Aalto University
School of Business

Profit maximization through budget allocation in display advertising

Master's Thesis
Suvi Mäkinen
Aalto University School of Business
Finance
Fall 2017

Author Suvi Mäkinen

Title of thesis Profit maximization through budget allocation in display advertising

Degree Master's degree

Degree programme Finance

Thesis advisor(s) Vesa Puttonen, Sami Torstila

Year of approval 2017 **Number of pages** 45 **Language** English

Abstract

Online display advertising provides advertisers a unique opportunity to calculate real-time return on investment for advertising campaigns. Based on the target audiences, each advertising campaign is divided into sub campaigns, called ad sets, which all have their individual returns. Consequently, the advertiser faces an optimization problem of how to allocate the advertising budget across ad sets so that the total return on investment is maximized. Performance of each ad set is unknown to the advertiser beforehand. Thus the advertiser risks choosing a suboptimal ad set if allocating budget to the one assumed to be the optimal. On the other hand, the advertiser wastes money when exploring the returns and not allocating budget to the optimal ad set.

This exploration vs. exploitation dilemma is known from so called multi-armed bandit problem. Standard multi-armed bandit problem consists of a gambler and multiple gambling-slot machines i.e. bandits. The gambler needs to balance between exploring which of the bandits has the highest rewards and simultaneously maximising the reward by playing the bandit having the highest return. I formalize the budget allocation problem faced by the online advertiser as a batched bandit problem where the bandits have to be played in batches instead of one by one. Based on the previous literature, I propose several allocation policies to solve the budget allocation problem. In addition, I use an extensive real world dataset from over 200 Facebook advertising campaigns to test the performance impact of different allocation policies.

My empirical results give evidence that the return on investment of online advertising campaigns can be improved by dynamically allocating budget. So called greedy algorithms, allocating more of the budget to the ad set having the best historical average, seem to perform notable well. I show that the performance can further be improved by dynamically decreasing the exploration budget by time. Another well performing policy is Thompson sampling which allocates budget by sampling return estimates from a prior distribution formed based on historical returns. Upper confidence and probability policies, often proposed in the machine learning literature, don't seem to apply that well to the real world resource allocation problem.

I also contribute to the previous literature by providing evidence that the advertiser should base the budget allocation on observations of the real revenue generating event (e.g. product purchase) instead of using observations of more general events (e.g. clicks of ads). In addition, my research gives evidence that the performance of the allocation policies is dependent on the number of observations the policy has to make the decision based on. This may be an issue in real world applications if the number of available observations is scarce. I believe this issue is not unique to display advertising and consequently propose a future research topic of developing more robust batched bandit algorithms for resource allocation decisions where the rate of return is small.

Keywords display advertisement, resource allocation, multi-armed bandit, machine learning

Tekijä Suvi Mäkinen

Työn nimi Tuoton maksimointi markkinointi budjetin optimoinnilla internet mainonnassa.

Tutkinto Kauppatieteiden maisteri

Koulutusohjelma Rahoitus

Työn ohjaaja(t) Vesa Puttonen, Sami Torstila

Hyväksymisvuosi 2017 **Sivumäärä** 45 **Kieli** Englanti

Tiivistelmä

Internet mainonta mahdollistaa ensi kertaa tarkan tuottoprosentin laskemisen mainos kampanjoille. Jokainen mainoskampanja on jaettu kohdeyleisöjen perusteella mainosjoukkoihin, joiden tuottoprosenttien perusteella määräytyy koko kampanjan tuottoprosentti. Mainostajan on päätettävä, miten jakaa mainoskampanjan kokonaisbudjetti eri mainosjoukkojen välillä, siten että koko kampanjan tuottoprosentti olisi mahdollisimman suuri. Koska mainostaja ei varmasti tiedä kunkin mainosjoukon todellista tuottoprosenttia, joutuu hän voiton maksimoimisen lisäksi saman aikaisesti käyttämään budjettia mainosjoukkojen testaamiseen selvittääkseen millä mainosjoukoista on paras tuotto.

Tämä kompromissi testauksen ja tuoton maksimoimisen välillä on tuttu ns. monikätinen rosvo ongelmasta. Ongelmassa uhkapelurilla on useita vanhan ajan pelikoneita, eli yksikätisiä rosvoja, joilla jokaisella on oma tuottojakaumansa. Peluri haluaa samanaikaisesti selvittää millä pelikoneista on paras tuottojakauma sekä maksimoida voittonsa pelaamalla tätä parasta konetta. Muotoilen perinteisen monikätinen rosvo ongelman budjetti allokointiin sopivaksi niputtamalla pelikoneiden pelaamisen ryppäiksi, siten että jokaisella päätöksenteko hetkellä pelaaja kokeilee useampaa pelikonetta samanaikaisesti sen sijaan että pelaisi niitä yksitellen. Aikaisempaan kirjallisuuteen pohjautuen, esitän useampaa algoritmia budjetti allokointi ongelman ratkaisemiseksi. Lisäksi testaan näiden algoritmien toimivuutta yli 200 oikeaa Facebook mainoskampanjaa käsittävällä aineistolla.

Empiiriset tulokseni osoittavat, että dynaaminen budjetin allokointi parantaa kokonaistuottoprosenttia verrattuna budjetin jakamiseen tasaisesti eri mainosjoukkojen välillä. Jo yksinkertaisesti painottamalla budjettia parhaan historiallisen keskiarvon omaaville mainosjoukoille, saatiin huomattava parannus tuottoprosenttiin. Vielä paremmat tulokset saatiin, kun painotusta parhaalle mainosjoukolle kasvatettiin ajan kuluessa. Myös Thompson otanta, missä mainosjoukkojen tuotto-odotukset arvioidaan otannalla oletetusta tuottojakaumasta, näytti toimivan hyvin budjetti allokointiin. Sen sijaan, paljon kirjallisuudessa tutkitut ylempään luottamusväliin ja todennäköisyys painotuksiin perustuvat mallit taipuivat huonosti tosielämän allokointi ongelmaan.

Tutkimukseni antaa viitteitä siitä, että budjetti allokointi tulisi tehdä perustuen havaintoihin todellisesta tavoitetapahtumasta (esim. ostotapahtumasta), eikä ylemmän tason tapahtumista (esim. mainoksen klikkaus). Toisaalta algoritmit näyttäisivät pärjäävän sitä huonommin, mitä vähemmän havaintoja niillä on käytössä. Tämä muodostuu ongelmaksi mm. mainostajan budjetti allokoinnissa, kun allokoinnin perusteena käytetään harvinaisempia tavoitetapahtumia. Havaintojen suhteellinen vähäisyys on todennäköinen myös muissa tosielämän sovelluksissa, joten ehdotan tulevaisuuden tutkimukselle mallien kehittämistä riippumattomammiksi havaintomäärästä.

Avainsanat Internet mainostaminen, budjetti allokointi, monikätinen rosvo, koneoppiminen

Table of Contents

1. Introduction 1

2. Objectives of the study 4

3. Context and previous literature 5

3.1. Basic concepts of display advertising 5

3.2. Multi armed bandit model 9

3.3. Related literature and previous findings 11

4. Hypotheses 16

5. Methods & data 18

5.1. Formalizing budget allocation as multi armed bandit problem 18

5.2. Stochastic bandit polices 19

5.2.1. Greedy approaches 19

5.2.2. Confidence bound estimation strategies 20

5.2.3. Probability matching 22

5.2.4. Randomized probability matching (Thompson sampling) 23

5.3. Data 25

5.4. Distributional simulation and sequential experiment 26

6. Results 29

7. Conclusions 37

8. References 40

1. Introduction

The emergence of Internet has had a huge impact on multiple industries. One notable change has been the revolution of the advertising industry. During the past years, online advertising has grown into a multibillion-dollar business, hitting over $72.5bn a year in 2016. With 16% compound annual growth rate online advertising has exceeded the global TV advertising and become the largest single advertising channel during 2016 (IAB/PwC, 2016). Nevertheless, taken into account the increasing number of internet users and the emergence novel business models such as e-commerce, the story of internet advertising appears to have just begun.

The strongest growth in online advertising spend is is currently seen in the field of display advertising. (IAB/PwC, 2016) The display advertising is generally defined as advertising where the targeted audience is reached out via some sort of visual advertisement, e.g. banner or video (Aksakallı, 2012; Sahin Cem Geyik and Dasdan, 2014). Traditionally the display advertisements have been targeted based on browser cookies of the internet users. However, the emergence of the social media sites during the past years has brought available wide range of targeting options based on the user profiles and user information from content browsing (Deshpande et al., 2014). This has led to the rapid evolvement of social media to a significant advertisement vehicle (Okazaki and Taylor, 2013).

One of the unique advantages of online advertising is its ability to provide accurate real-time feedback on customer behavior and closely monitor and measure the performance of the advertising campaigns (Roels and Fridgeirsdottir, 2009). This has effectively changed the whole nature of marketing to demand quantitative approaches and require more sophisticated tools and algorithms (Landry and Vollmer, 2010). The complexity of advertising optimization is especially prevalent in the social media display advertising. In the worst scenario, the advertiser is overwhelmed by the sheer amount of data and unable to utilize it correctly. However, the measurability of the advertising results has also a practical implication that the advertising can be automated and optimized to a great extent. The user specific information, available for advertisers in the social media, could in optimum provide inputs needed for highly automatized advertising. Indeed, if the right optimization strategies could be exploited, there might be a possibility to get relatively close to a system that automatically optimizes itself to

continuously yield profit for the advertiser at minimum cost. Taken into account the aforementioned, it is clear that making the optimal choices in online advertising can be highly beneficial.

One special field of attention is the ability to dynamically adjust the spending in online advertising. This showcases a new approach to budget allocation in contrast to the traditional up-front spending in advertising (Araman and Popescu, 2005). As in many economic applications, the company's (in this context advertiser's) actions are limited by the available budget. In the context of display advertising the budget will in most of the cases be split across multiple sub-campaigns. As the sub-campaigns vary in terms of audiences, content and performance, the advertiser is faced with a practical optimization problem of how to allocate the budget in a way that the return on investment is maximized.

While the display advertising business is experiencing expansive growth, the academic research has only recently started to pay attention to new field of study. Especially the academic research of the display marketing in social media context it is still scarce (Okazaki and Taylor, 2013). The current literature proposes several promising budget allocation methods for display advertising such as non-linear approximation (Aksakallı, 2012; Danaher, 2007), knapsack based multi-armed bandit allocation (Ding et al., 2013; Tran-thanh, 2012) and Bayesian bandit algorithm (Tkachenko, 2014). The research also indicates that relatively simple allocation heuristics can have a positive effect on the performance of the display advertising campaigns (Feldman et al., 2010; Tran-thanh, 2012).

Although the literature has developed several theoretical applications of allocation models, the empirical testing of these models has in general been limited to very specific settings. While the previous literature has some real life experiments, they are based on very limited data containing one or two campaigns and thus lack the applicability to wider generalization (e.g. Sahin Cem Geyik and Dasdan, 2014; Schwartz et al., 2017). Another branch of empirical studies is focused on numerical simulations (e.g. Ding et al., 2013; Tkachenko, 2014; Tran-thanh, 2012) Only few papers have run simulations based on real life data (Aksakallı, 2012; Schwartz et al., 2017) and none has utilized data from social media advertising.

In order to contribute to the scarcity of the empirical evidence of the theoretical models, I use an extensive real-world data set to test the budget allocation methods suggested by the previous literature. I utilize an extensive data from Facebook campaigns that provides several advances compared to the previous literature. Firstly, taken into account that nowadays the

display advertising is dominated by Facebook (Pew Research Center, 2015) using Facebook as the source of the data appears well motivated. This is especially as we test the applicability of the models to the real world optimization problems faced by the practitioners. Also, with 56% compound annual growth rate in social media advertising it is clear that a social media advertising platform context is relevant for the display advertising research.

Secondly, the Facebook data has extensive information of multiple conversion points of the advertisers and thus enables the measurability of the performance of the display advertising with a notable accuracy. While majority of the previous literature has limited the measurement of ad performance to the ad clicks, our dataset extends the performance measurement to the actual conversions that the advertiser wishes to optimize towards. This gives my study a preferable approach from the perspective of the practitioners as focusing on the revenue generating conversions takes full advantage of the traceability of display advertising performance.

Because the focus of this research is to study if the display advertising performance can be improved by optimal budget allocation, the perspective used is mainly the one of the advertiser's. However, it is good to note that our findings can have wider implications in the context of resource allocation optimization under uncertain reward distribution. The resource allocation problem is not a unique dilemma for online advertising but has wide applications among other fields of science such as finance, econometrics and industrial sciences. In the era of automation, computational learning algorithms applied in this study can be utilized in many concepts. Some examples are e.g. portfolio allocation, option pricing or electricity supply management. In the best scenario the empirical results of the allocation algorithms in the context of this study will give guidance for the future research in a wider scope.

