Exploiting Channels Interdependence in Internet Advertising Campaigns Optimization

AI & R Lab
Laboratorio di Intelligenza Artificiale e Robotica del Politecnico di Milano

Supervisor: Prof. Marcello Restelli
Cosupervisor: Francesco Trovò
Cosupervisor: Alessandro Nuara

Master Graduation Thesis by:
Nicola Sosio, Student id 876137
Maria Chiara Zaccardi, Student id 875520

Academic Year 2017-2018
Contents

Abstract XI

Sommario XIII

Ringraziamenti XV

1 Introduction 1

1.1 Problem .................................................. 2

1.2 Research Target .......................................... 2

1.3 Document Structure ...................................... 3

2 Theoretical Background 5

2.1 Digital Advertising ....................................... 5

2.2 Marketing Funnel ......................................... 6

2.3 Multichannel marketing .................................. 7

2.3.1 Channels in Online Advertising ....................... 8

2.4 Attribution Models ....................................... 12

3 Methodologies 15

3.1 Augmented Dickey Fuller Test ............................ 15

3.2 Wald Test .................................................. 16

3.3 Granger Causality ......................................... 17

3.3.1 Standard Approach ................................. 18

3.3.2 Surplus Lag Approach ............................... 19

3.4 Gaussian Process .......................................... 21

4 State of the art 25

4.1 Causality detection ....................................... 25

4.2 Bid/Budget optimization ................................. 29
# List of Figures

2.1 Marketing funnel. .................................................. 7
2.2 Main channels in online advertising. .......................... 9
2.3 Infographic representing Google AdWords ranking system. 10
2.4 Infographic representing Facebook’s Relevance Score. .... 12

3.1 Time series cross-validation. ................................. 20
3.2 Samples from prior and posterior of a GP with Squared Exponential covariance function. ..................... 23

7.1 Interaction between agent and environment. .......... 56
7.2 Interdependence graph $G$ for the Experimental Setting 1. 58
7.3 Revenue $R_t$ over time for the Setting 1. ............... 61
7.4 Interdependence graph $G$ for the Experimental Setting 2. 62
7.5 Revenue $R_t$ over time for the Setting 2. ............... 64
7.6 Interdependence graph $G$ for the Experimental Setting 3. 65
7.7 Revenue $R_t$ over time for the Setting 3. ............... 68
7.8 Interdependency graph inferred from the data from Campaign 1. .................................................. 69
7.9 Interdependency graph inferred from the data from Campaign 2. ................................................. 70
7.10 Interdependency graph inferred from the data from Campaign 3. ............................................. 71
7.11 GPs estimation of the potential number of impressions of sub-campaigns $C_5$ and $C_6$ depending on the influence index. ........ 74
7.12 Daily cost, the daily CPL, and the daily conversions of the sub-campaigns of the sub-graph $S_1$. .......................... 75
7.13 Relationship between the daily conversions and the daily cost of the sub-graph $S_1$. .............................. 76
7.14 Daily cost, the daily CPL, and the daily conversions of a real-world campaign $C$. ................................. 77
7.15 Relationship between the daily conversions and the daily cost of a real-world campaign $C$. \hfill 78
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Parameters values for Setting 1</td>
<td>59</td>
</tr>
<tr>
<td>7.2</td>
<td>Click probability $p_{ij}^{(cl)}$ and conversion probability $p_{ij}^{(co)}$ for Setting 1</td>
<td>59</td>
</tr>
<tr>
<td>7.3</td>
<td>Research probability $p_{ij}^{(res)}$ for Setting 1</td>
<td>59</td>
</tr>
<tr>
<td>7.4</td>
<td>Percentages of rejecting the non-causality hypothesis for Setting 1 with $\tau = 64$, and $\tau = 80$</td>
<td>61</td>
</tr>
<tr>
<td>7.5</td>
<td>Percentages of rejecting the non-causality hypothesis for Setting 1 with $\tau = 96$</td>
<td>62</td>
</tr>
<tr>
<td>7.6</td>
<td>Parameters of the Settings 2</td>
<td>63</td>
</tr>
<tr>
<td>7.7</td>
<td>Click probability $p_{ij}^{(cl)}$ and conversion probability $p_{ij}^{(co)}$ for Setting 2</td>
<td>63</td>
</tr>
<tr>
<td>7.8</td>
<td>Research probability $p_{ij}^{(res)}$ for Setting 2</td>
<td>63</td>
</tr>
<tr>
<td>7.9</td>
<td>Percentages of rejecting the non-causality hypothesis for Setting 2 with $\tau = 64$, $\tau = 78$, $\tau = 96$</td>
<td>64</td>
</tr>
<tr>
<td>7.10</td>
<td>Parameters of the Settings 3</td>
<td>65</td>
</tr>
<tr>
<td>7.11</td>
<td>Click probability $p_{ij}^{(cl)}$ for Setting 3</td>
<td>66</td>
</tr>
<tr>
<td>7.12</td>
<td>Conversion probability $p_{ij}^{(co)}$ for Setting 3</td>
<td>66</td>
</tr>
<tr>
<td>7.13</td>
<td>Research probability $p_{ij}^{(res)}$ for Setting 3</td>
<td>66</td>
</tr>
<tr>
<td>7.14</td>
<td>Percentages of rejecting the non-causality hypothesis for Setting 3 with $\tau = 100$</td>
<td>67</td>
</tr>
<tr>
<td>7.15</td>
<td>Percentages of rejecting the non-causality hypothesis for Setting 3 with $\tau = 150$ and $\tau = 200$</td>
<td>67</td>
</tr>
<tr>
<td>A.1</td>
<td>Percentages of rejecting the non-causality hypothesis for System 1</td>
<td>88</td>
</tr>
<tr>
<td>A.2</td>
<td>Percentages of rejecting the non-causality hypothesis for System 2</td>
<td>88</td>
</tr>
<tr>
<td>A.3</td>
<td>Percentages of rejecting the non-causality hypothesis for System 3</td>
<td>89</td>
</tr>
<tr>
<td>A.4</td>
<td>Percentages of rejecting the non-causality hypothesis for System 4</td>
<td>90</td>
</tr>
</tbody>
</table>

VII
A.5 Percentages of rejecting the non-causality hypothesis for System 5

90
# List of Algorithms

<table>
<thead>
<tr>
<th></th>
<th>Algorithm</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IDL</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>IntegrationOrder Subroutine</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>OPT($\hat{n}$, $\hat{v}$, $\hat{w}$, $\hat{D}$)</td>
<td>46</td>
</tr>
<tr>
<td>4</td>
<td>DiscardParetoDominatedSolutions</td>
<td>47</td>
</tr>
<tr>
<td>5</td>
<td>DiscardSolutions</td>
<td>47</td>
</tr>
</tbody>
</table>
Abstract

Over recent years, Online Advertising has become one of the main tools in which companies invest to promote and sell their products and/or services. Generally, in the context of Online Advertising, companies handle several advertising campaigns to reach as many people as possible. An Internet advertising campaign can include up to thousands of sub-campaigns on multiple channels, like search, social, and display. Companies should exploit all the marketing channels to create more opportunities to generate revenue. To do so, sub-campaigns’ bid and daily budget need to be optimized every day, subject to a cumulative budget constraint. Such an optimization process is often unaffordable for humans and its complete, or partial, automation can be crucial. As also shown by marketing funnel models, the sub-campaigns are usually interdependent, e.g., display ads induce awareness, increasing the number of impressions—and, thus, also the number of conversions—of search ads. This interdependence is widely exploited by humans in the optimization process, whereas, to the best of our knowledge, no algorithm for Internet Advertising Campaign Optimization takes it into account. In this work, we provide the first algorithm which exploits the sub-campaigns interdependence. The challenge is designing a model that guarantees a satisfactory trade-off between accuracy and amount of data required for training it. To do that, we provide an algorithm, called IDL, that, employing Granger Causality and Gaussian Processes, learns the campaign interdependence model from past data and returns an optimal stationary bid/daily budget allocation. Finally, we provide empirical evidence that the proposed algorithm leads to better performances in both realistic and real-world settings when compared with previous approaches.
Sommario

Internet è diventato negli ultimi anni uno dei canali più utilizzati dalle aziende che vogliono promuovere i loro prodotti e servizi. Investire nella pubblicità on line permette di raggiungere un numero notevole di persone, di mettere a punto una comunicazione efficace e di rivolgersi a target di consumatori diversificati con offerte ad hoc. Una campagna pubblicitaria online può comprendere migliaia di sottocampagne che si inseriscono in diversi canali, come search, social e display. Sfruttando tutti i canali a disposizione, le aziende possono creare maggiori opportunità di guadagno. La bid e il budget di ogni sottocampagna sono soggetti a un vincolo di budget complessivo e richiedono di essere ottimizzati ogni giorno. Questo processo di ottimizzazione è spesso troppo complesso per essere gestito da un essere umano, quindi, la sua totale o parziale automatizzazione, è cruciale. Come mostrato dai modelli di marketing funnel, le sottocampagne sono spesso dipendenti tra di loro. Ad esempio, la pubblicità sul canale display aumenta il numero di impression e quindi il numero di conversioni del canale search. Queste relazioni sono ampiamente sfruttate dagli esperti del settore nel processo di ottimizzazione, tuttavia, in passato nessun algoritmo presente nella letteratura scientifica prende in considerazione tali dipendenze. La sfida è di tale modellizzazione è quindi quella di trovare un modello che garantisca un compromesso soddisfacente tra precisione e quantità di dati richiesti. Per fare ciò, proponiamo un algoritmo, chiamato IDL, che, utilizzando la Granger Causality e i Gaussian Process, apprende un modello dai dati passati e restituisce un’allocazione ottima e stazionaria di bid/budget giornalieri. Infine, forniamo evidenze empiriche che l’algoritmo proposto raggiunge risultati migliori sia con dati sperimentali, sia in un contesto reale rispetto agli approcci precedenti.
Ringraziamenti

Ringraziamo il Prof. Marcello Restelli per la sua disponibilità e per aver reso possibile la concretizzazione di questo lavoro. Grazie anche a Francesco Trovò e ad Alessandro Nuara per averci seguito ed aiutato in ogni momento, sostenendoci sempre. Infine, un ringraziamento speciale ai nostri genitori che ci hanno permesso di arrivare fino a qui.
Chapter 1

Introduction

"Now when we speak of an information-rich world, we may expect, analogically, that the wealth of information means a dearth of something else - a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

Herbert A. Simon, 1971

Since the early stages of the Internet, one of the most remunerative ways to economically exploit this novel media channel is online advertising. However, technology has changed the way consumers access information over time. Traditional media, such as TV, have been overtaken by the endless source of digital content accessible anywhere at any time. In today’s digital landscape, the reality is that most internet users are Multi-Device users. As reported by [Mandar, 2017], on average, the typical digital consumer now owns 3.23 of them. This shift in information’s access has led to the attention of the economy field. Herbert A. Simon is thought to be the first person to address attention economics in a study called The Empirical Economics of Online Attention. In the context of Online Advertising users are exposed to large volumes of digital ads; however, most of them are ignored. Thus, the attention of consumers has become a valuable commodity because of its scarcity. In essence, it has become the currency of the online world.

