{\mu f_j} \quad \text{and} \quad \sum_{j=1}^{J} f_j = 1.$$

First, we recall that by construction $\lambda_j(T - \varpi_j^T) \leq s\kappa_j$ (as the average number of ads being displayed at every point in time is less than the slots available by design). Thus, $\lambda_j(T - (T_j - \frac{N_j\kappa_j}{\mu f_j})) \leq s\kappa_j$ which implies that the constraints $f_j \leq \lambda_j N_j / s\mu$ are always satisfied. Recalling that ϖ_j^T is decreasing in κ_j, we conclude that for a given demand rate λ_j the constraint on the delay function determines completely κ_j as a function of f_j. We denote by $\kappa_j(f_j; \lambda_j)$ the value of κ_j for each f_j given λ_j. It is not hard to see that $\kappa_j(f_j; \lambda_j)$ is increasing in f_j while ϖ_j^T is decreasing in κ_j and thus the constraint $\sum_{j=1}^{J} f_j \leq 1$ ought to be binding. The above minimization problem is separable and thus the optimality conditions can be re-written as follows

$$\frac{\partial}{\partial f_j}\Big[c_j N_j / T_j \lambda_j \varpi_j(\lambda_j, \kappa_j(f_j; \lambda_j))\Big] = m_A, \tag{c8}$$
$$\lambda_j = \arg\max_{\lambda_j} \Big\{\lambda_j p_j(\lambda_j) - c_j N_j / T_j \lambda_j \varpi_j(\lambda_j, \kappa_j(f_j; \lambda_j))\Big\}, \tag{c9}$$
$$\sum_{j=1}^{J} f_j = 1, \tag{c10}$$

where m_A (a Lagrange multiplier) is a constant independent of j. Practically, we first solve the second set of conditions, which is an optimization problem similar to the single plan problem. That solution gives the optimal demand rates as a function of the proportion, $\lambda_j(f_j)$. We integrate these in the first set of conditions. For each reasonable value of m_A, the set of equations in (c8-c10) admit a unique solution f_j, (if there are multiple solutions then one picks the largest one knowing that κ is in increasing in f and ϖ^T is decreasing in κ). Hence, for each value of the constant m, we can construct $\sum_{j=1}^{J} f_j$ and then pick the value of m that yields $\sum_{j=1}^{J} f_j = 1$.