The rest of the paper is organized as follows. The section 2 goes through the basic concepts of display advertising and multi armed bandit models as well as reviews the research currently done related to budget allocation problem and relevant multi armed bandit models. Section 3 outlines the research question for this thesis as well as the four hypotheses to be answered. The budget allocation problem is formalized as an multi armed bandit problem and required adjustments to the stochastic bandit polices are introduced in the section 4. In addition, section 4 describes the empirical methods and used data in more detail. Section 5 outlines the findings from the empirical experiments as well as discusses their implications. Final conclusions and further research suggestions are provided in section 6.

2. Objectives of the study

Unlike many other forms of advertising, internet display advertising enables accurate tracking of desired actions and attributing them back to the advertisements themselves. This accountability enables advertiser to receive real time data of the return on investment for each sub-campaign. As the total daily advertising budget is limited, the advertiser has to decide how to split the budget across sub-campaigns so that the total return on investment is maximized. A simple allocation is to split the budget equally across the sub-campaigns. In this study, I aim at finding out if the return on investment can be improved by dynamically adjusting the daily budget allocation based on past performance data from the sub-campaigns.

To test this, I use multiple applications of multi-armed-bandit models. In parallel to investigating if the equal allocation is outperformed by dynamic allocation, my objective is also to find out what kind of models do the job best. This will contribute to both the display advertising literature, by providing better understanding of how the performance of the advertising campaigns can be optimized, as well as to the machine learning literature by giving insight on how the current models preform in real world applications and what kind of needs empirical settings may have for the models. Namely, I investigate

The idea of improving display advertising performance by dynamic budget allocation has been investigated in the previous empirical literature (e.g. Aksakallı, 2012; Sahin Cem Geyik and Dasdan, 2014; Schwartz et al., 2017) However, the previous empirical studies have never been made in the context of social media display advertising, nor have they incorporated more than one advertising campaign. Thus my findings in this study will provide further evidence of whether the online advertisers benefit from dynamically allocating the advertising budget as well as of how the budget allocation should be done in order to maximize the benefit.

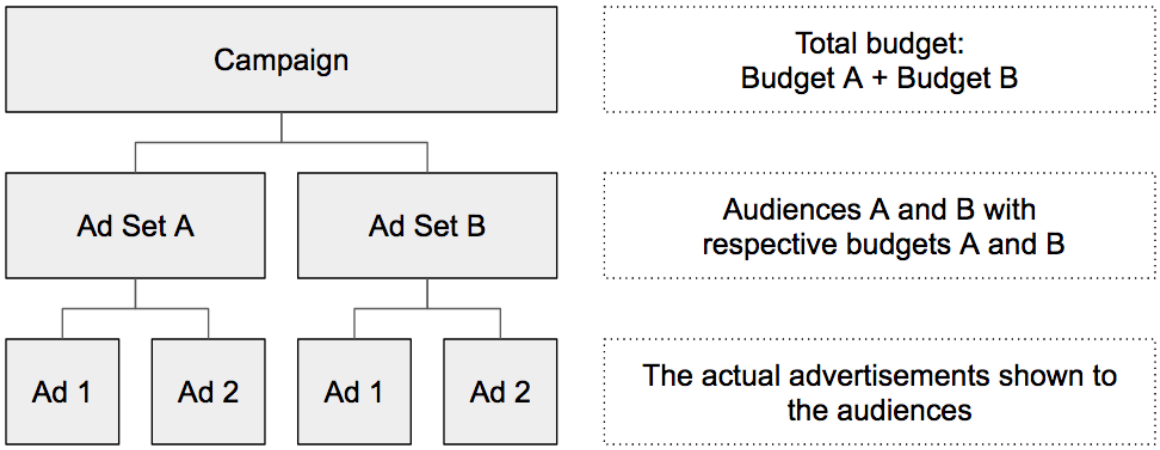
3. Context and previous literature

3.1. Basic concepts of display advertising

Hollis (2005) presents a traditional division of display advertising into two groups based on the ultimate objective of the advertisement, branding and direct response advertising. Branding is defined as long term advertisement that has goals such as generating new customer leads, nurturing existing customer relations and boosting the brand awareness. Direct response advertising on the other hand aims at achieving a measurable and immediate response which in general translates to maximizing attributable revenue for the ad. Although Hollis (2005) suggests that branding and direct response are not mutually exclusive, the latest technological advances in the attribution of conversions has made the online environment especially lucrative for direct response advertising. While the long term effects of branding advertising remain somewhat hard to measure even in the internet display advertising, straight forward measurement of return on investment in direct response campaigns has likely contributed to the popularity of direct response advertising in the online arena (Aksakallı, 2012).

The online advertisers aim to display the best ad for a given user in the best online context. In display advertising this can be done by setting constraints to whom and where the ad can be displayed i.e. audience for the ad. For this purpose the structure of an advertisement campaign is three fold as demonstrated in Figure 1. The advertisements are run within campaigns that are the top level item containing the total budget for the said campaign. The campaign is further divided into sub campaigns i.e. ad sets that define the placement and the targeted audience for the ads. As each of the ad sets have their own individual budget, the advertiser needs to split the total advertising budget across the ad sets. The performance of each ad set on the other hand derives from the combined performance of the ads underneath it. This reduces the budget allocation problem solely on the ad set level.

Figure 1: Structure of a display advertising campaign



Display ads have traditionally been sold through pre-negotiated long-term contracts between publishers and advertisers. However, during the past decade the spot markets have rapidly gained the popularity providing increased liquidity for the publishers and increased reach with granular targeting for the advertisers (Muthukrishnan, 2009). The publishers of the online advertisements typically sell the placements for the advertisements through real time auctions where the advertisers bid for the impressions i.e. the opportunity to show their ads. The bids of the advertisers represent the estimate of utility the advertisers will get from winning the auction. This is ensured by so called sealed second-price auction where the winner of the auction has the highest bid but pays the amount equaling the second highest bid. As the advertiser bids are based on the true value of the impressions, the advertiser's capability to bid is ultimately dictated by the budget of the advertising campaign. (Sahin Cem Geyik and Dasdan, 2014)

Advertisers, on the other hand, want to show their ads so that the number of desired actions will get maximized taken into account the budget they have available. This means that they need to optimize what ad is shown to whom. E.g. a company selling pet toys, would want to pay only for showing an advertisement of dog toys to people who actually have a dog and an advertisement of cat toys to people who have a cat. In addition, the advertiser might want to optimize the context where the ad is shown (e.g. mobile or desktop) and specify the time when the ad is shown. As these settings can all be defined on an ad set level, differentiating ad sets allows advertiser to only show relevant content to target audiences and increase the likelihood of desired actions per the amount of money. As each ad set participates to the auction by itself,

the next optimization question becomes with how to allocate the total budget across the ad sets. This meaning allocating less of the total budget to poorly performing ad sets and more budget to well performing ad sets taken into account that the performance of each ad set is unknown beforehand.

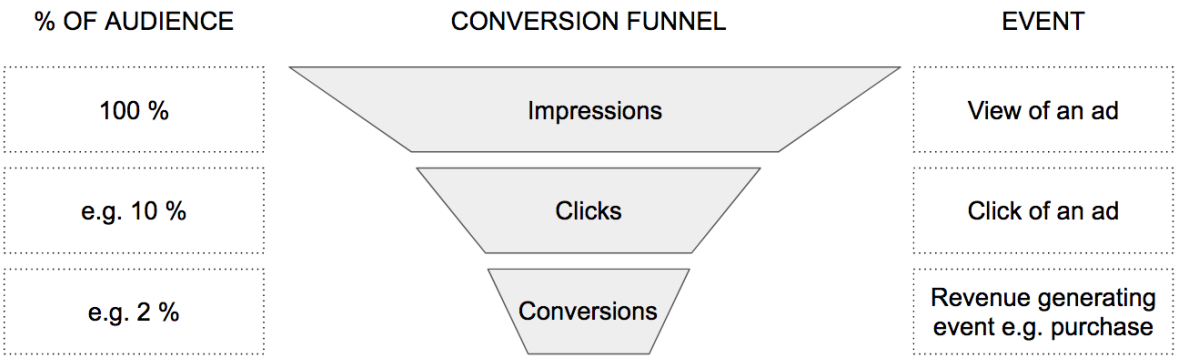
Before being able to optimally allocate the budget of the ad sets, the advertiser needs to define the goal which the advertiser optimizes. The goal itself is defined by the advertiser and the advertising objective. (Lee et al., 2012) For example, for a branding advertiser the optimization goal can very well be just a simple impression i.e. an opportunity show the ad to an user.

Many of the previous literature has been using the number of clicks i.e. user interactions with the ad as a performance measure for the online advertising. This can be widely attributable to the historical popularity of search advertising (Danaher et al., 2010). Search advertising refers to advertisements bought and shown to the audience based on keyword searches in the internet. While search engine based paid-search advertising still remains the biggest online advertising format in terms of revenues, it has been continuously losing share to the display advertising (IAB/PwC, 2016). Clicks, on the other hand, rarely pose the ultimate optimization goal for the display advertiser but are only a medium to the ultimate desired revenue generating action such as purchasing a product or subscribing to an email list.

The context of internet allows advertisers to measure the effectiveness of the advertisement in a highly specific level of target's actions. Indeed, in display advertising the advertiser is able to track actions of the target audience from clicks to website visits to purchase of the advertiser's product. Reaching the desired action for the advertisement is in general referred as conversion. The advertiser defines the conversion based on the advertisement goal but in many cases of direct marketing it is usually the revenue generating action of a customer which actually gives the monetary value to the conversion.

The actions of the target audience can be seen to form a conversion funnel where actions follow each other with decreasing likelihood of happening. For example, an ad that has gotten 100 impressions i.e. times the ad was displayed to target audience may have 10 clicks i.e. times when one from the audience has interacted with the ad and 2 conversions e.g. times when one from the audience has purchase an item the advertisement showcased. Conversion funnel is demonstrated in Figure 2

Figure 2: Conversion funnel in display advertising



The one advance of using Facebook data is that it reports returns for different optimization goals. This allows me to test the impact of budget allocation for different points of the conversion funnel. This broader view completes the previous research that has been on majority concentrating to the optimization of clicks.

In reality conversions are seldom happening directly after the user sees an advertisement. It may for example be that an internet user sees an advertisement of a product but doesn't immediately react to the advertisement. However, it may be that she still recalls the advertisement and later on visits the website of the advertiser and buys the product. This purchase of the product can be attributed as a conversion for the advertisement. In reality it is of course impossible to say if the advertisement was the sole reason for the product purchase, but for practicality this is generally assumed in the context of online marketing. The empirical research also supports this assumption and shows that internet display marketing has a significant positive impact on consumer behavior despite of a lack of clicks (Fulgoni and Mörn, 2009). In particular, it appears that immediate conversion after an internet user sees or clicks an ad is not how internet users make purchases. Instead, they prefer to make purchases on their own accord which can be referred as view-through effect (Bruner and Gluck, 2006).

The previous research has documented little of the used attribution model in the research. In this study, I'll define the attribution window (i.e. the time frame during which a conversion is attributed to an ad) as one day. This is to keep the conversions consistent with the budget allocation interval which is one day with daily budgets. For the data used, the advertiser has been able to specify whether to attribute only conversions after clicks or to take into account also the view through effect. My research will not take any take any position on this but simply follows the advertisers' choices.

Another attribution related dilemma is choosing to which advertisement a conversion is attributed in a case where the user has seen or clicked multiple advertisements. Relatively recent work by Sahin Cem Geyik and Dasdan (2014) describes a multi touch attribution based budget allocation model. The authors point out that the budget allocation should take into account the attribution model that is used to assign conversions to each ad set. Instead of giving all of the credit to the advertisement last clicked or seen, they implement a multi touch attribution based optimization where the conversion is attributed to multiple ads. While their empirical experiment demonstrates superior performance, the multi-touch attribution is in many cases unavailable to display advertisers. As Facebook (together with other major display advertising platforms) is using the last touch attribution of conversions, this study will not address the aforementioned multi touch attribution model.

3.2. Multi armed bandit model

In practice the budget allocation problem can be divided in two parts: how to compute the expected performance for each ad set and how to allocate the budget based on the expectation. As the allocation of the budget is based on the future performance, the advertiser maximizes profit if the best performing ad sets are recognized in the allocation.

A simple heuristic would be to use the past performance data as an indicator for the future and make the budget allocation based on them. Indeed, the previous literature shows evidence that an increase in online advertising performance can be obtained by simply following some simple performance metrics. For example Sahin Cem Geyik and Dasdan (2014) use relatively simple calculation of expected ROI and obtain well increased performance with allocation based on it. While evidence implies that the performance of budget allocation could be improved by pure reliance to the historical performance data, it is not justifiable to assume that the naïve approaches would be optimal.

One of the biggest shortfalls of naïve approaches in display advertising is that the historical data is insufficient to fairly allocate the budget and impressions for the ad sets. For example, allocating future impressions to the current “champion”, that has performed the best to date, is likely to be a myopic strategy. By following this greedy policy, the advertiser is likely to capitalize on chance instead of optimizing profits through learning. (Schwartz et al., 2013) This problem could be addressed by first running the campaign with unique ad set budgets for observation purposes (exploring) and then allocating the budget optimally (exploiting). However, as the exploration of the optimal allocation is costly, the strategy doesn’t seem

optimal in the context of display advertising. These kind of "test drives" seldom get separate budget under the highly dynamic environment of display advertising. Consequently, the display advertiser is faced by an exploration-exploitation dilemma where she needs to simultaneously balance between the cost of not optimizing the budget allocation and the cost of using a non-optimal budget allocation.