Companies must exploit all the available marketing channels to increase users’ attention and create a higher opportunity of generating revenues. The diversification of ad channels is a crucial degree of freedom that companies may exploit when setting up an advertising campaign. Indeed, different channels deeply affect each other’s performance as Internet users regularly
1.1. Problem

Multi-channel advertising is nowadays a common practice that allows reaching a wide population through different channels. In the context of multi-channel online advertising, where different advertising channels are involved, the use of assist sub-campaigns is a common strategy. In the marketing field, an assist campaign is a sub-campaign that does not provide direct conversions but increase the number of conversions of another sub-campaign. The assist sub-campaign generates either impression-assisted conversions or click-assisted conversions. In the former case, an impression generated by the assist sub-campaign leads to a conversion, while in the latter case a conversion is generated by a click on an assist sub-campaigns ad. Furthermore, both experts in the field of Internet advertising and studies in the Internet economic field demonstrated that impressions, clicks, and conversions from a sub-campaign might be influenced by the same quantities of the other sub-campaigns. For instance, when a user sees a display ad, is is more likely she clicks on the sponsored ad in the search engine results page. Capturing these interdependence among campaigns is a crucial task to provide effective methods to optimize an Advertising campaign in an automatic way.

1.2 Research Target

We extend the work by Nuara et al. [2017], designing an algorithm based on both learning and optimization techniques that can be adopted for the optimization of real-world Internet advertising campaigns and that, exploiting sub-campaigns interdependence, outperforms the state of the art algorithms of this specific optimization problem. To do that, we provide a novel model that, on the one hand, is expressive enough to capture the interdependences and, on the other hand, is simple enough to require few data for its estimation. Based on the proposed model, we design a data-driven
optimization algorithm, called IDL, which consists of two phases: the *Interdependence Graph Learning Phase* and the *Estimation and Optimization Phase*. In the former phase, the IDL algorithm learns the sub-campaigns interdependence structure (represented as a graph), identifying the pairs of sub-campaigns with the most significant interdependences by applying the Granger Causality test. This is crucial since the number of pairs of interdependent sub-campaigns dramatically increases the amount of data required to have accurate estimates of the model parameters. In the latter phase, the IDL algorithm computes the optimal joint bid/daily budget allocation exploiting Gaussian Process modeling [Rasmussen and Williams 2006] and an *ad hoc* dynamic programming procedure. In particular, Gaussian Processes allow us to assume some simple form of function regularity to describe the relationships among the problem parameters, without forcing them to be described by some specific family of curves. Finally, we show that neglecting the sub-campaigns interdependence can lead to massive losses even in simple and common scenarios and we theoretically bound the loss of our algorithm. Furthermore, we experimentally evaluate its performance in both realistic and real-world settings, showing the superiority of its performance compared to the previous approaches that neglect the sub-campaigns interdependence.

### 1.3 Document Structure

The remaining part of this document is organized as follows:

**Chapter 2** provides the basic knowledge required to better understand the context of Online Advertising.

**Chapter 3** describes the methodologies needed to introduce our method.

**Chapter 4** provides the state of the art of the research about causality detection and bid/budget optimization.

**Chapter 5** describes the problem formulation and the model formalization we adopt.

**Chapter 6** presents IDL, the algorithm we propose as solution to the problem of optimizing an Internet advertising campaign with sub-campaigns interdependence and provides an analysis on the theoretical properties of the IDL algorithm.

**Chapter 7** shows experimental results both in synthetic and real-world settings.

**Chapter 8** concludes the thesis with some observations and suggests possible future developments of this work.
Appendix A provides an analysis on the use of Granger Causality in the advertisement field.
Chapter 2

Theoretical Background

In this section, we provide the basic knowledge needed to better understand the context of online advertising. Initially, we present what is digital advertising and we provide a set of basic definitions in this field. Next, we describe what is the marketing funnel and how it works, then we explain what is multichannel marketing applied to online advertising and which are its main channels. Lastly, we provide a brief specification about attribution models.

2.1 Digital Advertising

Digital advertising, also known as online advertising or Internet advertising, uses the online channel as an opportunity to communicate marketing messages to users or visitors. There are many new and emerging formats for digital advertising, although it is commonly accepted that it would include: e-mail marketing, search engine marketing, social media marketing, display advertising, and mobile advertising. Digital advertising is a way to reach customers, possibly through different channels, on the Internet, which is the place where they spend the majority of their time. Before going into details of these concepts, we introduce some definitions required to fully understand the following chapters.

**Impression** occurs when an advertisement, or any other form of digital media, renders on a user’s screen. Thus, an impression is simply a view of an ad, it could be a banner ad, video ad, or text-ad. Digital marketing has made impression tracking significantly more quantitative than offline advertising. In fact, a billboard owner has no concrete way of estimating the number of
impressions his platform grants advertisers. Conversely, impression-based online campaigns can measure impressions objectively. These campaigns are generally sold in terms of cost-per-thousand (CPM) impressions.

**Click** occurs when a user clicks on an ad after she sees the ad and logs an impression. The click will bring the user into the landing page of the sponsored brand. Click-based online campaigns are generally sold in terms of cost-per-clicks (CPC). Pay-per-click is the most common model in case of click-based campaigns, where advertisers are only charged when a user actually clicks on their ad. The click-through-rate (CTR) is a common metric for measuring the number of clicks advertisers receive on their ads per number of impressions.

**Conversion** is an action that occurs when someone interacts with an ad (e.g., clicks a text ad or views a video ad) and then takes a particular action on the website. A conversion can include events, such as a visitor filling out a lead capture form or an online purchase. These actions are defined by the advertiser as the campaign goals.

In conversion-based online campaigns, the success of a campaign is measured in terms of **conversion rate**, that is defined as the number of conversions divided by the number of impressions served. The higher is the conversion rate the higher is the success of a campaign. So, in other words, the conversion rate is the percentage of visitors to a website that completes the desired goal (a conversion) out of the total number of visitors.

### 2.2 Marketing Funnel

The marketing funnel is a model describing the process of turning leads into customers, as understood from a marketing (and sales) perspective. The idea is that, like a funnel, marketers cast a broad net to capture as many leads as possible, and then slowly nurture prospective customers through the purchasing decision, narrowing down these candidates in each stage of the funnel. As people progress through the funnel, their intent to buy steadily increases.

In the ideal case, the marketing funnel would be a marketing cylinder where all the users that are aware of a specific campaign turn into customers. Thus, the main goal for a marketer is to turn as many leads into customers as possible, to make the funnel more cylindrical.

There are four main stages in the marketing funnel (Figure 2.1): awareness, interest, intent, and decision.
2.3. Multichannel marketing

Awareness is the uppermost stage of the marketing funnel. At this stage, the user becomes aware of a product or brand for the first time. Potential customers are drawn into this stage through marketing campaigns that have the sole purpose of introducing the brand and foster familiarity with the name of the brand. Once leads are generated, users move on the interest stage, where they learn more about the company, its products, and any helpful information and research it provides. Here, the content of advertising slightly changes: it is more targeted and specific. The main purpose of this stage is engaging users and showing them the real value of the company. To get to the intent stage, prospects must demonstrate that are interested in buying a brand’s product. Generally, this can happen in a survey, after a product demo, or when a product is placed in the shopping cart on an e-commerce website. Finally, in the last stage of the marketing funnel, a prospect has made the decision to convert and turns into a customer.

However, often happens that the marketing funnel does not end with a purchase in the decision stage. Online marketing continues to play an important role even after the conversion phase, to convince people to buy again. Most of what happens after the purchase depends on the quality of the brand’s product, on the service or/and on the customer service. The bright side is that if someone has already made a purchase, it is usually easier and cheaper to convince them to buy again. At this stage in the buying process, the goal of online marketing is to reinforce the positive things about the purchase and encourage people to come back for more.

2.3 Multichannel marketing

Multichannel marketing refers to the practice of engaging prospects, leads and customers across a combination of indirect and direct communication channel hoping to ignite interest in a brand and what that brand offers. It is important and helpful because it allows companies to be where customers
are. Buyers are in complete control of the sales process. There are more channels, more devices, and more options available when people need to look up for information they need to make a purchase decision. So, in other words, multichannel marketing is the practice of using multiple channels to reach customers and it allows them to complete desired conversions on the channel they are most comfortable with.

Multichannel marketing faces many challenges. The main difficulty is to keep consistency and coordination across all the channels or media used. This does not mean that different channels cannot be used for different tasks and with different aims. The problem is about finding the right mix of marketing channels that would provide the best results. Moreover, with multiple channels, it is more complex to monitor the effectiveness of each marketing channel. Thus, understanding which digital marketing channels contributed to a particular conversion, i.e. multichannel attribution (see Section 2.4 for more details about attribution models), become crucial.

On the other hand, the benefits of using more than just one channel for doing marketing are numerous. Firstly, even if the idea behind multichannel marketing is about simply advertising across different channels, its main advantage comes from creating cross-channel marketing strategies by considering the relationship among channels. Secondly, this approach to marketing across various channels is about moving and guide potential customers, in order to increase the conversions, through the marketing funnel (Figure 2.1), from awareness to conversion. Multiple channels are already used by most businesses’ marketing departments. However, in the majority of cases, multichannel use has not been strategized, thus has not been exploited at its full potential.

2.3.1 Channels in Online Advertising

In today’s modern era, the channels commonly used in the context of online advertising, are search, programmatic and social as Figure 2.2.

Search Advertising

In Paid search marketing the ad is placed within the sponsored listings of a search engine by paying either each time an ad is clicked (pay-per-click) or, less commonly, when an ad is displayed (cost-per-impression, CPM). Paid search advertising is one of the most popular forms of pay-per-click advertising using an auction-based mechanism. It allows showing digital ads on search engines results pages such like Google, Bing, and Yahoo. This is one of the most popular and effective types of online advertising and
2.3. Multichannel marketing

Figure 2.2: Main channels in online advertising.

the demand for the top ad ranking is enormous. This is the reason why search companies as Google trigger an auction anytime there are at least two advertisers bidding for keywords that are related to search queries that users consistently enter.