Similar sequential decision making problems under uncertainty are faced by multiple real world applications such as medical trials, communication networks and advertising. One of the most studied model for these is the multi-armed bandit problem that provides a theoretical model of exploitation-exploration tradeoff in learning. (Badanidiyuru, 2013) The standard multi-armed bandit problem (MBA) consist of a gambling-slot machine that has K arms each of which delivers rewards that are independently drawn from unknown distributions when the arm is pulled. The gambler can pull one arm at a time to get the respective reward. As the gambler wishes to maximize the sum of rewards in a sequence of pulls, she'll need to find the optimal arm to pull. As the reward distribution of each arm is unknown, the gambler needs to learn which of the arms yields the highest reward.

The fundamental dilemma in MAB problems is the tradeoff between the exploration and exploitation. This tradeoff is because the true reward distributions of each arm are unknown. If the arm selecting policy selects the arm it thinks is optimal (exploitation) it risks pulling a suboptimal arm due to wrongful assumption of the best arm. On the other hand, if the policy keeps trying all the arms and gathering information of the underlying reward distributions (exploration) it fails to exploit the best arm and maximize the total expected payoff. It can be easily seen that the above mentioned problem resembles the learning dilemma faced by the display advertiser. Allocating budget to an ad set is analogous to pulling the arm while the received conversions from the ad set represent the reward for the advertiser. Essentially the advertiser needs to learn the underlying reward distribution (exploration) in order to be able to efficiently allocate the budget (exploitation).

In order to address the exploration vs exploitations tradeoff, the research has suggested several pulling policies to maximize the total payoff. The simplest strategies presented in the literature rely on allocations based on averages. These so called greedy algorithms vary from simple naive algorithms such as ε–first (Even-Dar et al., 2002) or ε–greedy (Watkins, 1989) to more complex methods that are theoretically able converge the optimal policy, such as decreasing ε–greedy (Auer et al., 2002).

A widely investigated strand of pulling policies are so called optimistic strategies or upper confidence bound (UCB) strategies (Kaelbling, 1993). Instead of simply relying on the average return of the arms, UCB policy calculates the upper confidence bound for the average of each arm. The policy then chooses the arm with the highest upper confidence bound and so doing chooses the arm that has the highest optimistic expected reward. The literature has shown that the simple UCB policy is able to theoretically converge the optimal policy (Auer et al., 2002) and developed multiple applications to wider multi armed bandit settings (see e.g. Xia et al., 2017).

Another strand of the MBA allocation strategies are the Bayesian bandits where the views of the best arms are updated based on the observed new evidence. In Bayesian bandit problem, the gambler is assumed to have some knowledge about the estimated probability distribution based on the past experience (priori). After pulling the arm the gambler will observe the outcome and update her knowledge about the underlying distribution accordingly (posterior). On the next round the former posterior becomes the priori. This way the posterior gradually converges to the real underlying reward distribution as the gambler learns more about the optimal arm.

In general, the pulling policies for Bayesian bandits are using the probability matching methods that choose the arm based on a probability distribution reflecting how likely the arm is to be optimal (Vermorel and Mohri, 2005). The Bayesian framework is a natural way to deal with the exploration & exploitation tradeoffs where the accuracy of the underlying estimates need to be taken into account. This is because the approach incorporates the quality of the probability distributions to the final allocations. I.e. takes into account the increase of risk related to distributions that are based on fewer data. (Tkachenko, 2014). One example of the Bayesian approach is the Thompson sampling where the posterior distributions of the arms' rewards are sampled and the arm with the highest sample mean is chosen (Chapelle et al., 2013). Another Bayesian approach is so called Softmax method where the choice of the arm is made based on Gibbs distribution (Chapelle et al., 2013).

3.3. Related literature and previous findings

Despite of the significant market share of the internet display advertising, the literature of the optimization of internet display advertising is relatively scarce. While search advertising has been studied quite extensively, research concentrating to internet display advertising is still really limited. In addition, vast majority of the optimization research is focused on the process

of the ad platform not the advertiser (e.g. Balseiro et al., 2014; Feldman et al., 2010; Ghosh et al., 2009; Lee et al., 2012). While the optimization problems of the ad platforms are out of the scope of this study, it is insightful to look at the recent study of the search advertising.

Many of the research of search advertising optimization is concentrated on the budget optimization problem i.e. how to set the bid under the budget constraint so that the reward is maximized (e.g. Archak et al., 2010; Feldman et al., 2007; Zhou et al., 2008). Zhang et al. (2012) contributed to the previous research by showing the importance of jointly optimizing both bid and budget. They empirically showed with simulations that introducing budget allocation to the optimization problem can significantly improve the performance. Nevertheless, they highlight that in search advertising the optimization problem is not only limited to the budget allocation but requires also to bid optimization.

Although the budget optimization through optimal bidding is widely recognized as an important concept in the search advertising it is further shown that optimization of display advertising can be limited to pure budget allocation i.e. optimizing the budget between ad sets. This is because the paid search bids for keywords that can significantly differ in the levels of competition. Thus the required bid may significantly vary across the keywords and every keyword should have the optimal bid. In the concept of display advertising the advertisers' targeting options are significantly more complex than a pure keyword targeting. Also other factors such as ad quality may affect to the bid. Consequently, a combined optimization of both bid and budget is not in general feasible in the context of display advertising (Sahin Cem Geyik and Dasdan, 2014). In addition, this study focuses on the display advertising ecosystem of Facebook that specifically uses a budget pacing algorithm that automatically optimizes the bid value through the specified time span. Thus it is sensible to limit the optimization problem to budget allocation.

The first study to actually tackle the pure budget allocation problem in display advertising is presented by Danaher et al. (2010) who use multivariate negative binomial distribution to model internet media exposure and maximize the reach of internet display advertising campaigns with the help of non-linear programming. The authors use simulation to compare the model to allocation calculated by complete enumeration and find that they were able to achieve the optimal allocation in a fraction of time. While their model is suitable for optimizing the budget of branding campaigns that use the reach as an optimization goal, it lacks the applicability to direct response campaigns that use conversions as an optimization goal.

Aksakalli (2012) contributes the research by deriving a wider application of the model of Danaher et al. (2010). Likewise, Aksakalli (2012) uses a piecewise non-linear approximation of individual ad revenue functions to formulate a mixed integer program. Multivariate negative binomial distribution was used to model exposure distributions of conversion rates for each ad. Based on these distributions, the author computed the optimal budget allocations. The author shows through simulation that significant increase in the revenue can be obtained by allocating the budgets using the derived model.

While the empirical results of the above studies are encouraging, the provided optimization framework assumes that the ads are bought through guaranteed contracts. Guaranteed contracts are contracts where the advertiser buys a fixed number of impressions over a certain time period at a predefined price. practicality of the model may be questioned. While this kind of pricing is still used to some extent, the spot market of display advertisement as become the prominent market place (Ghosh et al., 2009). Also, the generalized model derived by Aksakalli (2012) requires substantial estimation time and is relatively complex. While this doesn't pose a problem in the context of guaranteed contracts, the framework is hardly applicable in the context of spot markets where the allocation decision is made on the daily basis and no separate exploration time frame is provided. Another shortfall of the aforementioned approach is that it assumes that the conversions follow the negative binomial distribution. Although the empirical literature implies that click through rates approximately follow this distribution (Danaher, 2007) there is no research showing that the distribution could be generalized to accurately measure other conversion rates.

A theoretical line of research with a novel approach to general budget allocation problem is started by Tran-thanh (2012) who is first to introduce a multi-armed bandits (MAB) with budgets or so called budgeted bandits. In the most common MAB setting, pulling an arm is not costly and thus any arm can be pulled arbitrary many times during the agent's operating time. However, this doesn't hold true in many real world applications where the arm pulling is limited by a cost and and a total budget. This is also the case in display advertising.

(Tran-thanh, 2012) extends the standard multi-armed bandit problems to include a fixed cost and budget limitation and uses ε-first, upper confidence based (UCB) and declining ε-greedy approaches to determine the best allocations. The simulations prove that although theoretically upper confidence based policy should be able to converge the optimal policy, it is overpowered by the weaker theoretical guarantees having ε-first policy. Both simplicity and

theoretical guarantees having declining ε-greedy policy is shown to have the best performance in simulations.

Work by Tran-thanh (2012) is followed by a considerable amount of literature deriving UCB aproaches to different kinds of budgeted bandit settings. Ding et al. (2013) contribute to the previous work by extending the upper confidence based multi-armed bandit model to model a variable cost instead of a fixed one. This is done by using the lower bound of the expected cost instead of assuming a known cost. (Xia et al., 2015a) contribute to the research by expanding the model from assuming discrete costs to assuming continuous costs. Other related studies are e.g. Slivkins (2013) Xia et al. (2016) and Xia et al. (2017). Inspired by the UCB literature revolving around the budgeted bandit problem Xia et al. (2015b) also extend another well studied multi-armed bandit policy, Thompson sampling for budgeted bandits

While the above mentioned budgeted bandit literature shows promising results through theoretical analysis and simple simulations, it lacks the application to budget allocation problem of many real world situations. Namely the budgeted bandit setting assumes that the arms are pulled consecutively and rewards and costs are observed immediately after each arm pull. Thus the budget is consumed simultaneously with observing the costs and rewards of the arms. This doesn't apply to display advertising setting where the whole daily budget is allocated to all of the ad sets and only after that the reward is observed. Recent work by Perchet et al. (2016) introduces an new MAB setting called batched bandits where the decisions to pull arms are made in batches and the rewards of the arms are observed simultaneously once the batch is played. The first setting only incorporates two armed batched bandits but Jun et al. (2016) extend the setting to n-armed bandits.

The current literature applying MAB policies to budget allocation in display advertising setting is limited to only few studies. Sahin Cem Geyik and Dasdan (2014) use a simple greedy algorithm to allocate the budget for ad sets. While their empirical study indicated that even this simple allocation policy improves the performance, it is only limited to one campaign and may thus not be generalized more widely.

A study by Tkachenko (2014) employs the probabilistic Bayesian bandit approach to the budget allocation problem. Following to widely used practice the paper uses soft max method to allocate the budget between the ad sets. The author also tests the derived algorithms via simulation and shows that when using probability matching in the budget allocation the amount of conversions is substantially closer to the optimum than when using simple greedy

allocation. The simulation however is not based on real data but generated by drawing conversions from beta binomial distribution.

The most comprehensive study by date is done by Schwartz et al. (2017). They continue the Bayesian approach and apply a Thompson sampling based method to allocate impressions across campaigns. The authors find significant improvement in their online field experiment. In addition, they test several MAB allocation policies with simulation and find equally encouraging improvement in campaign performance. However, the authors don't apply the batched bandit setting in the simulations but instead use simplification from budgeted bandits which may not be optimal. In addition, the paper uses impressions as the budget, not the actual costs. This again isn't directly applicable to the spot markets of display advertising where also the cost of impressions affects the optimal policy.

As the above literature review demonstrates, the empirical research of the performance of budget allocation polices in the context of display advertising is notably scarce. While the literature covers many theoretical models for budgeted multi armed bandit problems, it lacks the empirical comparison of models in real worlds applications. This paper aims to complete the line of MAB literature and provide a real world framework, namely display advertising, in which to evaluate the empirical performance of the models.

In addition to contributing to the model focused machine learning literature, I bring valuable insight to the relatively new online advertising research. This will have direct implications for the practitioners in this multibillion dollar business around the globe. While the approach of this paper is focused on the perspective of the online advertisers, it is also good to note that the resource allocation problem isn't only applicable on that setting. Thus the research can also be seen to contribute to a wider literature of the resource allocation under uncertainty. Providing a practical application for resource allocation problem is likely to pinpoint the benefits and shortages of the current research which hopefully will guide future research to the direction that can benefit multiple fields of science such as clinical trials, economics, finance, or industrial sciences.

4. Hypotheses

The main research question of this study is to find out if the performance of display advertising campaigns can be improved by optimal budget allocation. Answering this question gives us implications on if the practitioners actually should do dynamic budget allocation. If no performance improvement can be obtained through budget allocation, there is no point of using resources to it. Following the previous empirical research, we assume that already using some naïve approaches to the budget allocation can yield increase in the advertising performance. This gives us the first hypothesis

H1: Dynamically optimizing budget allocation has a positive effect to the returns of advertising

If the budget allocation can be optimized, we further want to know how to get closest to the optimal allocation. Even if we found that naïve budget allocation approaches increase the performance of display advertising, we don't assume them to be the optimal. This is because naïve approaches are likely to exhibit outlier behavior that leads to skewed estimates. For example, an ad set having one impression would have an extremely high CTR of 1 if that impression happened to convert to a click. (Tkachenko, 2014) Consequently, algorithms able to take into account the distribution of the return estimates should outperform those simply relying to the observed averages (so called greedy algorithms)

H2: Greedy budget allocation algorithms based on simple average returns are outperformed by more sophisticated models that take into account the uncertainty of return estimates

From the theoretical perspective the learning algorithm should perform better the more it is able to converge to the optimal policy as it accumulates observations. An algorithm with no adjustments on the exploration and exploitation weights can't even in theory tap the theoretical optimal returns as it will always spend some fixed amount of budget to exploration. This was even if the algorithm knew which of the ad sets was optimal. This leads us to my third hypothesis.