Paid search advertising is one of the most popular forms of pay-per-click advertising using an auction-based mechanism that works as follow. Search engine companies sell advertisement slots based on users’ search queries via an auction. The advertiser determines a set of keywords of their interest and then must create ads, set the bids for each keyword and provide a total, often daily, budget. An advertiser bids on a keyword, which can be a single word or composed by a set of words. When a user poses a search query, the Internet search company determines the advertiser whose keywords match the query and who still have budget left over, runs an auction among them, and show the set of ads corresponding to the advertisers who ‘win’ the auction. The way in which advertiser’s keywords match a query depends on the search engine. Note that an advertiser makes a single bid for a keyword that remains in effect for a period of time, and the keyword could match many different user queries throughout the day. Each user query might have a different set of advertiser competing for clicks. The advertiser could also bid different amounts on multiple keywords, each matching a (possibly overlapping) set of user queries. [Feldman et al., 2007].

The type of auction depends on the search engine company. Google’s advertising system is Google AdWords and it is shown in Figure 2.3. Advertisers bid on certain keywords to let appear their ads in Google’s search result. Google looks at two key factors to determine the ad ranks: maximum CPC or bid and quality score. The maximum bid is the maximum bid an advertiser specify in his keywords, while the quality score is a metric to determine how relevant and useful an ad is to the user in
2.3. Multichannel marketing

Figure 2.3: Infographic from Wordstream [2018b] representing Google AdWords ranking system.

terms of click-to-rate, relevance and landing page. Since paid search ads target specific search queries, users who see search ads are likely interested in evaluating the advertised item. For this reason, paid search advertising is placed at lower levels of the marketing funnel.

Display Advertising

Display ads are shown on specific Web pages sites and can be along the top of web pages such as the traditional banner ad, or the larger text billboard but they can also be videos. There are countless combinations of formats, sizes, and styles that represent the flexibility of these kinds of ads. Display ads appear on distinct sections of the site that are specifically reserved for paid advertising. Display ads also travel far, given the millions of websites reached by Google’s Display Network, which place display ads on a huge network of sites across the internet. The search giant can match your ads up to websites and apps based on keywords or your own targeting preferences. According to Google, the Display Network reaches over 90% of global internet users expanding across two million sites.

They are also fairly straightforward to measure: display advertising analytics allow to track the number of clicks, impressions, and conversions the ad has generated in real-time to provide an up-to-date picture of what is resonating with consumers. On the other hand, their viewability is affected by ad blocking, the use of which has skyrocketed over the past year, particularly in Europe. Display ads are so widely used, they are everywhere and for this reason, people tend to ignore them. The main difference between display
advertising and paid search advertising is that the former is a firm-initiated channel and the latter is a customer-initiated channel. A firm-initiated channel is a channel in which the company begins the communication, while in the customer-initiated channel consumers look for information on their own free will. The goal of display advertising is quite different: it is a way to introduce the brand and increase brand awareness among users. Display advertising is a brand awareness strategy that aims to move potential customers through the marketing funnel. Typically, when a user sees a display ad he may be more likely to click on the sponsored ad in the search engine results page, therefore measuring ROI is not as simple as just counting clicks.

**Social Advertising**

Social media marketing is a form of internet marketing that involves creating and sharing content on social media networks to achieve your marketing and branding goals. Social media marketing includes activities like posting text and image updates, videos, and other content that drives audience engagement, as well as paid social media advertising. This kind of advertising can be found is social media like Facebook, Twitter, Linkedin, Instagram, etc. One of the major benefits of social media advertising is that advertisers can take advantage of the users’ demographic information and target their ads appropriately. Again, the type of auction depends on social media.

On **Facebook**, quality score is known as Relevance Score. Facebook calculates this score based on how engaging an ad is. Of course, the more engagement, the better the Relevance Score. As Figure 2.4 shows, a high Relevance Score should be one of the primary goals for a Facebook advertiser. Higher Relevance Scores translate to greater impression shares and lower costs-per-engagement, making ads more visible and more cost-effective.

**Twitter**, on the other hand, is a goldmine for keyword targeting (an option that Facebook does not offer), which is very different from how works in the search channel. Keyword targeting on Google often feels like you’re marketing to robots. The keywords are rigid, blunt, and direct. Social keywords on Twitter are much more conversational, informational, and use hashtags. This could actually make targeting look more natural. Furthermore, on Twitter, targeted people are the ones that have used certain keywords in their Tweets, in addition to keyword searches on Twitter. As display ads, social media advertising is usually located at higher levels of the marketing funnel, with the same aim of introducing the brand and increase
brand awareness. The main difference is that social network advertising can use demographic information of users and target ads in a better way.

2.4 Attribution Models

In what follows, we introduce attribution models which can help in identifying the channels responsible for the final conversions, and stop channels that do not bring any significant result. Marketing attribution reveals that sale’s history helps marketers to map the journey from lead to customers, taking into account every touch point and interaction along the way. In digital advertising, attribution is the problem of assigning credit to one or more advertisements for driving the user to the desirable actions such as making a purchase. Rather than giving all the credits to the last ad a user sees, multiple-touch attribution allows more than one ads to get the credit based on their corresponding contributions. Multi-touch attribution is one of the most important problems in digital advertising, especially when multiple channels, such as search, display and social are involved, in order to discover the impact of each channel at different stages of the funnel. Typically, multiple advertising channels have delivered advertisement impressions to a user. When the user then makes conversions (e.g., a purchase decision or sign up to a service being advertised) the advertiser wants to determine which ads have contributed to the user’s decision. This step is critical in completing the loop so that one can analyze, report and optimize an advertising campaign. This problem of interpreting the influence of advertisements to the user’s decision process is called the **attribution problem**, ([Shao and Li, 2011]).
2.4. Attribution Models

The most famous and used free tool for web analytics is Google Analytics, offered by Google. Using attribution and attribution modeling in Google Analytics, it is possible to determine the relative value of the marketing channels in terms of conversions and Return On Investment (ROI). There are different possible models, in the following a brief explanation.

**Last non-direct click** This is the default model of Google Analytics. It gives full attribution credit to the channel that produces conversion unless that channel is direct. The general idea is to eliminate the unknowable "direct" from the equation and ignore brand recognition and the possibility that visitors came because they know the company. This model works best for sites that derive the majority of their conversions from paid or organic search.

**Last interaction** The last interaction model gives full attribution credit to the visitor’s last interaction prior to conversion. All other activities of the visitor on the website are discounted. The only difference between the previous model is that this one includes direct traffic.

**Last AdWords Click** Full conversion credit goes to the visitor’s last Adwords click. This model is used to estimate the efficiency of an individual ad campaign and can be easily set up. This model is suitable only for those whose entire marketing strategy relies on using Google AdWords (multi-channel strategies are not covered) otherwise is not very useful.

**First Click** First click gives full attribution credit to the first click interaction of the visitor with the site. According to this model, if a visitor clicks a display banner or a social media post and she converts after days, the entire credit would go to the first interaction. It is difficult to imagine a situation where this would be true, but this model’s intention is to reveal channels that help establish brand awareness.

**Linear** The linear interaction model gives equal credit to each channel. This model is not that useful, as it gives all channels the same amount of credit, regardless of their actual contribution, it represents only an approximation of the real contributions.

**Time decay** The time decay model gives most credit to the most recent interaction channel but still allows for earlier click to play a role. This is one of the most realistic models and represents the idea that the most recent
2.4. Attribution Models

touch points or interactions prior to conversion have more influence on the visitor’s decision, and should, therefore, receive more credit.

**Position-based** The position-based model gives interaction credit according to a channel’s position in the funnel. It is possible to customize the amount of credit given to each position. It is a combination of the first and last click models: the first and the last interaction each receive 40% credit for conversion, while all the steps in between share the remaining 20% equally.

Which model to choose is up to the advertiser and depends on the business model of the website. A possible option could also be a custom model which allows defining the proportion of attributions to each step in the conversion path. Probably it is the best option to design a realistic model for conversion attribution but requires knowledge of the business process and marketing strategy of the business.
Chapter 3

Methodologies

In the Sections 3.1–3.2 we describe the Augmented Dickey-Fuller Test and the Wald Test, that are used in the Granger Causality described in Section 3.3. Then, in Section 3.4 we illustrate the theoretical background behind Gaussian Processes.

3.1 Augmented Dickey Fuller Test

The Augmented Dickey Fuller Test (ADF) [Dickey, 1976] tests whether a time series is stationary or not. Firstly, we define a stationary time series as following:

**Definition 3.1.1 (Stationary).** Given a time series $x_t$, where $t \in \{1, \ldots, T\}$, $x_t$ is stationary if the process satisfies all the following conditions:

- a constant mean over time
  $$E(x_t) = E(x_s) = \mu, \ \forall t \neq s;$$

- a constant, finite variance over time
  $$\sigma^2(x_t) = \sigma^2(x_s) = \sigma^2 < \infty, \ \forall t \neq s;$$

- a constant, finite covariance over time
  $$\text{Cov}(x_t, x_s) = \text{Cov}(x_{t+1}, x_{s+1}) = \gamma_{t-s}, \ \forall t \neq s.$$

Given a time series $x_t$ where $t \in \{1, \ldots, T\}$, to test if $x_t$ is stationary, the following $AR(K)$ model is build:

$$x_t = \sum_{k=1}^{K} \beta_k x_{t-k}, \quad (3.1)$$
where $k \in [1, \ldots, K]$ is the lag of the time series. The $AR(K)$ model can be rewritten as the following regression model:

$$\Delta x_t = \gamma x_{t-1} + \sum_{k=1}^{K-1} \delta_k \Delta x_{t-k},$$  \hspace{1cm} (3.2)

where $\Delta$ is the first difference operator. Once the regression model has been estimated, the unit root test is equivalent to testing the null hypothesis $\gamma = 0$, against the alternative hypothesis $\gamma < 0$. In practice, the null hypothesis for the ADF test states that there is a stochastic trend in a time series that make it unpredictable, while the alternative hypothesis states that the time series can be considered stationary. Formally the test uses:

$$\begin{cases} H_0 : \gamma = 0 \\ H_A : \gamma < 0 \end{cases}.$$  \hspace{1cm} (3.3)

The test statistic corresponding to the ADF test is:

$$DF = \frac{\hat{\gamma}}{SE(\hat{\gamma})},$$  \hspace{1cm} (3.4)

where $SE(\hat{\gamma})$ denotes a consistent estimate of the standard deviation of the OLS estimator $\hat{\gamma}$. Under the null hypothesis, the test statistic in Equation (3.4) has not a given distribution. However, the critical values for the unit root tests have been calculated using Monte Carlo methods by [Dickey and Fuller, 1981]. If the test statistic is less than the critical value, with confidence $\alpha$, we can reject the null hypothesis and the time series is stationary.

### 3.2 Wald Test

The Wald test is a way of testing the significance of a particular explanatory variable, or group of explanatory variables, in a model. Let us partition the vector of explanatory variables into two components $X = [X_1, X_2]^T$ with respectively $N$ and $M$ elements. To evaluate if the first group of explanatory variables $X_1 = [x_1, x_2, \ldots, x_N]^T$ is significant, the following model is estimated:

$$y = c + \sum_{i=1}^{N} a_i x_i + \sum_{j=N+1}^{M} b_j x_j$$

$$= c + \theta_1 X_1^T + \theta_2 X_2^T.$$  \hspace{1cm} (3.5)
3.3. Granger Causality

Once the regression model defined in Equation (3.5) has been estimated it is possible to assess whether or not the group of explanatory variables \( X_1 \) is significant. More specifically, let \( \hat{\theta}_n \in \Theta \subseteq \mathbb{R}^{N+M} \) be the estimate of the real parameter \( \theta_0 = [\theta_1, \theta_2]^T \) of the regression model obtained by maximizing the log-likelihood over the whole parameter space \( \Theta \):

\[
\hat{\theta}_n = \arg \max_{\theta \in \Theta} \ln[L(\theta, X)],
\]

where \( L(\theta, X) \) is the likelihood function and \( X = [X_1, X_2]^T \). The test statistic \( W_n \) employed in the Wald test is:

\[
W_n = (\hat{\theta}_1 - \bar{\theta})' [\hat{V}_n]^{-1} (\hat{\theta}_1 - \bar{\theta}),
\]

where \( \hat{V}_n \) is a consistent estimate of the covariance matrix of \( \theta_n \) and \( \bar{\theta} \) is a zero vector of size \( N \). Under the null hypothesis

\[
H_0 : \hat{\theta}_1 = \bar{\theta},
\]

the Wald statistic \( W_n \) converges in distribution to a Chi-squared distribution with \( N \) degrees of freedom, i.e., \( W_n \overset{d}{\to} \chi^2_N \).

3.3 Granger Causality

The purpose of using Granger Causality is to find causality relations among time series. In Section 3.3.1 we explain the standard procedure for testing Granger causality [Granger, 1969]. However, this procedure assumes that the time series are stationary and not cointegrated. Therefore, in Section 3.3.2 we illustrate another approach that allows to apply the Granger Causality framework also to time series which presents non-stationarity and cointegration characteristics. At first, we require to introduce some key concepts, i.e., integration order and cointegration.

Definition 3.3.1 (Integration order). A variable \( x_t \) is said to be an integrated process of order \( I(d) \) if \( \Delta^d x_t \) is \( I(0) \), where the \( \Delta \) operator represents the first difference operator.

In the case a series is stationary, the order of integration is zero and this is denoted by \( I(0) \). Furthermore, if there are two or more time series with the same order of integration, they may be cointegrated.

Definition 3.3.2 (Cointegration). Given two time series \( x_t \) and \( y_t \), which are \( I(d) \), if there exists \( \beta \) such that \( y_t - \beta x_t = \mu_t \) where \( \mu_t \) is integrated \( I(d - b) \), i.e., of order \( d - b < d \), \( x_t, y_t \) are said to be cointegrated of order \( d - b \). \( x_t, y_t \) are denoted as \( CI(d, b) \).
3.3. Granger Causality

3.3.1 Standard Approach

The formal definition of Granger Causality [Granger, 1969] is the following:

**Definition 3.3.3** (Granger Causality). A variable $x_t$ is Granger causing variable $y_t$ if the presence of $x_t$ allow to better explain $y_t$.

Granger causality is based on prediction and it is a statistical concept of causality. According to Granger causality, if a variable $x_t$ Granger-causes (or $G$-causes) a variable $y_t$, then past values of $x_t$ should contain information that helps predict $y_t$ above and beyond the information contained in past values of $y_t$ alone. Often, we have that $x_t$ Granger-causes $y_t$ and $y_t$ Granger-causes $x_t$; in this case we say that there is a feedback.

To evaluate if there are causal relationship between $x_{1,t}$ and $x_{2,t}$ we estimate the following VAR($p$) model:

$$
\begin{align*}
  x_{1,t} &= a_0 + \sum_{i=1}^{p} a_i x_{1,t-i} + \sum_{i=1}^{p} b_i x_{2,t-i} \\
  x_{2,t} &= c_0 + \sum_{i=1}^{p} c_i x_{1,t-i} + \sum_{i=1}^{p} d_i x_{2,t-i}
\end{align*}
$$

(3.9)

After that, by means of a Wald test (see Section 3.2 for more details), it is possible to asses the causal relationship. The Wald test is a way to find out if the lagged values of an independent variable $x_t$ in a model are relevant to predict the dependent variable $y_t$. This means that the dependent variable adds something to the model, while if the test fails, the lagged values of $x_t$ can be removed without affecting the model in any meaningful way. Wald test consider as null hypothesis the relevance of a variable in the model. If the null hypothesis is rejected we can remove the variable $x_t$ from the model, otherwise, if the test shows that the parameters are jointly different from zero, we should include the variable in the model. In other words, the Wald Test can shows us whether or not the variable is significantly contributing in the model. In particular, two Wald tests, in which the null hypothesis states non-existence of causal relationship, are carried out:

$$
\begin{align*}
  H_0 : b_i = 0 \quad \forall i \in \{1, \ldots, p\} \\
  H_A : \exists b_k \neq 0
\end{align*}
$$

(3.10)

$$
\begin{align*}
  H_0 : c_i = 0 \quad \forall i \in \{1, \ldots, p\} \\
  H_A : \exists c_k \neq 0
\end{align*}
$$

(3.11)

When there is statistical evidence for rejecting the null hypothesis in Equation (3.10) with confidence $\alpha$, we say that $y_t$ Granger-causes $x_t$, while
when there is statistical evidence for rejecting the null hypothesis in Equation (3.11) with confidence $\alpha$, we say that $x_t$ Granger-causes $y_t$.

### 3.3.2 Surplus Lag Approach

The approach defined in Section 3.3.1 cannot always be applied, i.e., when the time series are both non-stationary and cointegrated. In such a situation, we require to use a test addressing the non-stationary and the cointegration of the variables. For this reason it is not possible to apply the Granger Causality test as is, since the test statistic would not follow the aforementioned Chi-squared distribution. A possible procedure to follow is the surplus lag approach, introduced in Toda and Yamamoto [1995]. In this work, the authors show that it is possible to estimate a VAR model formulated in levels and test general restrictions on the parameters matrices even if the processes may be integrated or cointegrated of an arbitrary order.

The key points for building this alternative model are the selection of the integration order $m_i$ for each variable and the order of the VAR model that has generated the time series, $p$. The estimated model will be a $(p+m_{\text{max}})$th-order VAR, where $m_{\text{max}} := \max_i m_i$ is the maximal order of integration that we suspect might occur in the process. Therefore, at first, we compute the integration order $m_i$ for each time series, i.e., the minimum number of applications of the finite difference operator $\Delta$ needed to get a stationary series, defined as $I(m_i)$. To test when a series, after being differentiated $m$ times, become stationary we use the Augmented Dickey Fuller Test (ADF) (described in Section 3.1).

Once we have all the integration orders $m_i$, we compute $m_{\text{max}}$, i.e., the maximum order of integration among the time series, which we suspect might occur in the model. Subsequently, we need to elect the order of the VAR model $p$. The order can be selected using either an information criteria (e.g., BIC [Schwarz and others 1978], AIC [Akaike 1973], FPE [Akaike 1969, 1970], HQIC [Hannan and Quinn 1979]) or by cross validation. When dealing with time series, the standard cross validation approach cannot be used because of the sequentiality of data. We use a series of test sets, each consists of a set of consecutive observations. The corresponding training set consists only of observations that occurred prior to the observation that forms the test set, thus, no future observations can be used in constructing the forecast. Figure 3.1 illustrates the series of training and test sets, where the yellow observations form the training sets, and the red observations form the test sets.

Then we use a VAR with additional $m_{\text{max}}$ lags, i.e., $k = p + m_{\text{max}}$, as
3.3. Granger Causality

follows:

\[ x_{m,t} = c_0 + \sum_{i=1}^{p} a_i x_{m,t-i} + \sum_{j=p+1}^{m_{\text{max}}} a_j x_{m,t-i} + \]
\[ + \sum_{n=1}^{N} \left( \sum_{i=1}^{p} b_{n,i} x_{n,t-i} + \sum_{j=p+1}^{m_{\text{max}}} b_{n,j} x_{n,t-i} \right), \]

(3.12)

where \( m, n \in [1, N] \) and \( m \neq n \). Once that the model has been fitted, we perform as many Wald test as the number of dependent variables, for each time series, formally:

\[ \begin{cases} 
H_0 : b_{n,i} = 0 & \forall i \in \{1, \ldots p\}, \forall n \in \{1, \ldots N\} | n \neq m \\
H_A : \exists b_{n,i} \neq 0
\end{cases} \]  

(3.13)

At last, we obtain as a result two matrices: an adjacency matrix \( G \), where \( g_{i,j} = 1 \) if the \( i \)–\( th \) variable Granger cause the \( j \)–\( th \) variable, and a matrix \( P \), where for each causal relation is shown the corresponding p-value.

In order to retain an overall confidence level \( \alpha \) in an analysis involving more than one comparison, the confidence level for each comparison must be tighter than \( \alpha \). This problem is known as multiple comparison problem. In fact, in statistics, when a set of statistical inferences are considered simultaneously or when a subset of parameters are selected based on the observed values, multiple testing correction is needed. In other words, when multiple tests are performed the p-values need to be adjusted and the Type I error rate, i.e, the probability of having a false positive, has to be controlled. In general, if we perform \( n \) hypothesis tests, the probability of having at least one false positive is

\[ P(\text{Type I error}) = 1 - (1 - \alpha)^n. \]  

(3.14)

For the purpose of preventing this from happening and allowing significance levels for single and multiple comparisons to be directly compared, several statistical techniques have been developed. These techniques generally
require a lower significance threshold for individual comparisons, so as to compensate for the number of inferences being made. Bonferroni correction [Dunn 1961], Holm’s method [Holm 1979] and Benjamin-Hochberg [Benjamini and Yekutieli 2001] are examples of methods used to counteract the problem of multiple comparisons.

The simplest and most conservative approach is the Bonferroni correction [Dunn 1961], which sets the alpha value for the entire set of n comparisons equal to the α value for each comparison divided by the number of tests n, i.e., we test each pair of variables with a confidence α′ = α/n.

The Surplus Lag approach is subsequently used in Chapter 6 to detect interdependence among channels with non-stationary and cointegrated time series.