H3: Algorithms dynamically adjusting between exploration and exploitation should outperform those using fixed proportion of budget for exploration.

In general, the previous research has concentrated on optimizing towards impressions or clicks. While these are the more conventional measurements for advertising, the current technology also enables tracking conversions, which are the actual revenue generating events

derived from the advertisements. As impressions and clicks usually are events preceding conversions, they are commonly used as a proxy for conversions also among practitioners. This may be due to e.g. technical knowledge and effort needed for tracking conversions compared to simpler metrics directly provided by the advertising platform. Nevertheless, it is insightful to investigate whether using other metrics as a proxy for the actual optimization goal is a viable solution. I predict that the performance can further be improved by optimizing directly towards the real optimization goal as there may be some fundamental differences between the optimality depending on which event is seen as the desired one. E.g. the ad set getting most impressions per budget doesn't necessarily get most conversions per budget. This gives us the last hypothesis.

H4: Optimizing towards the real optimization goal instead of using a proxy goal improves the performance of the allocation.

In the following section, I will first describe the data and the methodology to address the above hypotheses. I will then present the result of my study and further elaborate the implications that can be derived from there. The paper will conclude with a discussion of the practical applications of my results and further elaborate the proposed direction for the future research.

5. Methods & data

5.1. Formalizing budget allocation as multi armed bandit problem

The advertiser wants to allocate a predetermined budget across a number of ad sets such that the return on investment is maximized. Each ad set has an unknown reward and cost distributions and the advertiser's problem is to learn the unknown distributions while also being able to use the limited budget as efficiently as possible. The above problem can be translated into MAB problem as follows. The advertiser has a set of ad sets $i \in \{1, 2, \dots, K\}$. Here the ad sets are analogous to the arms of the slot machines in the standard MAB problem. At time slot t the policy is allowed to pull each arm multiple times as well as multiple arms at the same time. The arm pulling at round t is restricted by a budget B.

As many observations are allocated simultaneously at each round t, the problem is so called batched MAB, where the budget serves as the batch size. In batched MAB the arms are sampled in batches at each round and the reward is revealed only after the batch is played. Batched bandits have been investigated especially in the context of clinical trials. A more formal definition was provided recently by Perchet et al. (2016) who considered them in the context of two arm bandits. Jun and Jamieson (2016) extend the batched two arm bandits to k-arm bandits.

At each round the the advertiser sets allocations $w_{i,t}$ of budget for each ad set. Each ad set has an unknown reward distribution with an average reward of μ_i. The reward $R_{i,t}$ from allocating budget to an ad set i can be defined as the number of observed actions $r_{i,t}$ per allocated budget $w_{i,t}B_t$. At each round a random reward is observed based on the underlying distribution. The process is assumed to be stochastic i.e. the underlying reward distribution and μ_i remain constant over time. The goal is to allocate the budget for the ad sets such that the total reward is maximized subject to the budget constraint. Thus the advertiser's optimization problem can be formalized as:

$$\max \sum_{t=1}^{T} \sum_{k=1}^{K} R_{k,t}$$

$$subject\ to \sum_{k=1}^{K} w_k = 1$$

where the reward is defined as $R_{k,t} = \frac{r_{k,t}}{w_{k,t}B_t}$, where $r_{k,t}$ denotes the number of desired actions and $w_{k,t}$ denotes the budget allocation for ad set k at round t, and B_t denotes the total budget at round t.

In order to solve the above allocation problem, I will jointly test several multi armed bandit policies. Section 2.2 already outlined some of the most commonly used pulling policies. In the following I will go through in more detail how to apply them to the batched bandit problem of budget allocation in display advertising.

5.2. Stochastic bandit polices

5.2.1. Greedy approaches

The simplest pulling heuristic to multiple armed bandit problem is the so called greedy approach where the budget is allocated to the arm having the best historical performance. The downside of this policy is that it gives no budget for the exploration and thus is likely to demonstrate relatively poor performance. A variant to this policy is so called ε-first policy where the exploration phase is specifically split from the exploitation phase. In this policy, during the time horizon T the pulled arm is randomly selected at time εT (exploration) and then the best arm is greedily selected at time (ε-1)T (exploitation). While the ε-first incorporates the exploitation, it fails to asymptotically converge the optimal pulling behavior as it may incorrectly choose the suboptimal arm to pull based on the exploitation phase. Also, the policy assumes that the performance of each arm stays constant over the exploration and exploitation phases. This may not hold true in many applications, also not in the context of display advertising.

In order to address the above issues, Watkins (1989) introduced the ε-greedy policy where the policy commits exploration with probability ε by selecting a random arm to pull. The best arm is pulled with the probability of (1-ε). While this approach uses exploration, it's easily seen that it fails to converge the optimal policy as the the exploring becomes unnecessarily when the best arms are learned. Nevertheless, in finite time frames the ε-greedy policy has been shown to perform well (Vermorel and Mohri, 2005). In order to address the issue with asymptotic convergence Auer et al. (2002) proposed a decreasing ε-greedy algorithm that commits to exploration with probability $\min\{1, \varepsilon_t\}$ at time t and otherwise selects the best arm

according to the greedy policy. The ε_t denotes to $\frac{C}{t}$ where C is some positive number and ε_t decreases as the time t increases.

The greedy policies are easily applied to the batched bandit setting of budget allocation by simply using the ε as the proportion of the budget equally allocated to all ad sets and (1-ε) as the proportion of budget allocated to the ad set having the highest average reward μ_i. From the greedy policies I test the simple greedy, ε-greedy and decreasing ε-greedy policies.

5.2.2. Confidence bound estimation strategies

Another approach to the pulling strategies focuses on the theoretical guarantees of the best arm. These so called upper confidence bound (UCB) policies aim to solve the exploration problem by attributing an optimistic reward estimate to each arm and then greedily selecting the one with the best estimate. The reasoning is that unobserved arms will have an over-valued reward estimate and thus will be explored more frequently. The more an arm is pulled the closer it's optimistic estimate converges to the the true reward mean. In addition, no assumptions of the underlying reward distributions are needed.

The first UCB policy introduced by Auer et al. (2002) at the simplest follows a policy of pulling an arm that has the highest index consisting of two terms. The first term being the average reward and the second term being derived from the size of the one sided confidence interval for the average reward. The selected arm maximizes:

$$UCB_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{2 \ln M_t}{m_{i,t}}}$$

where $\hat{\mu}_i$ is the average reward value of an ad set i, M_t is the total spend of the campaign and $m_{i,t}$ is the total spend of the ad set i until the time t. While the policy theoretically converges to the optimal pulling policy, it is shown to to perform poorly in finite time applications (e.g. Auer et al., 2002; Vermorel and Mohri, 2005). Consequently, multiple variations of this policy have been developed in the later literature. One example is so called UCB-tuned algorithm that incorporates variance of the outcome and has been shown to perform better empirically (Auer et al., 2002)

$$UCBtuned_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{\ln t}{m_{i,t}} min\left\{\frac{1}{4}, V_{k,t}\right\}}$$

where $V_{k,t} = \sigma_{i,t}^2 + \sqrt{2(\ln M_t)/m_{i,t}}$ and $\sigma_{i,t}^2$ is the empirical sample variance of the reward of ad set i.

The upper confidence bound policy can be directly applied to the budget allocation problem by allocating the whole budget to the best arm at each round (e.g. Schwartz et al., 2016). However, this naïve reduction to standard MAB algorithm can hardly be seen optimal as it wastes the information from all the non-selected arms at each round.

Jun et al. (2016) propose a novel approach incorporating the confidence bounds and Racing algorithm used in top arm identification. They introduce BatchRacing algorithm, a variant to widely used Racing algorithm first proposed by Maron and Moore (1993). The idea of the BatchRacing algorithm is to take advantage of the confidence intervals in order to determine with high probability which or the arms are or are not the optimal ones. While BatchRacing is shown to be theoretically optimal in identifying top-k arm, it's hardly optimal to the budget allocation problem as it focuses on the exploration of the top arms with a certain confidence instead of actually maximizing the accumulated reward.

Niculescu-Mizil (2009) proposes independently an algorithm similar to the BatchRacing which theoretically converges to the optimal policy and is more suitable for the budget allocation problem. The algorithm maintains a set of surviving arms which is initialized as follows $S_1 = [K]$ At each round, all arms are allocated the same amount of budget and then the upper and lower confidence bounds are computed for each arm similar to the UCB algorithm. Based on the confidence bounds, a set of dominated arms can be identified. An arm i is said to be dominated if there exists another arm j such that the lower confidence bound of j is higher than the upper confidence bound of the arm i. The set of dominated arms can then be safely excluded from the set of surviving arms as with high probability the best arm will not be among them. I adjust the proposed algorithm of Niculescu-Mizil (2009) to be applicable on the budget allocation problem and call it Racing UCB. A more detailed description is shown in Algorithm 1.

Algorithm 1: Racing UCB

Initialize cumulative spend for campaign $M = 0$ and for all ad sets $i \in [K]$ cumulative number of actions $z_i = 0$ and cumulative spend $m_i = 0$

for $t = 1$ to T **do**

Compute $ub_i = \frac{z_{i,}}{m_{i,t}} + \sqrt{\frac{2 \ln M_t}{m_{i,t}}}$ for all $i \in [K]$

Compute $lb_i = \frac{z_i}{m_{i,t}} - \sqrt{\frac{2 \ln M_t}{m_{i,t}}}$ for all $i \in [K]$

$S_t \leftarrow \left\{ i \middle| ub_i > \max_{\forall j \in [K]} lb_j \right\}$

Allocate B_t equally for each arm in S_t

Observe payoffs $r_{i,t}$ for all arms $i \in S_t$

$z_i \leftarrow z_i + r_{i,t}$ for all arms $i \in S_t$

$m_i \leftarrow m_i + \frac{B_t}{S_t}$ for all arms $i \in S_t$

$M \leftarrow M + B_t$

end

5.2.3. Probability matching

Third pulling policy commonly suggested in the literature balances between exploration and exploitation by randomly pulling arms in such way that those arms that are expected to have the higher rewards are pulled with higher probability. Vermorel and Mohri (2005) denote this concept as probability matching. Simpler heuristics such as ε-greedy and decreasing ε-greedy can be considered wasteful as they use simple random sampling for the basis of the exploration. Probability matching tackles this issue by using stratified sampling which under-samples the arms that are likely to be sub-optimal. (Scott, 2010) Two kinds of methods are commonly applied in the literature: Softmax and Softmix.

Luce (1959) was first to introduce so called Softmax policy that is frequently used in the machine learning literature. The policy determines the best arm based on Gibbs distribution. The SoftMax policy chooses an arm i at time t with a probability

$$p_i(t) = \frac{e^{\frac{\widehat{\mu}_{i,t}}{\tau}}}{\sum_{j=1}^{K} e^{\frac{\widehat{\mu}_{j,t}}{\tau}}}$$

where the τ is a tuning parameter that determines the degree of exploration. The choice of the τ‘s value is left to the user. The larger the τ, the more equal the weights between the arms are and the greater is the degree of exploration. On the other hand, as $\tau \rightarrow 0$ the policy converges to simple greedy algorithm. Similarly to the greedy algorithm, the Softmax policy fails to

theoretically converge towards the optimal policy as the tuning parameter is a constant and thus the level of exploration also remains constant over time.

In order to address the above issue Cesa-Bianchi and Fischer (1998) suggest a Softmix policy in which the value of τ decreases over time. They introduce a temperature decreasing with a factor $\frac{\ln t}{t}$. Another common variant is to decrease the temperature similar to decreasing greedy policy with a factor $\frac{1}{t}$ (Tran-thanh, 2012).

Softmax and softmix are in theory directly applicable to the batched budget allocation problem. However, due to the limits of computer memory we can't in practice decrease the tuning parameter for softmix indefinitely as the exponent term becomes infinite and the allocation proportion non numeric. In order to decrease these cases I use the factor $\frac{\ln t}{t}$ which decreases less aggressively. Once the allocation hits the computational limit the algorithm switches to simple greedy algorithm as an approximation to situation where tuning parameter is converging to zero.