### 3.4 Gaussian Process

Gaussian processes are the extension of multivariate Gaussians to infinite-sized collections of real-valued variables. In particular, this extension allows us to think of Gaussian Processes as distributions not over random vectors but as distributions over functions. More formally:

**Definition 3.4.1.** A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution.

In general we are given a set \( O \) of \( N \) observations, \( O = \{(x_i, y_i)|i = 1, \ldots, N\} \), where \( x_i \) is a \( d \)-dimensional input vector \( x_i = [x_1, \ldots, x_d]^T \) and \( y_i = f(x_i) \) denotes the output or target. A Gaussian process is completely specified by a mean function \( m(\cdot) \) and a covariance function \( k(\cdot, \cdot) \):

\[
\begin{bmatrix}
  f(x_1) \\
  \vdots \\
  f(x_n)
\end{bmatrix} \sim \mathcal{N}
\begin{bmatrix}
  m(x_1) \\
  \vdots \\
  m(x_n)
\end{bmatrix},
\begin{bmatrix}
  k(x_1, x_1) & \cdots & k(x_1, x_n) \\
  \vdots & \ddots & \vdots \\
  k(x_n, x_1) & \cdots & k(x_n, x_n)
\end{bmatrix},
\]

(3.15)

denoted as:

\[
f(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot)),
\]

(3.16)

where

\[
m(x) = \mathbb{E}[f(x)],
\]

\[
k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))].
\]

(3.17)

If we have a prior knowledge about the process we can use any real-valued function \( m(\cdot) \), otherwise we set \( m(x) = 0, \forall x \in X \). The covariance function
3.4. Gaussian Process

$k(·,·)$ characterizes correlations between different points in the process and for any set of input points $x_1, \ldots, x_n \in X$, the resulting matrix:

$$K = \begin{bmatrix}
    k(x_1, x_1) & \ldots & k(x_1, x_n) \\
    \vdots & \ddots & \vdots \\
    k(x_n, x_1) & \ldots & k(x_n, x_n)
\end{bmatrix} \tag{3.18}$$

must be semidefinite positive. The matrix $K$ is also known as the Gram matrix.

One of the most used kernel functions is the **Squared Exponential Kernel**, defined as:

$$k_{SE} = \sigma^2 \exp\left(-\frac{1}{2l^2}\|x - x'\|\right), \tag{3.19}$$

where $\sigma^2$ is the precision and $l$ is the length-scale. With a Squared Exponential Kernel, for any pairs of points $x$ and $x'$:

- $f(x)$ and $f(x')$ have high covariance when $x$ and $x'$ are close in the input space (i.e. $\sigma^2 \exp\left(-\frac{1}{2l^2}\|x - x'\|\right) \to 1$ when $\|x - x'\| \to 0$);
- $f(x)$ and $f(x')$ have low covariance when $x$ and $x'$ are distant in the input space (i.e. $\sigma^2 \exp\left(-\frac{1}{2l^2}\|x - x'\|\right) \to 0$ when $\|x - x'\| \to \infty$).

Gaussian Process can be used both for regression and classification. In the following, we briefly explain them in the regression case since is the most relevant for what will be presented in what follows. Let $S = \{(x_i, y_i)|i = 1, \ldots, N\}$ be a training set of i.i.d. samples from an unknown distribution such that:

$$y_i = f(x_i) + \varepsilon_i, \quad i \in \{1, \ldots, N\}, \tag{3.20}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. If we assume additive independent identically distributed Gaussian noise $\varepsilon$ we obtain the **likelihood**, i.e. the joint distribution of $y = (y_i, \ldots, y_N)$ conditioned on $f = (f_i, \ldots, f_N)$, defined as:

$$p(y|f) = \mathcal{N}(f, \sigma^2 I_N), \tag{3.21}$$

where $I_N$ is the $N \times N$ identity matrix. Integrating over the function variables gives the **marginal likelihood**

$$p(y) = \int p(y|f)p(f)df = \mathcal{N}(0, K + \sigma^2 I_N). \tag{3.22}$$

Once we have computed the prior distribution we are interested in com-
3.4. Gaussian Process

puting the posterior predictive distribution over a test set \( f_T \), in particular, given the training \( f \) and test \( f_T \) points, we have:

\[
p(f, f_T) = \mathcal{N} \left( f, \begin{bmatrix} K_N & K_{NT} \\ K_{TN} & K_T \end{bmatrix} \right), \tag{3.23}
\]

where \( K_N = K(f, f) \), \( K_{NT} = (f, f_T) \), \( K_{TN} = K_N^T \) and \( K_T = (f_T, f_T) \).

From the assumption of i.i.d. Gaussian noise we have that:

\[
\begin{bmatrix} \varepsilon \\ \varepsilon_T \end{bmatrix} = \mathcal{N} \left( 0, \begin{bmatrix} \sigma^2 I_N & 0 \\ 0^T & \sigma^2 I_N \end{bmatrix} \right). \tag{3.24}
\]

The sums of independent Gaussian random variables is also Gaussian, thus, we obtain:

\[
p(y, y_\ast) = \mathcal{N} \left( 0, \begin{bmatrix} K_N + \sigma^2 I_N & K_{NT} \\ K_{TN} & K_T + \sigma^2 I_N \end{bmatrix} \right). \tag{3.25}
\]

Now by means of the Gaussian conditioning’s rule we get the predictive Gaussian process as follows:

\[
p(y_\ast | y) = \mathcal{N}(\mu_\ast, \Sigma_\ast), \tag{3.26}
\]

where

\[
\mu_\ast = K_{TN} [K_N + \sigma^2 I_N]^{-1} y,
\]

\[
\Sigma_\ast = K_T - K_{TN} [K_N + \sigma^2 I_N]^{-1} K_{NT} + \sigma^2 I_N. \tag{3.27}
\]

Figure 3.2a shows samples from a zero-mean GP prior with Squared Exponential covariance function (where \( l = 1 \) and precision \( \sigma = 1 \)). Figure 3.2b

**Figure 3.2**: Samples from prior and posterior of a GP with Squared Exponential covariance function.
shows samples from the posterior of a zero-mean GP where training data come from $y_i = \sin(x_i) + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 0.01)$. The black line denotes the mean of the posterior predictive distribution, and the grey region denotes the 95% confidence region based on the model’s variance estimates.

One of the main advantages of the GPs is the ability to choose the hyperparameters, i.e., the length-scale $l$ and precision $\sigma^2$ in Equation (3.19), directly from the training data. Other models require cross-validation, while with GP one can find the optimal hyperparameters by minimizing the negative log marginal likelihood $\mathcal{L}(\theta)$ with respect to hyperparameters $\theta$ as following:

$$
\mathcal{L} = -\log p(y|\theta) = \frac{1}{2} \log \det C(\theta) + \frac{1}{2} y^T C(\theta)^{-1} y + \frac{N}{2} \log(2\pi),
$$

where

$$
C = K + \sigma^2 I_N.
$$

In Chapter 6, GPs are used for the estimation of some parameters because they allow us to directly capture model uncertainty and exploit regularities in the input space.
Chapter 4

State of the art

In this section, we comment on the main contributions to our work. We divide the chapter into two sections to cover both causality detection and bid/budget optimization.

4.1 Causality detection

In the context of Online Advertising, marketing channels deeply affect each other’s performance as Internet users regularly surf from one to another. In what follows, we provide the state of the art related to intra-channel and/or cross-channel causality relations.

Hoban and Bucklin, Duffett analyze the impact of a specific channel, i.e., display and social advertising, on the marketing funnel. Instead, Rutz and Bucklin investigate intra-channel dependencies in the search advertising channel.

Hoban and Bucklin [Hoban and Bucklin, 2015] examine the effects of Internet display advertising using cookie-level data coming from a large financial services firm, taking into account only the effect of display advertising on-site visit without considering the effects of other channels. In their experiment, a small portion of people was randomly assigned to the control group. This group was served ads in precisely the same manner as the treatment group, but the ad copy was for an unrelated charity instead of the local firm. Differently from the treatment group, control group was served ads for all unrelated charity instead of the local firm. They use the comparison between the different ad responses in the two groups for estimating the effect of display advertising at different stages of the purchase funnel. The identified stages in the purchase funnel are the following: non-visitor,
4.1. Causality detection

They examine the difference in probability of site visit between treatment and control groups for each funnel stage. The results achieved show that the impact is positive and significant for three of the four stages, but it is ineffective for users who previously visited the firm’s website and did not create an account. Thereafter, using a binary logic model, they show that the effect of display advertising is positive and significant for non-visitors, authenticated users, and converted customers, while it has no discernible effect on prior visitors. Their approach assumes that the funnel stage remains constant within an observation. That is, a person cannot be an authenticated user during the first period and a converted customer in the next. However, in real applications, it is difficult to find such a situation. Furthermore, in our case tracking users at an individual level is not possible.

A deeper analysis of the social channel has been done by Duffett [Duffett, 2015]. This study is only related to Facebook advertising channel and conducted only in South Africa through a quantitative research. The authors investigate the influence of behavioral attitudes towards the social medium Facebook, among Millennials in South Africa, and determine whether various usage and demographic variables have an impact on intention-to-purchase and purchase perceptions. In particular, they analyze if Facebook advertising, usage factors (length of usage, log on duration and frequency, and profile update incidence) and demographic factors (gender, age, and ethnic orientation) have an effect on intention-to-purchase among South African Millennials and what is their impact on purchase. The research was conducted through a survey over a sample of over 3500 respondents in SA. This inquiry reveals that advertising on Facebook has a favorable impact on behavioral attitudes but at the minimal level, and found that log on duration and profile update incidence had an influence on intention-to-purchase and purchase perceptions of advertising on Facebook, whereas how Facebook was accessed, length of usage and log on frequency had no influence. Furthermore, this study found that ethnic orientation had an impact on the behavioral attitudes of black Millennials in SA, but no effect on age and gender.

Rutz and Bucklin [Rutz and Bucklin, 2011] analyze only the intra-channel dependencies, in particular, the ones from generic queries to branded queries in the search channel. The authors propose that a generic search can create awareness that the brand is relevant to the goals of the search and consequently spill over to influence subsequent branded search. This awareness
can then lead to future branded keyword searches in which the user searches the brand’s offering in more detail or perhaps goes on to complete a purchase. This hypothesis is supported by the fact that stand-alone metrics usually show that generic keywords have higher apparent costs to the advertisers than branded keywords, a generic search may create a spillover effect on subsequent branded search. This paper proposes a dynamic linear model to capture the potential spillover from generic to branded paid search, and then, using a Bayesian estimation approach, the authors apply the model to data from a paid search campaign for a major lodging chain. They run Granger causality tests in order to check the founded relation, and what they found is that generic impressions did not Granger cause any branded search activity but generic clicks did Granger-cause branded impressions and branded clicks. They also try an alternative model the standard VAR model that allows the data to reveal relationships among the different search activities by treating each as endogenous.

Cross-effects were analyzed also by Dinner et al. [Dinner et al., 2014], which consider interdependence between offline and online advertisement. The authors empirically validate the presence of cross-channel advertising effects, quantify their magnitude, and generate insights into what determines this magnitude. The analysis considers how these advertising expenditures translate directly into sales, as well as indirectly through intermediate search advertising metrics (impressions and click-through rate). They develop a multi-equation model that estimates sales in both online and offline channels as a function of online display, online paid search, and traditional advertising. The model is estimated using a procedure that accounts for the endogeneity of advertising, dynamic advertising effects, multivariate dependent variable, potential autocorrelation, and competitive advertising effects. The four dependent variables that describe the linear model of four equations are offline sales, online sales, paid search impressions and paid search click-through rate. Furthermore, for each channel, an AdStock variable is defined. AdStock is the cumulative value of a brand’s advertising at a given point in time. The authors find that cross effects exist and are important and that cross-effect elasticities are almost as high as own-effect elasticities. Online display and, in particular, search advertising is more effective than traditional advertising. This result is primarily due to strong cross effects on the offline channel.

Kireyev et al. [Kireyev et al., 2016], instead of considering only a specific channel, investigate interactions among different online channels. The au-
4.1. Causality detection

Authors develop a vector error correction model to analyze the interaction between paid search and display ads using weekly data coming from a bank that uses online ads to acquire new checking account customers. Initially, they check correlation among variables and exclude correlated variables. Then they use persistence modeling technique to capture dependencies among variables. Persistence modeling consists of multiple steps. As an initial step unit root test are used in order to test whether a time series variable is non-stationary and possesses a unit root. Granger causality tests, conducted pair-wise, are used in order to investigate causality between two variables. Then Cointegration tests are done so as to identify stationary linear combinations of non-stationary variables. After that, they build a vector error correction model. Impulse response analysis has been used to investigate the impact of display and search advertising. Impulse response functions trace the impact of a shock to one variable through other variables, thereby providing a cumulative view of all dynamic interactions that took place. The Granger causality test has been done without taking into account the presence of integrated and/or cointegrated time series.

In the following works, interdependence among channels are incorporated within attribution models. The multi-touch attribution model is one of the most important problems in digital advertising, especially when multiple media channels, such as search, display, social, mobile and video are involved. The goal of attribute modeling is to pinpoint the credit assignment of each positive user to one or more advertising touch point. The resulting user-level assignment can be aggregated along different dimensions including media channel to derive overall insights.

Shao and Li [Shao and Li, 2011] propose a new bivariate metric. One component of this metric measures the variability of the estimate, and the other measures the accuracy of classifying the positive and negative users. Then they develop a bagged logistic regression model, which they show achieves a comparable classification accuracy as a usual logistic regression, but a much more stable estimate of individual variable contributions. They also propose a probabilistic model to compute the attribution of different variables based on a combination of first and second order conditional probabilities. Lastly, they evaluate both models using the proposed bivariate metric. Their model provides some important insights that could help the advertiser to gauge the true effectiveness of each media channel and root out those gaming tactics. They estimated that, by this change alone, the advertiser could improve the overall campaign performance of 30%.
4.2. Bid/Budget optimization

Li and Kannan [Li and Kannan, 2014] develop a framework to measure customers’ consideration of online channels, visits through these channels over time, and subsequent purchases at the website to estimate the carryover and spillover effects of prior touches at both the visit and purchase stages. They propose a three-level measurement model based on individual-level path data of customer’s touches. The stages considered in the purchase decision are the followings: consideration stage, visit stage and purchase stage. Furthermore, they divide channels between customer-initiated channels and firm-initiated channels. The former consider channels in which consumers look for information on their own free will, while the latter consider channels in which the firm initiate marketing considerations. They consider the carryover and spillover effects of customer’s prior visit to a specific channel in the same channel and across other channels at both visit and purchase stage. At visit stage, they model carryover and spillover effect through their impact on the cost of visiting a channel, e.g. search costs, opportunity costs, and exposure effects. While at purchase stage they model the impact of channel’s visits through the cumulative informational stock. In particular, the cumulative informational stock increase or decrease the overall utility of making a purchase. Their empirical results show that there are significant carryover and spillover effects at both visit stage and purchase stage, and the magnitude varies across channels. Furthermore, their models help to understand the impact of different channels on conversions. Lastly, they assume to have the individual-level path data of customer’s touches that, generally, it is difficult to obtain.

The Granger Causality test has been also used in many other different fields to infer the structure among datastreams, e.g., sensor networks by Alippi et al. [Alippi et al., 2014] and economics by Calderón and Liu [Calderón and Liu, 2003].

4.2 Bid/Budget optimization

The choice of the ads to be displayed and their placement on a webpage are made through auctioning mechanisms. The role of the advertiser is to determine the daily budget and bid for each sub-campaign. In what follows, we present the state of the art related to the bid and budget optimization.

Zhang et al. [Zhang et al., 2012] study a scenario similar to the one we analyze. The authors take into account the problem of the joint bid/budget optimization of sub-campaigns in an offline fashion. The authors propose a
novel algorithm to automatically address the advertisers’ issue of campaign management. Many advertisers are dealing with a very large number of campaigns, bid keywords, and bid prices at the same time, and this is a great challenge to the optimality of their campaign management. Firstly, they perform a data study on the effectiveness of campaign budget allocation and keyword bid price setting in sponsored search and pointed out the importance of jointly optimizing them. Then, they model the problem as a constrained optimization problem, which takes the campaign budgets and the keyword bid prices as variables and finds their optimal values by maximizing the advertiser revenue, with the constraint of the total budget of the advertiser. Simulation results on the sponsored search log from a commercial search engine show that the proposed technology can effectively help advertisers improve their campaign performance under several metrics like click number, cost per click, and advertiser revenue. Their proposed method is also robust to the second-order effects caused by the advertisers’ dynamical bid changes.

Thomaidou et al. [Thomaidou et al., 2014] propose an integrated approach for automated advertising campaign creation, management, and monitoring for profit optimization under budget constraints. In particular, their algorithm automatically generates keywords based on the landing page and automatically generates text-based ads through automatic text summarization. They formulate the optimization problem with an MKCP problem and separate the optimization of the bid from that of the budget, using a genetic algorithm to optimize the budget and then applying some bidding strategies.

Geyik et al. [Geyik et al., 2014b] propose a budget allocation scheme that distributes money from the campaign top-level to the sub-campaigns according to their performance and spending capabilities. They examine both last-touch attribution, i.e. a user’s action is attributed to the last ad he sees, and multi-touch attribution, i.e. a user’s action is attributed fractionally to a subset of the ads he sees, to determine sub-campaign performance. They reduce the budget allocation to two sub-problems: spending capability calculation for a sub-campaign, and return-on-investment calculation for a sub-campaign. They also show that sub-campaign performance values, calculated via the multi-touch attribution, leads to better allocation of budgets.

Nuara et al. [Nuara et al., 2017] provide an online joint bid/daily budget optimization algorithm. They formulate the optimization problem as a
4.2. Bid/Budget optimization

combinatorial-bandit problem, where the different arms are the bid/budget pairs. In a combinatorial-bandit problem, differently, from the standard multi-armed bandit problem, there are a set of superarms, i.e., an element of the power set of arms, here corresponding to a combination of bid/budget pairs for each sub-campaign, whose element satisfy some set of combinatorial constraints, in this case, knapsack-like. Their model can be divided into three phases. In the first phase, named Estimation, the algorithm learns, from the observations of previous days, the model of the user behavior for each sub-campaign. More precisely, the model provides a probability distribution over the number of clicks \( n_j(x, y) \) as the bid \( x \) and the budget \( y \) vary and over the value per click \( v_j \). For each sub-campaign, the observations that the algorithm gets are: the actual number of clicks, actual total cost of the sub-campaign, time when the daily budget finished (if so), and the actual value per click. Subsequently, the model of each sub-campaign is updated using those observations which include all the previous days until the day before. In the second phase, named Bandit Choice, using the updated model, the algorithm chooses the values for the function \( n_j(\cdot, \cdot) \) and the parameter \( v_j \). More precisely, for each sub-campaign, the algorithm initially selects the bids and budgets that are feasible, then, according to the probability distributions of the model, the samples of the function \( n_j(\cdot, \cdot) \) are chosen. In the third phase, named Optimization, the algorithm uses the values of \( n_j(\cdot, \cdot) \) and of \( v_j \) as parameters of a variant of Multiple-Choice Knapsack problem. Bid/budget pairs to be set for the current day in each different channel are returned.

Less related works, which are still relevant for the Internet advertising optimization field to be cited here, concern the optimization of the daily budget [Xu et al., 2015; Italia et al., 2017], learning the quality score of the ads [Gatti et al., 2015b, 2012], bidding strategies in display advertising [Wang et al., 2016; Weinan et al., 2016; Zhang et al., 2014; Lee et al., 2013], video advertising [Geyik et al., 2016], and targeting [Gasparini et al., 2018]. Finally, some works deal with the attribution problem in display advertising [Geyik et al., 2014a; Kireyev et al., 2016].
4.2. Bid/Budget optimization
Chapter 5

Problem formulation

5.1 Problem

Given an advertising campaign, the advertiser has to determine the daily budget and bid for each sub-campaign in order to maximize a given metric subject to a total daily budget constraint. Generally, the effectiveness of a campaign is measured in terms of the total number of conversions. Note that, in our scenario, the advertiser set a single bid for a sub-campaign that lasts for a period of time, usually one day.

The ad of the $j$-th sub-campaign may appear or not in one of the available slots of a web page. Each slot is characterized by different probabilities of being observed and different probabilities of being clicked. To place an ad in a given page, an auction is carried out among all the advertisers that want to place an ad. Usually, the action is based on the Vickrey-Clark-Groves mechanism (see [Mas-Colell et al., 1995, Gatti et al., 2015a] for more details): the advertisers are ranked in decreasing order of bid, and each advertiser is assigned a price equal to $x_{k+1} + \rho_k$ where $x_{k+1}$ is the bid of the advertiser below him and $\rho_k$ is the probability that the ad is clicked given the probability of being observed in the $k$-th position. In Vickrey-Clark-Groves auctions specifying higher bids are displayed in higher positions. Once an ad is placed, it can be clicked or not by a user. Each click has a cost, denoted as cost-per-click (CPC). The advertiser has a constraint on the total daily expenditure, thus some solutions of bid and budget may violate this constraint. When all the total daily budget is spent, all the ads stop being shown and the advertiser cannot gather more clicks.

An advertising campaign can be more effective by combining multiple channels. Indeed, multi-channel advertising is fundamental to reach a wider population through different channels. Thus, in the bid/budget optimiza-
tion problem described above, we should take into account dependencies among sub-campaigns. Sub-campaigns of different channels are usually interdependent, e.g., display ads induce awareness, increasing the number of impressions—and, thus, also the number of conversions—of search ads. Optimally balancing the budget among interdependent sub-campaigns may lead to better performances with respect to a budget allocation that greedily invests only in the sub-campaigns that perform better alone.