5.2.4. Randomized probability matching (Thompson sampling)

The Bayesian ideas for solving multi armed bandit problems date back to over 80 years. Thompson (1933), was first to introduce an algorithm based on posterior sampling. Here the idea is to start with fictitious prior distributions of the rewards which get updated to more and more accurate posterior distributions as real data from the rewards is gathered. Updating the posterior distribution continuously adds information about the true unknown reward parameter. The posterior distributions can then be used to calculate the likelihood of an arm being optimal by sampling from the posterior of each arm and then choosing the arm proportionally to the times of it being optimal. This procedure makes sure that the arms more likely to be optimal are chosen more often. The literature has shown that randomized probability matching is easy to apply in general settings and tends to balance well in exploration and exploitation by allocating observations efficiently (e.g. Chapelle and Li, 2011; Gopalan et al., 2014; Scott, 2010). Randomized probability matching is also compatible with batch updates of the posterior distribution so it's easily applied to the budget allocation problem.

Algorithm 2: Thompson Sampling

Initialize cumulative number of actions $k_i = 0$ and cumulative spend $\theta_i = 0$ for all ad sets $i \in [K]$. Total number of days $T = 100$ and total number of times sampled $N = 100$
for $t = 1$ to T **do**
  Initialize best arm counts $s_i = 0$ for all $i \in [K]$
  for $n = 1$ to N **do**
    For each arm $i = 1,2, \dots K$ **do**
      sample $\hat{r}_{i,n}$ from the $\Gamma(k_i, \theta_i)$ distribution
    end
    Select best arm $i^* := \arg\max \hat{r}_{i,n}$
    Set best arm count $s_{i^*} = s_{i^*} + 1$
  end
  For each arm $i = 1,2, \dots K$ **do**
    $w_{i,t} = \frac{s_i}{N}$
    Allocate budget for each arm $b_{i,t} = w_{i,t} B_t$ and observe reward $r_{i,t}$
    $k_i = k_i + r_{i,t}$
    $\theta_i = \theta_i + b_{i,t}$
  end
end

A starting prior distribution for the rewards is needed for the algorithm. In general, the choice could as its simplest be a uniform distribution, but choosing a distribution closer to the assumed real distribution can make the model converge quicker to the actual underlying distribution. Following to Tkachenko (2014) I assume that the conversions follow a Poisson distribution. Poisson is well suited for this purpose as it expresses the probability of given number of events occurring in a fixed interval of space, in this case the probability of actions per money spent. A vast majority of the literature has investigated the simplest case of multiple armed bandits where the problem is Bernoulli distributed (i.e. the reward is binomial). However as Agrawal and Goyal (2013) show, the Thompson sampling can be generalized to be used with any kind of prior distribution. Following the Poisson distributed reward assumption I choose the gamma distribution as the priori for the rewards.[1]

The Thompson sampling for budget allocation is described in the Algorithm 2. For each ad set I assume that priori the reward is gamma distributed $\mu_{i,t} \sim \Gamma(k, \theta)$. Initially I choose a vague priori with shape parameter $k = 0$ and scale parameter $\theta = 0$. At each round t the

[1] In Bayesian inference (where the probability of a hypothesis is updated based on the accumulated evidence) Gamma distribution is the conjugate prior to Poisson distribution. Conjugate prior is a distribution that has the same algebraic form as the posterior and can be used for an algebraic convenience in order to avoid numerical iteration. (Fink, 1997)

outcome is observed and the priori of the reward is updated to a posteriori as $\Gamma(k + r_{i,t}, \theta + b_{i,t})$, where $r_{i,t}$ denotes to the observed actions and $b_{i,t}$ to the allocated budget for the ad set i. After this the algorithm samples expected rewards for each ad set from the posteriors. For each sampling round the algorithm observes which of the ad sets performs best. Counting together the times an ad set is optimal and dividing the sum by the total number of sampling gives the probability of an ad set being optimal based on the knowledge accumulated till the time. This probability is then used by the algorithm to allocate the budget for the next round.

5.3. Data

The empirical experiments are run against real world conversion data obtained from a Facebook marketing partner. The original dataset contains ad sets of campaigns that have been active between March 2016 and December 2016. As the purpose is to focus on the algorithm performance based on historical data I limit the ad sets to those that have at least 30 days of data during the observation period. For the campaigns I require at least 100 days. As in the real world, the ad sets may be added to the campaign after it's start, thus the requirements for the consecutive days for ad sets and campaigns differ. For each day of a campaign I require more than two ad sets for which to allocate the budget. Due to the limitations in the computational power I also disregard days for campaigns having more than 100 ad sets.

As the audience of the ads is limited, each ad set has a spending capability that is not easily foreseen beforehand. A theoretical spend limit can be calculated from the ad set's realized cost per impression and estimated reach by $CPM_i \times E(reach_i)$. However, when ads are bought from the spot market at run rate, it is not guaranteed that the budget is spent up to this theoretical limit. This is because in general impressions to persons who are more likely to react positively to an ad are more expensive than impressions to persons who tend to ignore the ads. Consequently, the price per impression tends to grow the more of the audience the advertiser actually wants to reach. If the advertiser doesn't increase the bid of the ad set there is no guarantee that the ad set is able to spend up to its theoretical limit. Also, the estimated reach in general includes users who are not online on a daily basis. As a result, making any assumptions about the ad sets spending capability above its real spend can be seen dubious.

In order to solve the above issue, I will limit the dataset to days when ad sets have been able to spend at least 100 units of budget. The 100 units are then used as the total budget for the campaign. This way I confirm that each ad set would have been able to spend the allocated budget with the realized conversion rate. For real world applications the spend capability could

naturally be taken into account by observing the point where the ad set is unable to spend its budget and not allocating budget above that point. For the simulation purposes, the total budget choice is arbitrary as the performance is measured as regret against the optimal policy. Thus downscaling the simulation budget will not affect the results.

As stated above, some of the impressions may be more valuable in terms of reward potential than others. In spot markets the impressions are sold based on the bids that advertisers set for the ad sets. In general, this means that the budget of an ad set should not impact on its conversion rate if the bid remains constant. As the scope of this study is to investigate the impact of budgeting alone, I eliminate the impact of bid changes by treating an ad set after a bid change as a new ad set. When the bid of an ad set is held constant it is reasonable to assume that the underlying reward distribution also remains constant.

After the above mentioned adjustments and removing possibly duplicate rows I am left with 267 campaigns having altogether 3 211 ad sets with a total of 179 460 days of returns.

5.4. Distributional simulation and sequential experiment

Two sets of simulations are run for the empirical data set. In the first, I construct a distribution based on the empirical data set to be used in the simulation. In the second, the historical data set is directly used in sequential experiment to test the allocation methods.

For the distributional simulation, I use the historical data of actions per unit of budget $\{R_{i,t}\}$ to construct an empirical distribution p_i for each ad set i. For each round, the actions per unit of budget are drawn from these distributions and the reward is simulated based on that. (e.g. Chapelle and Li, 2011)

For the sequential experiment I follow e.g. Amin et al., 2012 and use the sequential data directly instead of generating a simulated data. For each ad set, we can compute the exact optimal allocation by knowing the sequence $\{R_{i,t}\}$ a priori. The performance of the model can then be benchmarked against this optimal allocation.

The algorithms to be tested are equal allocation (denoted as *equal*), greedy policies, probability matching policies, upper confidence policies and Thompson sampling. For greedy policies I run the following algorithms: simple greedy denoted as *greedy* allocating all of the budget to the ad set having best average return, ε-greedy algorithms with epsilon of 0.1 and 0.5 denoted as *ε-greedy 0.1* and *ε-greedy 0.5*, and decreasing ε-greedy algorithms with constants 1 and 10 denoted as *decreasing ε-greedy 1* and *decreasing ε-greedy 10*. For probability matching

policies I run softmax and softmix algorithms with tuning parameters 1 and 5. These are denoted as *softmax 1*, *softmax 5*, *softmix 1* and *softmix 5* respectively. For upper confidence interval policies I use the racing UCB algorithms denoted as *ucb* and *ucb-tuned* depending on the method used to calculate the confidence intervals. The Thompson sampling policy is denoted as *Thompson.*

The simulations are run by calculating allocations day by day for consecutive 100 days. Each day each campaign has 100 units of budget to allocate to its ad sets. At each time t the allocation algorithms are given the historical information of the number of actions for each ad set $\{r_{i,1} \dots r_{i,t-1}\}$. If the algorithm does not allocate any budget to an ad set i at time t, the number of actions $r_{i,t}$ will not be available for the algorithm on any following time periods. As in reality it is not possible to allocate an infinitively small amount of a budget for an ad set and still observe the rate of return for the allocation, I limit the minimum amount of allocable budget to 1. If the algorithm was to allocate less than this, the allocation is rounded to zero and the underlying return for the time remains unknown for the algorithm. When no history is available, the budget is allocated equally. The same goes with new ad sets in the campaigns, their initial allocation equals their share of the budget as if it was equally allocated.

Following to established practice in the multi armed bandit literature (e.g. Scott, 2010; Vermorel and Mohri, 2005) the performance of the the allocation algorithm can be measured as cumulative regret. The regret ρ after T rounds can be defined as the cumulative expected lost reward relative to always allocating to the optimal ad set from the beginning of the experiment. For each campaign, the regret at time t can be computed as the following

$$\rho_t = \sum_{i=1}^{K} b_{i,t}(R_t^* - R_{i,t})$$

where $b_{i,t}$ is the allocated budget and and $R_{i,t}$ is the reward obtained from an ad set i at time t. The reward of the optimal ad set at time t is denoted by $R_t^* = \max(R_{i,t})$. From the above we get the cumulative regret at time T as

$$\mathrm{P} = \sum_{t=1}^{T} \rho_t$$

As the absolute regret depends on the rewards of the ad sets, comparing them directly would cause overweighting the algorithm performance of campaigns with high rate of return while downplaying the importance of the campaigns with lower rate of returns. As we are merely

interested on how the algorithms perform compared to the optimal allocation strategy, I compute a relative regret for each campaign. The relative regret at time T is defined as

$$\frac{\text{P}}{\sum_{t=1}^{T} R_t^*}$$

where R_t^* is the rewards of a campaign at time t if the budget was allocated to the optimal ad set.

In order to investigate the effect of different optimization goals, I run separate simulations using the number of impressions, clicks and conversions per units of budget as rewards. Each of these simulations then optimizes towards one of the goals respectively. For the sequential simulations the computational time is around 20 minutes. For the distributional simulations I run 100 scenarios which takes about 34 hours.

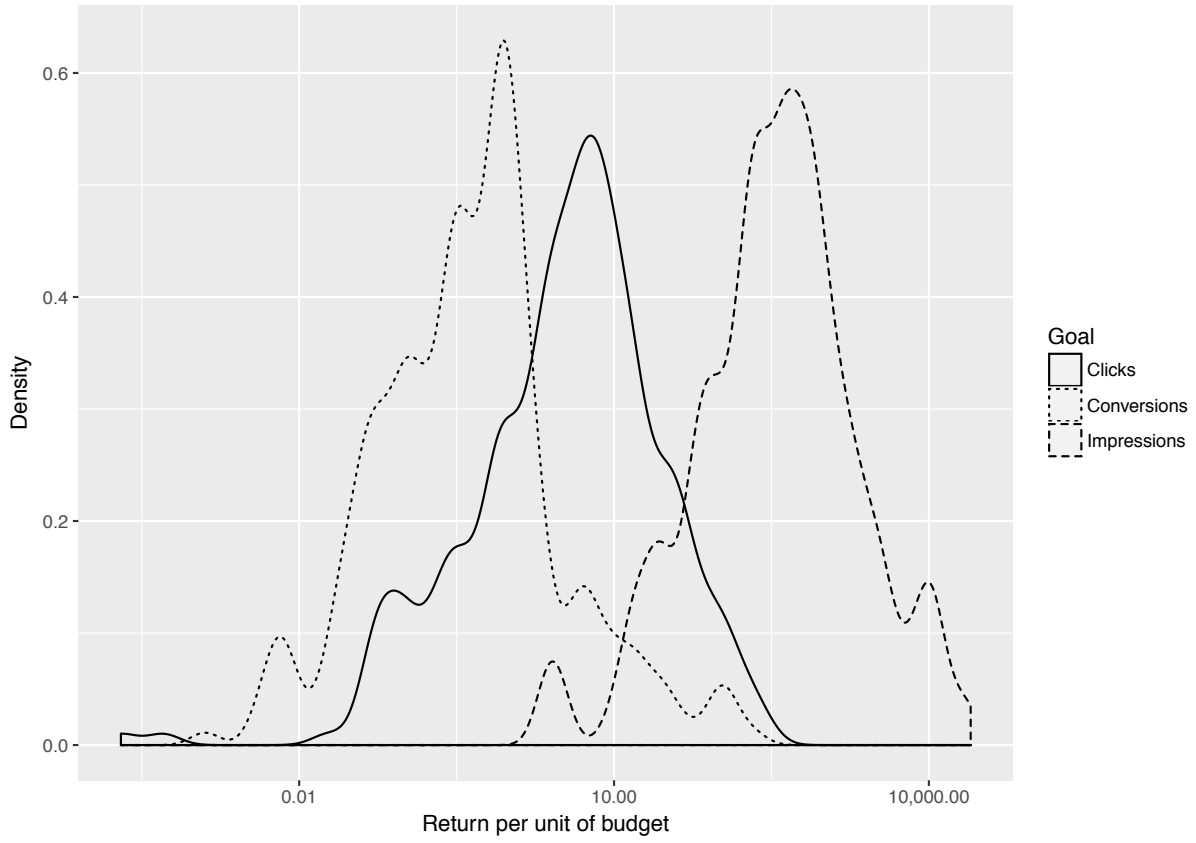
6. Results

The baseline for the analysis is set by the reward for the optimal allocation which shows the theoretical maximum reward obtainable by the campaigns. This is calculated by allocating the whole daily budget to the best performing ad set while knowing beforehand which of the ad sets will perform best at that day. The summary statistics of maximum daily reward for each optimization goal are show in Table 1. We can see that the differences in the average rewards are notable for different optimization goals. This is natural as the actions for conversions and clicks may occur only after the ad has gotten an impression. Conversions, on the other hand, are more unlikely to occur than clicks. This is easy to see as e.g. making a purchase requires a lot more consideration and commitment than simply clicking an advertisement.

Table 1: Summary statistics of daily return per unit of budget with optimal allocation

Goal	Mean	SD
Impressions	1437.891	3262.556
Clicks	15.827	34.317
Conversions	4.652	19.525

Figure 3: Distributions of average returns with optimal allocation



The x axis is log scaled due to the skewness of the observations

Figure 3 gives more insight of the distribution of rewards with different optimization goals. The rewards for the campaigns are in general heavily skewed to the right so the figure is using log scaled x axis. We see that apart from the number of actions per budget the differences between optimization goals are small. Clicks seem to have somewhat higher variance but in general the distributions resemble each other. Based on this we can anticipate that the choice of optimization goal should not have a notable impact on the budget allocation performance unless the amount of actions has a notable impact on the predictability.

The results for distributional simulations are presented in

Table 2 and Figure 4. The base line for the performance is set by the equal policy which simply allocates budget equally across all of the ad sets. In line with the Hypothesis 1, we see that dynamically allocating budget appears to improve performance as the other allocation policies outperform the equal allocation quite consistently.

Unlike Hypothesis 2 predicted, the naïve approaches are doing notable well. In fact, the greedy algorithms are among the top performers regardless of the optimization goal. This indicates that the average return seems to predict quite well the forthcoming returns. We can also see that for the greedy approaches more aggressive policies outperform the more cautious policies. Decreasing ε-greedy policies outperform constant ε-greedy policies and the smaller the epsilon the better the algorithm does. However, results also indicate that reserving some budget for exploration seems to pay out as the simple greedy algorithm is outperformed by ε-greedy 0.1 and decreasing ε-greedy 1 policies. This was anticipated as the greedy algorithm won't be able to observe the returns of those ad sets it doesn't allocate budget for and thus receives only part of the information received by the less aggressive policies. Nevertheless, it seems that in the context of display advertising the very first returns of the starting day are good enough estimates for the greedy algorithm to predict the future performance notably well.

The probability matching policies follow the same trend as greedy policies. Using smaller tuning parameter results to smaller regrets as well as decreasing the tuning parameter by time. While optimizing towards impressions the algorithms clearly beat the equal allocation, moving the optimization goal to clicks and conversions makes them to converge the reward of equal allocation. By looking at the equation of softmax algorithm we can see that this is most likely caused by smaller absolute differences in the return rates of clicks and conversions which is due to the fact that the average return rates are smaller for those optimization goals. This on

the other hand could be offset by choosing a smaller, more appropriate tuning parameter for those optimization goals. This brings us a practical problem with the probability matching and tuning algorithm. The performance of the algorithms seems to be heavily dependent on the choice of tuning parameter which on the other hand depends on the beforehand unknown rate of return. Thus the real world application of the model appears cumbersome.

Table 2: Average relative regret of allocation policies in distributional simulation

	Impressions		Clicks		Conversions	
Algorithm	Mean	SD	Mean	SD	Mean	SD
Equal	0.359	0.172	0.387	0.172	0.526	0.205
Greedy	0.119	0.097	0.211	0.139	0.422	0.214
Epsilon-greedy 0.1	0.123	0.085	0.205	0.130	0.395	0.206
Epsilon-greedy 0.5	0.228	0.109	0.287	0.133	0.452	0.197
Decreasing epsilon-greedy 1	0.109	0.089	0.196	0.135	0.390	0.214
Decreasing epsilon-greedy 10	0.177	0.091	0.247	0.130	0.422	0.197
Softmax 1	0.196	0.139	0.381	0.170	0.525	0.205
Softmax 5	0.276	0.153	0.386	0.172	0.526	0.205
Softmix 1	0.147	0.145	0.341	0.169	0.518	0.210
Softmix 5	0.165	0.125	0.372	0.168	0.524	0.206
UCB	0.149	0.127	0.372	0.174	0.525	0.207
UCB-tuned	0.121	0.099	0.333	0.186	0.516	0.215
Thompson	0.109	0.092	0.243	0.144	0.441	0.212

Relative regret refers to the ratio of lost returns to the theoretical maximum returns. The smaller the relative regret, the closer the algorithm is to the return of the theoretical optimal allocation.

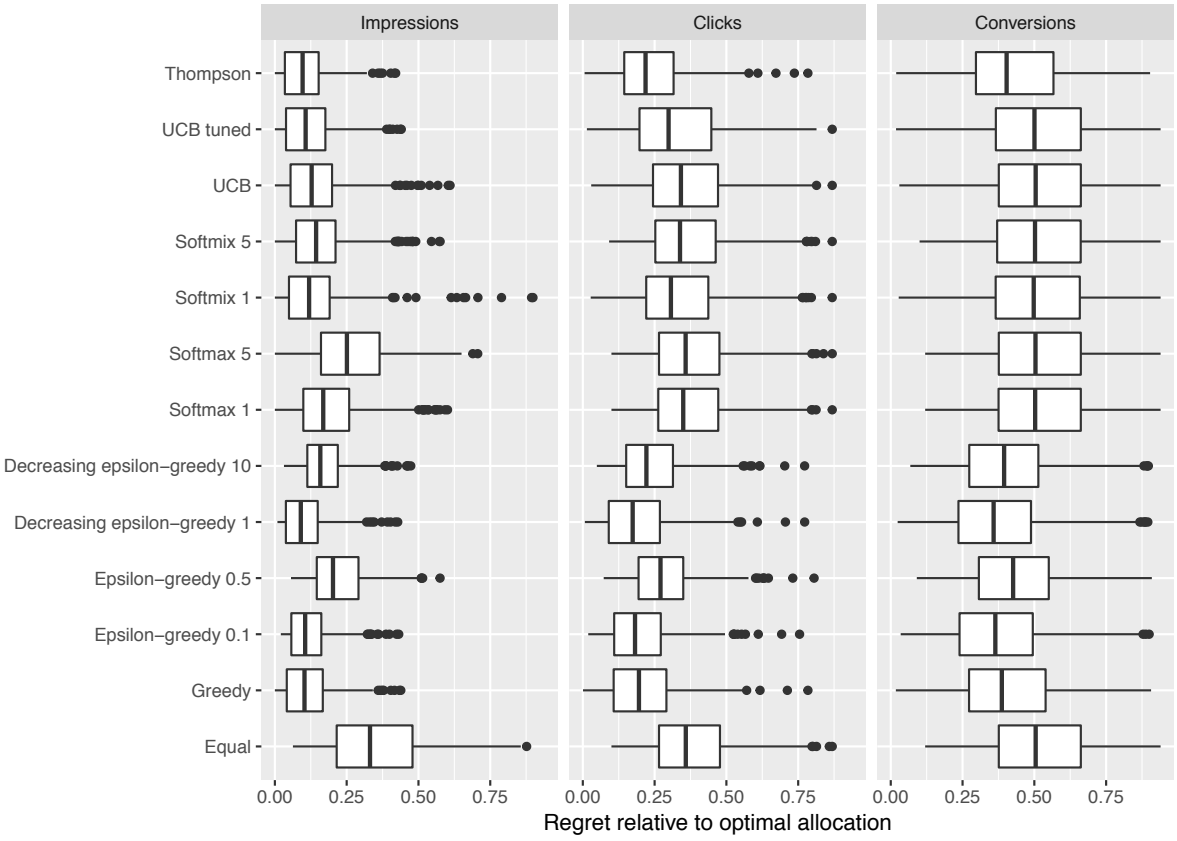
Table 3: Average relative regret of allocation policies in sequential experiment

	Impressions		Clicks		Conversions	
Algorithm	Mean	SD	Mean	SD	Mean	SD
Equal	0.332	0.177	0.349	0.177	0.497	0.219
Greedy	0.090	0.110	0.174	0.159	0.398	0.258
Epsilon-greedy 0.1	0.083	0.067	0.158	0.136	0.372	0.237
Epsilon-greedy 0.5	0.195	0.103	0.243	0.135	0.427	0.220
Decreasing epsilon-greedy 1	0.072	0.077	0.152	0.142	0.369	0.243
Decreasing epsilon-greedy 10	0.142	0.082	0.204	0.135	0.398	0.226
Softmax 1	0.161	0.134	0.342	0.174	0.496	0.220
Softmax 5	0.242	0.152	0.348	0.176	0.497	0.219
Softmix 1	0.107	0.137	0.298	0.174	0.489	0.226
Softmix 5	0.128	0.115	0.332	0.172	0.495	0.221
UCB	0.118	0.124	0.333	0.177	0.495	0.222
UCB-tuned	0.090	0.106	0.293	0.191	0.487	0.230
Thompson	0.076	0.091	0.196	0.156	0.412	0.245

Relative regret refers to the ratio of lost returns to the theoretical maximum returns. The smaller the relative regret, the closer the algorithm is to the return of the theoretical optimal allocation.

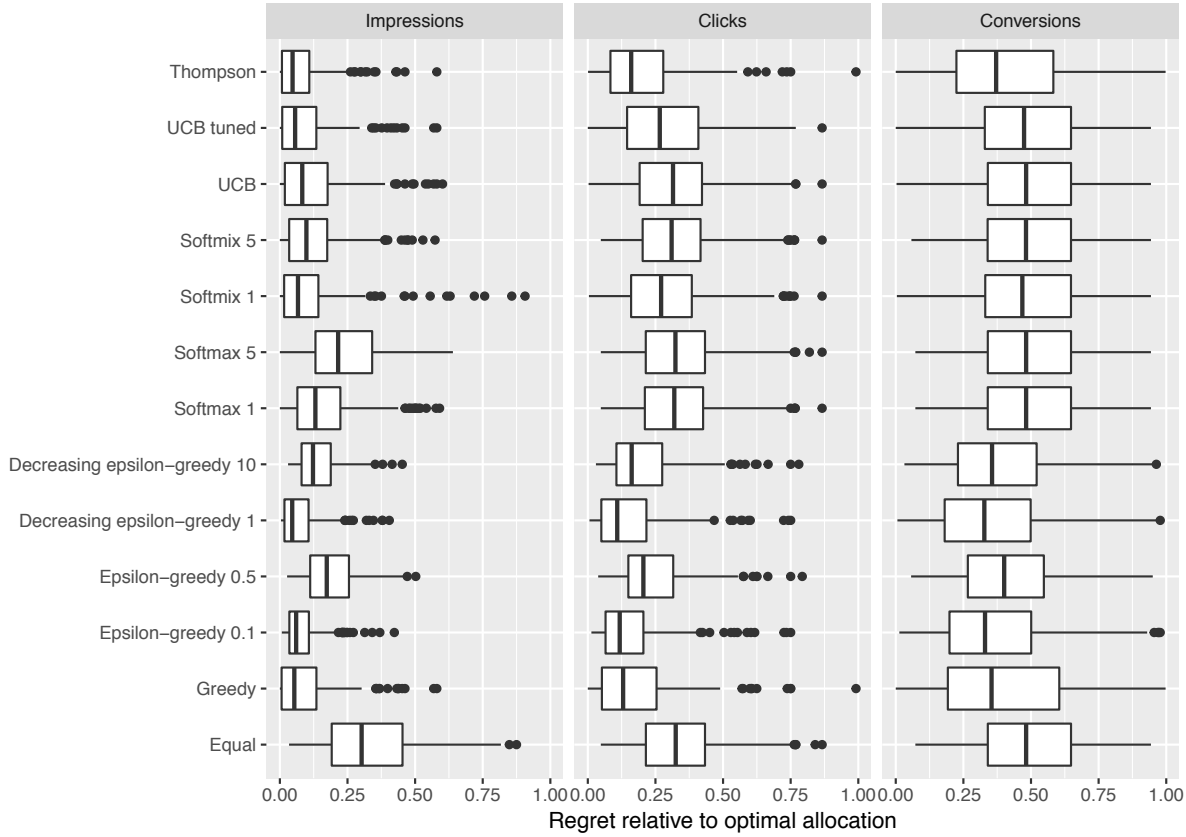
The UCB algorithms seem to suffer from the same issue as the probability matching. With conversions as optimization goal they practically retain the equal allocation despite of having a slightly longer tail to smaller regret than the probability matching policies. Again the problem seems to be inbuilt with the algorithm. When the average reward is small the confidence intervals become relative big and algorithms fail to reject any of the ad sets as poor performers. With impressions as optimization goal, the average reward is much higher and the algorithms perform notable better. In line with the previous literature, incorporating the reward variance to the confidence intervals seems to yield better results as UCB-tuned outperforms the plain UCB algorithm.