Capturing interdependence among campaigns is a crucial task, indeed not considering some relations between sub-campaigns can lead to lower performances in terms of conversions. Nonetheless, manually detecting all the relevant interdependencies and setting bid and budget for all sub-campaigns could lead to a sub-optimal solution. Thus, an automatized approach, both for identifying dependencies and for choosing the optimal bid and the optimal budget, is less time consuming and provides potentially higher conversions. The main advertising tools used by companies do not allow to keep track of all the actions of each user perfectly, instead, they record aggregated user advertising quantities, i.e., click, impressions, conversions and cost. Therefore, we choose to analyze dependencies among sub-campaigns on daily aggregated data.

The goal of our work is to provide an automatic method to detect interdependence among sub-campaigns and exploit them in order to maximise the total conversions. After an initial exploration during which we collect data of all sub-campaigns, we estimate interdependence from the collected data. Then, at each day, we compute the optimal bid/budget allocation considering those dependencies.

5.2 Model

In Section 5.1 we described the problem of bid/budget optimization with sub-campaigns interdependence. To formalize the optimization problem we define the following:

- an Internet advertising campaign \( C = C_1, \ldots, C_N \), with \( N \in \mathbb{N} \), where \( C_j \) is the \( j \)-th sub-campaign;
- a bid \( x_{j,t} \in [x_j, \overline{x}_j] \) for each sub-campaign \( C_j \) at time \( t \) where \( x_j, \overline{x}_j \in \mathbb{R}^+ \) are the minimum and the maximum bid of sub-campaign \( C_j \);
- a daily budget \( y_{j,t} \in [y_j, \overline{y}_j] \) for each sub-campaign \( C_j \) at time \( t \) where \( y_j, \overline{y}_j \in \mathbb{R}^+ \) are the minimum and the maximum daily budget of sub-campaign \( C_j \);
5.2. Model

- a total daily budget $Y \in \mathbb{R}^+$, that is the maximum amount that the advertiser want to spend in one day.

We also define for each sub-campaign $C_j$ the following quantities:

- a value per click $v_j$ for each sub-campaign $C_j$;
- a click-through rate $w_j$ for each sub-campaign $C_j$;
- the expected number of impressions $n_j(x_{j,t}, y_{j,t}, X_{t-k}^{t-1}, Y_{t-k}^{t-1})$ given bid $x_{j,t}$, daily budget $y_{j,t}$, $X_{t-k}^{t-1} = (x_{t-k,0}, \ldots, x_{t-1,N})$ is the vector indicating all the bids that have been set in all the other sub-campaigns from $t - k$ to $t - 1$, and $Y_{t-k}^{t-1} = (y_{t-k,0}, \ldots, y_{t-1,N})$ is the vector indicating all the budgets allocated in all the other sub-campaigns from $t - k$ to $t - 1$.

Formally, given a campaign $C$ and a daily budget of $Y$, we aim to find, at day $t$, the value of bid $x_{j,t}$ and the value of daily budget $y_{j,t}$, for every sub-campaign $C_j$ that maximize the revenue:

$$ R_t := \sum_{j=1}^{N} f(v_j, w_j, n_j) \quad (5.1) $$

by solving the following optimization problem:

$$ \max_{x_{j,t}, y_{j,t}} R_t \quad (5.2a) $$

s.t.

$$ \sum_{j=1}^{N} y_{j,t} \leq Y \quad (5.2b) $$

$$ \underline{x}_j \leq x_{j,t} \leq \overline{x}_j \quad \forall j \ \forall t \quad (5.2c) $$

$$ \underline{y}_j \leq y_{j,t} \leq \overline{y}_j \quad \forall j \ \forall t. \quad (5.2d) $$

In principle one might use arbitrarily complex shapes for the function $f(\cdot)$, which might depend on the overall allocation as well as by the allocations selected in the past days. However, since this function is a priori unknown, its estimation should be feasible in terms of number of days required to gather the necessary data. This can be done either by resorting to a large amount of historical data, if they are available and if the environment is stationary, or by sacrificing the adherence of the model to the real process, for instance by using a simpler dependence of $f(\cdot)$ from the allocation. The consideration on how to properly model the function $f(\cdot)$ will be provided in Chapter 6.
Chapter 6

Proposed solution

6.1 Model Assumption

The model described in Chapter 5 encodes the dependencies with the other sub-campaigns’ bid and budget set in the previously $k$ days. However, with a finite number of data samples in a high-dimensional feature space with each feature having a range of possible values, a large amount of training data is required to ensure that there are several samples with each combination of values. Thus, we approximate the dependency with a function of $X_{t-k}^t$ and $Y_{t-k}^t$. In particular, we define for each sub-campaign $C_j \in \mathcal{C}$ an influence index $u_{j,t}$. The influence index represents the impact of other dependent sub-campaigns on the $j$-th sub-campaign. The expected number of impressions is rewritten as function of $u_{j,t}$, and it is defined as $n_j(x_{j,t}, y_{j,t}, u_{j,t})$ given the bid $x_{j,t}$, the daily budget $y_{j,t}$, and the number of users $u_{j,t}$ for each sub-campaign $C_j$. We formalize the interdependence among sub-campaigns through a directed graph $G := (\mathcal{C}, D)$ where nodes are present if the interdependence is present. $D$ is the adjacency matrix with $d_{ij} = 1$ if there is a directed edge from the $i$-th node to the $j$-th node. More formally, given the inter-dependency graph $\mathcal{G}$, we define the influence index $u_{j,t}$ as:

$$u_{j,t} := \sum_{i=1}^{N} \sum_{h=t-K}^{t-1} d_{i,j}(1 - v_i w_i) n_i(x_{i,h}, y_{i,h}, u_{i,h}), \quad (6.1)$$

where $K$ is the maximum lag order, meaning that users are influenced by ads at most for $K$ days. Notice that the first sub-campaign $C_1$, being influenced by no other sub-campaign, has $u_{1,t} = 0$ since the first summation in Equation (6.1) is over an empty set. The term $(1 - v_i w_i)$ removes from $n_i(x_{i,h}, y_{i,h}, u_{i,h})$ those users who converted directly when the ad of the sub-campaign $C_i$ was displayed. In the case each user might perform
multiple purchase of the same product, the model can be adapted removing
the multiplicative term \((1 - v_i w_i)\), i.e. including as potential users in other
sub-campaigns the user who converted in the past. The above definition of
\(u_{j,t}\) is based on the assumption that the increase of impressions provided by
a user coming from any sub-campaign influences the number of impressions
of \(C_j\) in the same way. While this assumption might seem simplistic, it is
necessary to keep at a pace the complexity of training the model. Indeed,
a more complex model, e.g., where there is a different influence index for
every pair of sub-campaigns, might be an option, but this would introduce
further complexity to the model, which could be tolerated only if a very
large amount of data is available. Furthermore, in Equation [6.1] we have
included only the interdependence among impressions. However different
models, e.g., including the interdependence between the clicks and the con-
versions, are straightforward extensions of what is proposed in this section.

In the optimization phase of the algorithm we require that the graph \(G\)
induced by the adjacency matrix \(D\) is a Directed Acyclic Graph.

**Definition 6.1.1.** A direct acyclic graph \(G := (C, D)\) is a graph with directed
edges in which there are no cycles. A directed acyclic graph is referred to as
DAG.

This assumption is supported by the model of marketing funnel (Section
2.2) in which the user flows from the top to the bottom, and different
advertising channels are positioned at different levels of the funnel. For
the sake of notation, we assume that the order over the indices of the sub-
campaigns is one of the topological order induced by \(G\).

**Definition 6.1.2.** A topological ordering of a directed acyclic graph \(G := (C, D)\) is an ordering of its nodes \(c_1, c_2, \ldots, c_N\) so that for every edge \((c_i, c_j)\)
we have \(i < j\).

Thus, we redefine the optimization problem Equations [5.2a]–[5.2d] as follow:

\[
\max_{x_{j,t}, y_{j,t}} \sum_{j=1}^{N} v_j w_j n_j(x_{j,t}, y_{j,t}, u_{j,t}) \quad (6.2a)
\]

\[
\text{s.t. } \sum_{j=1}^{N} y_{j,t} \leq Y \quad (6.2b)
\]

\[
x_j \leq x_{j,t} \leq \overline{x}_j \quad \forall j \quad (6.2c)
\]

\[
y_j \leq y_{j,t} \leq \overline{y}_j \quad \forall j. \quad (6.2d)
\]
We denote with \((x^*, y^*, u^*)\) the solution to the optimization problem. The optimization problem in Equations (6.2a)–(6.2d) can be solved using dynamic programming techniques, once all parameters are known.

6.2 IDL Algorithm

We design the Interdependence Detect and Learning (IDL) algorithm that, given a set of past observations, models interdependencies and returns an optimal allocation of the bid/daily budget on the sub-campaigns maximizing the revenue. The pseudo-code of the IDL algorithm is provided in Algorithm 1. It requires a dataset \(Z := \{z_{j,t}\}\) of \(\tau\) samples that provide, for each day \(t \in \{1, \ldots, \tau\}\) and each sub-campaign \(C_j \in C\) with \(j \in \{1, \ldots, N\}\), the following values:

\[z_{j,t} := (\tilde{x}_{j,t}, \tilde{y}_{j,t}, \tilde{n}_{j,t}, \tilde{c}_{j,t}, \tilde{c}_{o,j,t}, \tilde{c}_{j,t}),\]

that is a tuple, with the used bid \(\tilde{x}_{j,t}\), and daily budget \(\tilde{y}_{j,t}\), the received impressions \(\tilde{n}_{j,t}\), clicks \(\tilde{c}_{j,t}\), values of the conversions \(\tilde{c}_{o,j,t}\), and costs \(\tilde{c}_{j,t}\). We require that data collected up to day \(\tau\) to be exploratory enough to properly model the sub-campaigns interdependence. Furthermore, two confidence levels \(\alpha_{ADF} \in (0,1)\) and \(\alpha_{GC} \in (0,1)\) are needed. The IDL is divided in two separate phase:

- the first phase of the algorithm (Lines 1–8) is called Interdependence Graph Learning phase and is devoted to learning the interdependence graph of the sub-campaigns. The output of this phase is an estimate \(\hat{D}\) of the actual adjacency matrix \(D\),

- the second phase of the algorithm (Lines 9–13) is called Estimation and Optimization Phase and is devoted to the estimation of the parameters for each sub-campaign \(C_j\) (i.e. \(\hat{n}_j(\cdot, \cdot), \hat{v}_j, \hat{w}_j\)), using Gaussian Process modeling (see Section 3.4 for more details), and solving the optimization problem in Equations (6.2a)–(6.2d), once the parameters have been replaced with their estimates. The outputs of this phase are \((x^*, y^*, u^*)\), i.e., the optimal bid, daily budget, and influence index for each sub-campaign.