Figure 4: Distributional simulation: Relative regrets of algorithms by optimization goal



Regret refers to the lost return opportunity compared to the optimal allocation. The optimal allocation is the theoretical maximum return obtained had the advertiser known the return of the ad sets beforehand.

Figure 5: Sequential experiment: Relative regrets of algorithms by optimization goal



Regret refers to the lost return opportunity compared to the optimal allocation. The optimal allocation is the theoretical maximum return obtained had the advertiser known the return of the ad sets beforehand.

From the more sophisticated algorithms, only Thompson sampling seems to be able to challenge the performance of greedy algorithms. We can also see that it retains its performance regardless of the optimization goal. This makes the algorithm to appear as a lucrative option to the greedy ones. From the theoretical point of view, it should not inherently carry as much risk for under exploring as the greedy algorithms do. Nevertheless, the empirical evidence shows that Thompson sampling is also able to compete the greedy algorithms when an aggressive allocation pays off.

Despite of good performance of Thompson sampling, it doesn't notably outperform the greedy algorithms. As neither do probability matching strategies nor UCB strategies we will reject Hypothesis 2. Based on the evidence, it appears that already simple greedy algorithms can give the advertiser the same benefits as models having theoretically better guarantees. On the contrary, dynamically adjusting the exploration and exploitation weights does seem to make a difference. In general, all of the algorithms give indication that converging towards a pure exploitation policy yields higher returns than keeping the exploration vs exploitation allocation

constant. For both greedy and probability matching policies, increasing the exploitation proportion as time passes improves the performance. This leads us to retain the Hypothesis 3.

The results for the sequential experiment are presented in the Table 3 and Figure 5. We see that these are well in line with the results obtained through distributional simulation and the above analysis appears to apply also for the sequential experiment. As the consistency between the two experiments demonstrates the robustness of the sequential experiment I base the following more detailed analysis on the sequential experiment. This allows us to do more accurate analysis with data that correctly describes the relations between different points of time as well as different optimization goals.

As the Figure 4 and Figure 5 demonstrate, the predictability of the returns gets worse when moving the optimization goal from impressions to clicks and clicks to conversions. A highly likely explanation is that the amount of observed actions correlates with the performance of the algorithm. The less observations of the underlying reward distribution the model has, the less educated guesses of the future performance it can make. This is demonstrated in the

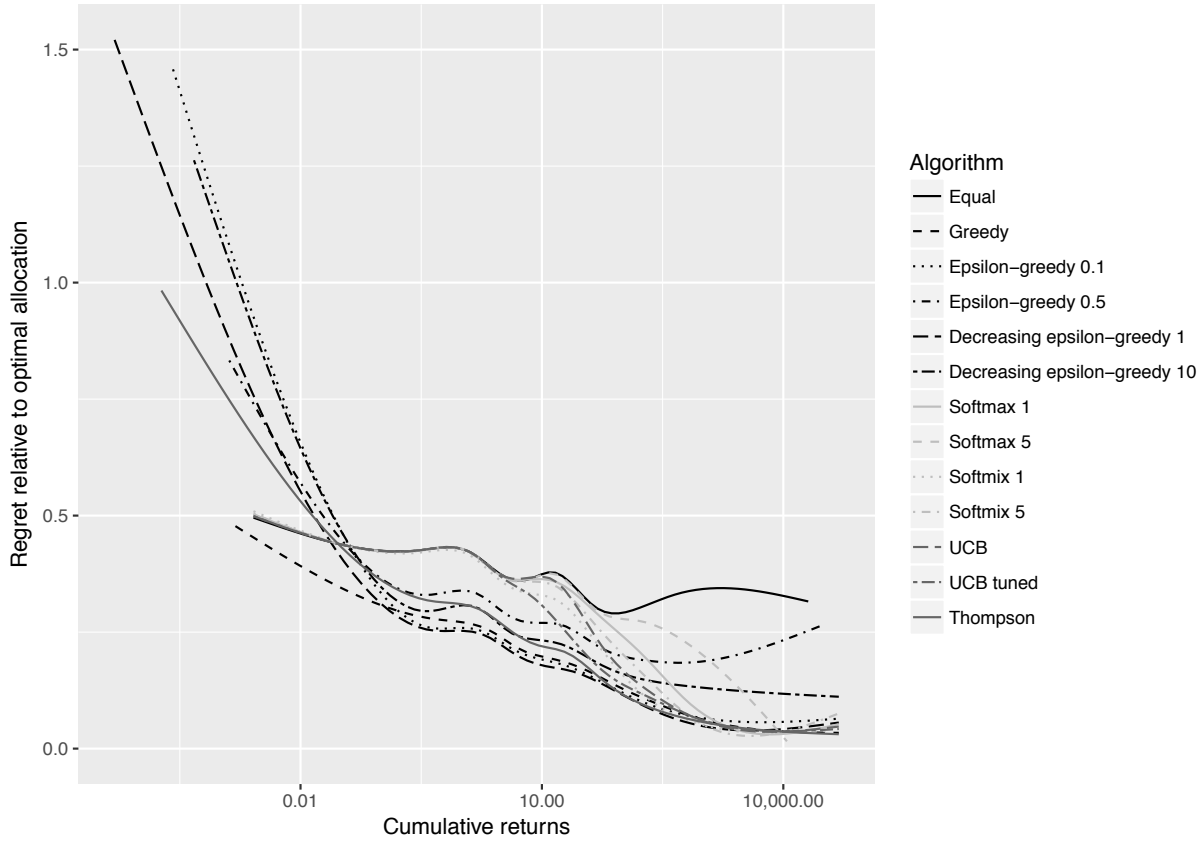
Figure 6 which shows smoothed average regrets against cumulative reward.

In general, the algorithms perform better the higher the cumulative reward is i.e. the more observations of the actions they are having. The greedy algorithms as well as Thompson sampling have the steepest curves and appear to be most robust to the scarcity of the observations. Although with very low number of observations their regret is high, they start to outperform the equal allocation quite quickly as the number of observations grows. Probability matching and UCB algorithms on the other hand are shown to practically follow the equal allocation until the cumulative reward reaches a certain threshold. This in line with my previous analysis and implies that while the algorithms are able to outperform equal allocation they require enough observations to do so. This can be an issue with many real life applications.

As the scarce number of accumulated actions is an issue for the predictability it doesn't seem viable to optimize towards an event that has too low reward. In many cases, however, from the advertisers' point of view, the conversions e.g. purchase events are the very purpose of the advertising and thus the true goal they'd like to optimize towards. One widely used workaround is to approximate the conversions with some event that is earlier in the conversion funnel. E.g. if most of the people making a purchase have also clicked the advertisement the clicks could be used as a proxy for conversions and optimizing towards clicks would simultaneously optimize towards conversions while bypassing the problem of too few actions

in conversions. My dataset enables testing for this theory in practice by applying the reward calculated based on conversions to the allocations calculated when optimizing towards impressions or clicks. The reward obtained by using a proxy goal can then be benchmarked against the reward obtained when optimizing directly towards the real goal.

Figure 6: Average regret versus cumulative reward

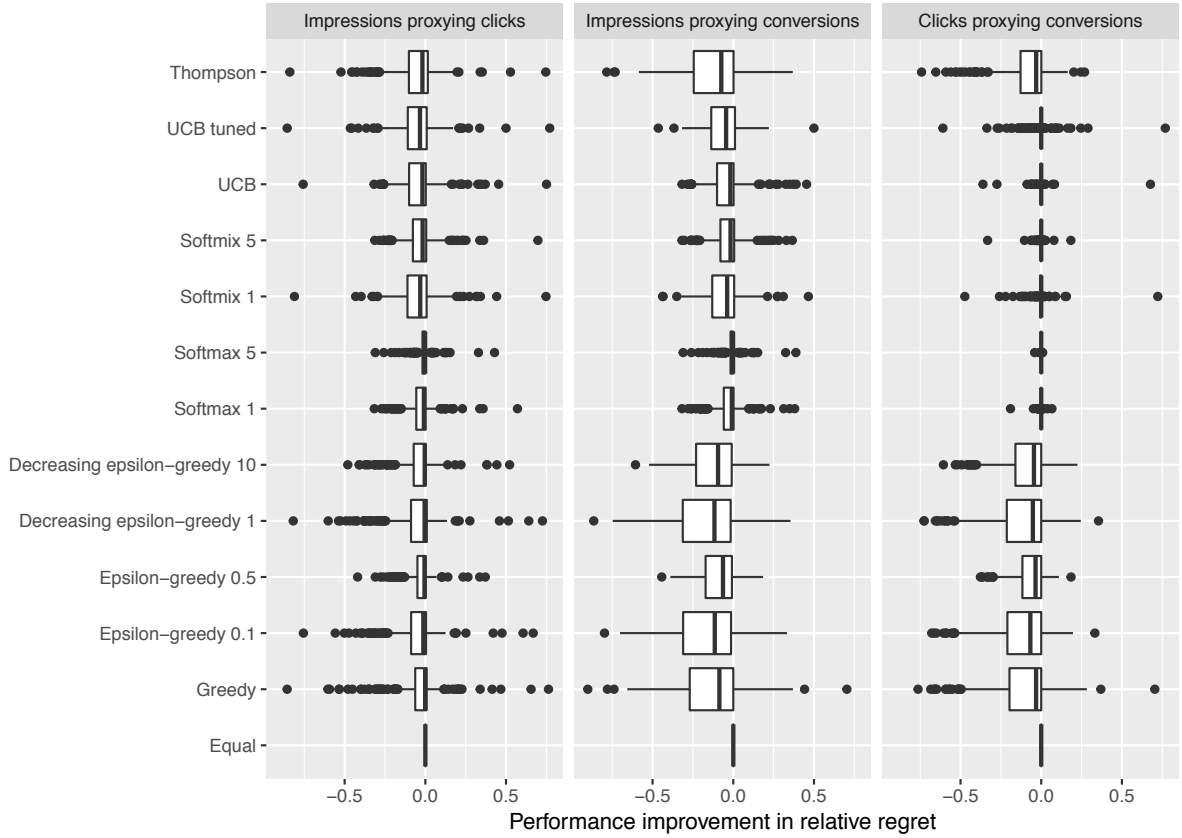


Relative regret refers to the ratio of lost returns to the theoretical maximum returns. The smaller the relative regret, the closer the algorithm is to the return of the theoretical optimal allocation. The x axis is log scaled due to the skewness of the observations

The results are presented in Figure 7. We can see that using impressions as a proxy optimization goal does not improve the performance when maximizing clicks. When trying to maximize conversions by optimizing towards impressions or clicks the performance gets reduced even more notably. Especially for those algorithms that performed best when optimizing towards conversions, the performance is lost when changing the optimization goal upwards in the funnel. The results imply that even if small amount of actions seems to be an issue for the allocation algorithms, it's still worthwhile to optimize towards the real goal than

change the optimization goal to an event with more returns. Consequently, we can retain Hypothesis 4.

Figure 7: Impact of alternative optimization goals to the performance of algorithms



Each panel describes the distributions of changes in performance when changing the optimization goal to a proxy goal while still measuring the performance with the original goal. E.g. the first panel shows the performance changes for the campaigns when measuring the performance with clicks but using impressions instead of clicks as the optimization goal for the algorithms.

This has also implications for practitioners doing budget allocation optimization. First of all, using the real performance metric for the advertising as the optimization goal for budget allocation pays off even with small number of actions. Secondly, measuring the revenue generating event directly instead of approximating it with events earlier in the conversion funnel increases the advertising performance. Thus implementing a proper tracking for the conversion funnel is likely to pay out even if the conversion measurement required a little more work than impression or click measurement. From the theoretical point of view, we can see a clear need for developing models that would be more robust to also smaller rewards.

7. Conclusions

The unique empirical research done in this paper has contributions to two lines of academic literature. Firstly, I have presented both theoretical and empirical extensions to the machine learning research done around multi armed bandit models. Majority of the literature on MAB problems has focused on sequential pulling policies where the rewards from actions can be observed immediately and the arm pulling is limited by the number of sequential pulls. This model however isn't directly applicable to many real world resource allocation problems where the resources are allocated in batches instead of one by one. Perchet et al. (2016) formalize a batched bandit problem but practical applications have still been missing from the literature.

I have brought together a scattered literature of different multi armed bandit problems and formalized several alternative pulling policies for the budgeted batched bandit problem with unknown cost and reward ratios. Schwartz et al. (2017) already introduced a budgeted batched bandit applications for greedy algorithms and Thompson sampling. However, they reduced the UCB algorithm to simple MAB problem. I add on this by presenting a batched application of UCB policy based on the works by Niculescu-Mizil (2009) and Jun et al. (2016). Based on the theoretical work by Tran-thanh (2012), I expand the testing of greedy algorithms to include decreasing greedy algorithms. In addition, inspired by Tkachenko (2014) I suggest applications for both softmix and softmax policies for the budgeted batched bandit problem.

In addition to adding practical applications for the pulling policies, I have tested the policies in an empirical setting based on extensive real world dataset. This gives valuable insight on the performance of the current models in practice. In line with the previous literature I found that in general, even the simplest MAB policies are able to outperform an equal allocation. This is in line with the findings by e.g. Sahin Cem Geyik and Dasdan (2014), Schwartz et al. (2017) and Tran-thanh (2012). Also in line with the previous empirical research by Schwartz et al. (2017) I found supporting evidence that simple greedy algorithms are able to outperform many more sophisticated models. This implies that, in real world resource allocation settings, underexploring may not be as big of an issue as suggested in the theoretical literature. In line with the theory presented by Tran-thanh (2012) my evidence shows that the decreasing ε-greedy algorithms are outperforming the simple ε-greedy policies. This shows that exploring in the early phase an the converging to simple greedy strategy applies well to display advertising setting.

The empirical research related to probability matching methods is rather scarce. Tkachenko, (2014) got promising results by numerical simulations which were not based on empirical data. However, my results demonstrate poor applicability of probability matching methods in practical settings. Namely, the choice of temperature term proved to be difficult for the optimality of the model. While too high temperatures resulted in computational infeasibility, too low temperatures ended up mimicking equal allocations. In order to be able to use these models in practical resource allocation problems they would need some sort of extension that would e.g. automatically be able to adjust the temperature term based on the average reward.

Last notable finding from the batched multi armed bandit models was their drop in performance with small rewards. While this may not be a common issue in traditional multi armed bandit setting, in the context of budgeted batched bandits it's likely to emerge quite often. This is because the reward is defined as rate of return i.e. a number of desired actions per units of budget and not all settings have high enough rate of return. At least in the context of online advertising, having more robust models for lower rates of returns could be highly beneficial. I imagine this same issue may arise in other industries as well if MAB policies were used to optimize resource allocation.

In addition to the contributions to the machine learning literature, my empirical study has also clear implications to the marketing research as well as practitioners namely at the field of online advertising. In line with the previous literature, I give more evidence that dynamically optimizing advertising budgets can substantially improve the advertiser's return on investment. In my empirical simulations, optimizing budget allocation was able to improve the average performance of campaigns by almost 30 %. Notable performance improvement was achieved already by greedily allocating majority of the budget to the ad set having the best historical average rate of return. Another lucrative allocation approach seemed to be Thompson sampling, which was able to compete with the greedy algorithms and at least in theory may be able to better avoid myopic behavior of giving too much weight on isolated observations. I also found evidence that optimizing towards the real desired advertising goal yields better results in terms of return on investment than using some earlier point in the conversion funnel as a proxy for the real goal. The findings are well in line with those of Schwartz et al. (2017) and give indication that their findings apply also to social media display advertising.

There are some limitations to this study that could potentially be taken account by the following research. Firstly, my simulations and the used pulling policies assume that the reward distributions are stochastic and remain constant over time. This may not always hold true in

real world applications and thus it would be interesting to investigate the effect of changing distributions to the performance of the models. Also, the rewards of this study are attributed to the ad sets with last touch attribution due to the nature of Facebook tracking. However, as the display advertising tools get more advanced, using multi touch attribution becomes more available for both the practitioners and the scientific community. The choice of the attribution model has a great potential to reveal totally new aspects of the optimality between the ad sets and thus also change the game with the budget allocation policies.

8. References

Agrawal, S., Goyal, N., 2013. Further optimal regret bounds for thompson sampling, in: Proceedings of the 16th International Conference on Artificial Intelligence and Statistics. pp. 99–107.

Aksakallı, V., 2012. Optimizing direct response in Internet display advertising. Electron. Commer. Res. Appl. 11, 229–240. doi:10.1016/j.elerap.2011.11.002

Amin, K., Kearns, M., Key, P., Schwaighofer, A., 2012. Budget Optimization for Sponsored Search: Censored Learning in MDPs, in: Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI-12). Catalina, CA, pp. 543–553.

Araman, V.F., Popescu, I., 2005. Stochastic Revenue Management Models for Media Broadcasting, Business Week.

Archak, N., Street, W., York, N., Mirrokni, V.S., 2010. Budget Optimization for Online Advertising Campaigns with Carryover Effects Categories and Subject Descriptors. Sixth Ad Auction. Work.

Auer, P., Cesa-Bianchi, N., Fischer, P., 2002. Finite-time analysis of the multiarmed bandit problem. Mach. Learn. 47, 235–256. doi:10.1023/A:1013689704352

Badanidiyuru, A., 2013. Bandits with knapsacks. Act. Learn. 1–51. doi:10.1109/FOCS.2013.30

Balseiro, S., Feldman, J., Mirrokni, V., Muthukrishnan, S., 2014. Yield optimization of display advertising with ad exchange. Manage. Sci. 60, 2886–2907. doi:10.1287/mnsc.2014.2017

Bruner, R.E., Gluck, M., 2006. Best Practices for Optimizing Web Advertising Effectiveness.

Cesa-Bianchi, N., Fischer, P., 1998. Finite-time regret bounds for the multiarmed bandit problem, in: Proceedings of the Fifteenth International Conference on Machine Learning. pp. 100–108.

Chapelle, O., Li, L., 2011. An Empirical Evaluation of Thompson Sampling. Adv. Neural Inf. Process. Syst. 2249–2257.

Chapelle, O., Manavoglu, E., Rosales, R., 2013. Simple and scalable response prediction for display advertising, people.csail.mit.edu. doi:10.1145/0000000.0000000

Danaher, P.J., 2007. Modeling Page Views Across Multiple Websites with an Application to Internet Reach and Frequency Prediction. Mark. Sci. 26, 422–437. doi:10.1287/mksc.1060.0226

Danaher, P.J., Lee, J., Kerbache, L., 2010. Optimal Internet Media Selection. Mark. Sci. 29, 336–347. doi:10.1287/mksc.1090.0507

Deshpande, N., Ahmed, S., Khode, A., 2014. Web based Targeted Advertising: A Study based on Patent Information. Procedia Econ. Financ. 11, 522–535. doi:10.1016/S2212-5671(14)00218-4

Ding, W., Qin, T., Zhang, X., Liu, T., 2013. Multi-Armed Bandit with Budget Constraint and Variable Costs, in: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. pp. 232–238.

Even-Dar, E., Mannor, S., Mansour, Y., 2002. PAC Bounds for Multi-Armed Bandit and Markov Decision Processes, in: Fifteenth Annual Conference on Computational Learning Theory (COLT). pp. 255–270.

Feldman, J., Henzinger, M., Korula, N., Mirrokni, V.S., Stein, C., 2010. Online stochastic packing applied to display ad allocation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 6346 LNCS, 182–194. doi:10.1007/978-3-642-15775-2_16

Feldman, J., Muthukrishnan, S., Pal, M., Stein, C., Acm, 2007. Budget Optimization in Search-Based Advertising Auctions. Ec'07 Proc. Eighth Annu. Conf. Electron. Commer. 40–49. doi:10.1145/1250910.1250917

Fink, D., 1997. A Compendium of Conjugate Priors. doi:10.1.1.157.5540

Fulgoni, G.M., Mörn, M.P., 2009. Whither the Click? How Online Advertising Works. J. Advert. Res. 49, 134. doi:10.2501/S0021849909090175

Ghosh, A., McAfee, P., Papineni, K., Vassilvitskii, S., 2009. Bidding for representative allocations for display advertising. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 5929 LNCS, 208–219. doi:10.1007/978-3-642-10841-9_20

Gopalan, A., Mannor, S., Mansour, Y., 2014. Thompson Sampling for Complex Online Problems, in: Proceedings of the International Conference on Machine Learning 2014. pp. 1–9.

Hollis, N., 2005. Ten Years of Learning on How Online Advertising Builds Brands. J. Advert. Res. 45, 255–268. doi:10.1017/S0021849905050270

IAB/PwC, 2016. Internet advertising [WWW Document]. Internet Ad Revenue Report, FY 2016. URL http://www.pwc.com/gx/en/global-entertainment-media-outlook/segment-insights/internet-advertising.jhtml

Jun, K., Jamieson, K., Nowak, R., Zhu, X., 2016. Top Arm Identification in Multi-Armed Bandits with Batch Arm Pulls, in: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS). pp. 139–148.

Kaelbling, L.P., 1993. Learning in Embedded Systems. MIT Press.

Landry, E., Vollmer, C., 2010. HD Marketing 2010 : Sharpening the Conversation.

Lee, K., Orten, B., Dasdan, A., Li, W., 2012. Estimating Conversion Rate in Display Advertising from Past Performance Data, in: KDD '12 Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, pp. 768–776. doi:10.1145/2339530.2339651

Luce, R., 1959. Individual Choice Behavior: A Theoretical Analysis, Individual Choice Behavior: A Theoretical Analysis. doi:10.2307/2282347

Maron, O., Moore, A., 1993. Hoeffding Races: Accelerating Model Selection Search for Classification and Function Approximation. Adv. Neural Inf. Process. Syst. 59–66.

Muthukrishnan, S., 2009. Ad exchanges: Research issues. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 5929 LNCS, 1–12. doi:10.1007/978-3-642-10841-9_1

Niculescu-Mizil, A., 2009. Multi-armed bandits with betting, in: COLT 2009 Workshop. pp. 133–138.

Okazaki, S., Taylor, C.R., 2013. Social media and international advertising: theoretical challenges and future directions. Int. Mark. Rev. 30, 56–71. doi:10.1108/02651331311298573

Perchet, V., Rigollet, P., Chassang, S., Snowberg, E., 2016. Batched bandit problems. Ann. Stat. 44, 660–681. doi:10.1214/15-AOS1381

Pew Research Center, 2015. State of the News Media 2015. doi:10.1017/CBO9781107415324.004

Roels, G., Fridgeirsdottir, K., 2009. Dynamic revenue management for online display advertising. J. Revenue Pricing Manag. 8, 452–466. doi:10.1057/rpm.2009.10

Sahin Cem Geyik, A.S., Dasdan, A., 2014. Multi-Touch Attribution Based Budget Allocation in Online Advertising. Proc. Eighth Int. Work. Data Min. Online Advert. 1–9. doi:10.1145/2648584.2648586

Schwartz, E.M., Bradlow, E., Fader, P., 2013. Working Paper Customer Acquisition via Display Advertising Using Multi- Armed Bandit Experiments.

Schwartz, E.M., Ross, S.M., Bradlow, E., Fader, P., Bradlow, E.T., Fader, P.S., 2017. Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments. Mark. Sci. 36, 1–23.

Scott, S.L., 2010. A modern Bayesian look at the multi-armed bandit. Appl. Stoch. Model. Bus. Ind. 26, 639–658. doi:10.1002/asmb.874

Slivkins, A., 2013. Dynamic Ad Allocation: Bandits with Budgets.

Thompson, W., 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. Biometrika 25, 285–294. doi:10.2307/2332286

Tkachenko, Y., 2014. Optimal allocation of digital marketing budget: The empirical Bayes approach. J. Mark. Anal. 2, 162–172. doi:10.1057/jma.2014.14

Tran-thanh, L., 2012. Budget – Limited Multi – Armed Bandits Thesis.

Vermorel, J., Mohri, M., 2005. Multi-armed bandit algorithms and empirical evaluation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 3720 LNAI, 437–448. doi:10.1007/11564096_42

Watkins, C.J.C.H., 1989. Learning from delayed rewards. Cambridge University. doi:10.1016/0921-8890(95)00026-C

Xia, Y., Ding, W., Zhang, X., Yu, N., Qin, T., 2015a. Budgeted Bandit Problems with Continuous Random Costs, in: JMLR: Workshop and Conference Proceedings. pp. 317–332.

Xia, Y., Li, H., Qin, T., Yu, N., Liu, T.-Y., 2015b. Thompson Sampling for Budgeted Multi-armed Bandits, in: 24th International Joint Conference on Artificial Intelligence. pp. 3960–3966.

Xia, Y., Qin, T., Ding, W., Li, H., Zhang, X.-D., Yu, N., Liu, T.-Y., 2017. Finite Budget Analysis of Multi-armed Bandit Problems. Neurocomputing 0, 1–17. doi:10.1016/j.neucom.2016.12.079

Xia, Y., Qin, T., Ma, W., Yu, N., Liu, T., 2016. Budgeted Multi-armed Bandits with Multiple Plays. Proc. Twenty-Fifth Int. Jt. Conf. Artif. Intell. 2210–2216.

Zhang, W., Zhang, Y., Gao, B., Yu, Y., Yuan, X., Liu, T.-Y., 2012. Joint optimization of bid and budget allocation in sponsored search. Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '12 1177. doi:10.1145/2339530.2339716

Zhou, Y., Chakrabarty, D., Lukose, R., 2008. Budget constrained bidding in keyword auctions and online knapsack problems. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 5385 LNCS, 566–576. doi:10.1007/978-3-540-92185-1_63