6.2.1 Interdependence Graph Learning Phase

The task of learning \(\hat{D}\) is obtained by resorting to the Granger Causality test (see Section 3.3.1 for more details). Given as input a tuple \((\tilde{n}_1, \ldots, \tilde{n}_N)\),
6.2. IDL Algorithm

**Algorithm 1 IDL**

**Input:** dataset $Z$, confidence $\alpha_{ADF}$, confidence $\alpha_{GC}$

**Output:** optimal bid/budget/new user allocation $(\hat{x}^*, \hat{y}^*, \hat{u}^*)$

▷ **Interdependence Learning Phase**
1: for $j \in \{1, \ldots, N\}$ do
2: $d_j \leftarrow \text{IntegrationOrder}(\tilde{n}_j, \alpha_{ADF})$
3: $d_{\text{max}} \leftarrow \max_j \{d_j\}$
4: $\hat{P} \leftarrow 0$
5: for $j \in \{1, \ldots, N\}$ do
6: for $i \in \{j + 1, \ldots, N\}$ do
7: $\hat{p}_{i,j} \leftarrow \text{GCT}(\tilde{n}_j, i, j)$
8: $\hat{D} \leftarrow \text{DAG}(\hat{P}, \alpha_{GC})$

▷ **Optimization Phase**
9: for $j \in \{1, \ldots, N\}$ do
10: $\hat{n}_j(\cdot, \cdot, \cdot) \leftarrow \text{GP}(Z, \hat{D}, j)$
11: $\hat{w}_j \leftarrow \frac{1}{t} \sum_{h=1}^{t} \hat{\omega}_{h,j}$
12: $\hat{v}_j \leftarrow \frac{1}{t} \sum_{h=1}^{t} \hat{\nu}_{h,j}$
13: $(\hat{x}^*, \hat{y}^*, \hat{u}^*) \leftarrow \text{OPT}(\hat{n}, \hat{v}, \hat{w}, \hat{D})$
14: return $(\hat{x}^*, \hat{y}^*, \hat{u}^*)$

where $\tilde{n}_j = [\tilde{n}_{j,1}, \ldots, \tilde{n}_{j,t}]^T$ is the vector of observed impressions of sub-campaign $C_j$ we estimate a Vector AutoRegressive model of order $K_{GR} + d_{\text{max}}$, where $d_{\text{max}} \in \mathbb{N}$ is the maximum integration order of the time series that we analyze and $K_{GR} \in \mathbb{N}$ is a lag order which is estimated from the data. The use of $K_{GR} + d_{\text{max}}$ lag ensures that the test statistic used in the Granger Causality test for non-stationary and cointegrated time series has the same asymptotic distribution of the stationary case, and, therefore, statistically valid conclusions can be drawn. More specifically, to compute $d_{\text{max}}$ in (Lines 1–3), we compute the order of integration $d_j$ for each time series $\tilde{n}_j$ and then we take the maximum, i.e. $d_{\text{max}} = \max_j \{d_j\}$. Given the time series $\tilde{n}_j$, the order of integration $d_j$ is obtained by iteratively applying the ADF test (see Section 3.1 for more details) with a significance level $\alpha_{ADF}$ as shown in Algorithm 2, where $\Delta^d$ operator represents the first difference operator. Once we have both the maximum order of integration $d_{\text{max}}$ and lag order $K_{GR}$, we perform the Granger Causality test on each pair of sub-campaigns (Lines 5–7). To test if the impressions of sub-campaign $C_i$ influence the impressions of sub-campaign $C_j$, we estimate the parameters
Algorithm 2 IntegrationOrder Subroutine

**Input:** observed impressions $\tilde{n}_j$, confidence $\alpha_{ADF}$

**Output:** order of integration $d_j$

1: $d_j \leftarrow 0$
2: $Stationary \leftarrow ADF(\tilde{n}_j, \alpha_{ADF})$
3: while not $Stationary$ do
4:   $d_j \leftarrow d_j + 1$
5:   $Stationary \leftarrow ADF(\Delta^d \tilde{n}_j, \alpha_{ADF})$
6: return $d_j$

$a_{jlm}$, for each $m \in 1, \ldots, K_{GR} + d_{max}$, of the model:

$$\tilde{n}_{j,t} = \sum_{l=1}^{N} \sum_{m=1}^{K_{GR}+d_{max}} a_{jlm} \tilde{n}_{l,t-m} \quad \forall h \in \{1, \ldots, N\}$$ \tag{6.3}

and we test, for each sub-campaign $C_i$, for the hypothesis:

$$\begin{cases} 
H_0 : \forall m \in \{1, \ldots, K_{GR}\} & a_{jim} = 0 \\
H_1 : \forall m \in \{1, \ldots, K_{GR}\} & a_{jim} \neq 0 
\end{cases}$$ \tag{6.4}

When there is statistical evidence for rejecting the null hypothesis in Equation (6.4) with confidence $\alpha < \alpha_{GR}$, we say that sub-campaign $C_i$ influence sub-campaign $C_j$. The result of this procedure is a matrix $\hat{P}$ containing the p-values of the pairwise tests, which is used to generate a valid estimate of the adjacency matrix $\hat{D} \in \{0,1\}^{N \times N}$. This operation is performed by $DAG(\hat{P}, \alpha_{GC})$ (Line 8) by selecting the largest subset $S$ of the p-values:

$$\hat{p}_{ij} < \frac{2\alpha_{GC}}{N(N-1)},$$

s.t. the matrix $\hat{D} := \{d_{j,i} = 1 \text{ iff } p_{j,i} \in S\}$ to correspond to a DAG. Since the Granger Causality might provide adjacency matrices $\hat{D}$ whose corresponding graph presents cycles, the DAG procedure, in such cases, generates an approximated adjacency matrix $\tilde{D}$ starting from $\hat{D}$ such that the corresponding graph is acyclic. In particular, given the cycles of the graph induced by $\hat{D}$, it iterates over all the cycles and for each cycle iteratively removes the edges with the highest p-values until that cycle is removed. This procedure ensures an overall confidence $\alpha_{GC}$ on the Granger Causality test, thanks to the Bonferroni correction for multiple tests, and it avoids the presence of cycles.
6.2.2 Estimation and Optimization Phase

The second phase of the IDL algorithm exploits predictive models to estimate unknown functions and quantities in the optimization defined in Equations (6.2a)–(6.2d), and solves it in a dynamic programming fashion with an *ad hoc* procedure.

**Estimation**

In the estimation process we resort to Gaussian Processes (GPs) to compute, for each sub-campaign $C_j$ the function $\hat{n}_j(x, y, u)$ estimating the expected number of impressions $n_j(x, y, u)$, given the chosen bid $x$, the allocated budget $y$, and the influence index $u$ generated by the sub-campaigns influencing sub-campaign $C_j$ (Line 10). GPs allow us to directly capture the model uncertainty and add prior knowledge and specifications about the shape of the model by selecting different kernel functions. Furthermore, they avoid to estimate the value of $n_j(\cdot, \cdot, \cdot)$ for every value of bid, budget and influence index. Indeed, GPs exploit regularities in the input space and correlation between different points of the input space.

Estimating the number of impressions $n_j(\cdot, \cdot, \cdot)$ through a 3-dimensional GP would require a large number of samples, thus, we resort to a factored model. The dependency of $n_j(x, y, u)$ from bid $x$, budget $y$ and influence index $u$ is modelled in a factorized fashion by two 2-dimensional GPs as follows:

$$n_j(x, y, u) := n_{j}^{sat}(x, u) \min\left\{1, \frac{y}{c_{j}^{sat}(x, u)}\right\}, \quad (6.5)$$

where, given a bid $x$ and an influence index $u$, the two GPs employed for a given sub-campaign $C_j$ are:

- the maximum number of impressions $n_{j}^{sat} : \mathcal{X}_j \times \mathcal{U}_j \rightarrow \mathbb{R}^+$ that can be obtained with a given bid $x$ without any budget constraint (or equivalently if we let $y \rightarrow +\infty$);

- the maximum cost incurred $c_{j}^{sat} : \mathcal{X}_j \times \mathcal{U}_j \rightarrow \mathbb{R}^+$ that can be obtained with a given bid $x$ without any budget constraint (or equivalently if we let $y \rightarrow +\infty$);

where the bid space is defined as $\mathcal{X}_j := [\underline{x}_j, \overline{x}_j]$ with $\underline{x}_j := \min_t \underline{x}_{j,t}$ and $\overline{x}_j := \max_t \overline{x}_{j,t}$, and the influence index space is $\mathcal{U}_j := \mathbb{R}^+$.

The basic idea behind the factorized model defined in Equation (6.5) is that, given a bid $x$ and an influence index $u$, the number of impressions
increases linearly in the budget until the maximum number of impressions is achieved. The coefficient of this dependence is the average cost per impression $c_{j \text{sat}}^\text{sat} (x, u)$.

Both the maximum number of impressions $n_{j \text{sat}}^\text{sat}(x, u)$ and the maximum cost incurred $c_{j \text{sat}}^\text{sat}(x, u)$ are modelled through a 2-dimensional GP. The maximum number of impressions of a given sub-campaign $C_j$ over the bid space $X_j$ and the influence index space $U_j$ is defined as:

$$n_{j \text{sat}}^\text{sat} (Z) := \mathcal{GP}(m(Z), k(Z, \cdot)) \quad \forall Z \in X_j \times U_j,$$

where $m : Z \rightarrow 0$ is the mean of the GP, and $k : Z \times Z \rightarrow \mathbb{R}$ is the covariance function. The maximum cost incurred is modelled similarly as follows:

$$c_{j \text{sat}}^\text{sat} (Z) := \mathcal{GP}(m(Z), k(Z, \cdot)) \quad \forall Z \in X_j \times U_j.$$

After the initial exploration of $\tau$ days, in the following days $t = \tau, \tau + 1, \ldots$ the Gaussian Processes are updated by making use of:

- a vector of the bids $\tilde{x}_{j,t-1} = [\tilde{x}_{j,1}, \ldots, \tilde{x}_{j,t-1}]^T$;
- a vector of influence indexes $\tilde{u}_{j,t-1} = [\tilde{u}_{j,1}, \ldots, \tilde{u}_{j,t-1}]^T$;
- a vector of maximum number of impressions $\tilde{n}_{j,t-1}^\text{sat} = [\tilde{n}_{j,1}^\text{sat}, \ldots, \tilde{n}_{j,t-1}^\text{sat}]^T$;
- a vector of maximum costs $\tilde{c}_{j,t-1}^\text{sat} = [\tilde{c}_{j,1}^\text{sat}, \ldots, \tilde{c}_{j,t-1}^\text{sat}]^T$;

and by means of the Gaussian conditioning’s rule we get the mean $\mu_{j,t-1}^\text{imp}$ and the variance $\sigma_{j,t-1}^2$ of the maximum number of impressions defined as:

$$\mu_{j,t-1}^\text{imp} (Z) = K(Z, \tilde{Z}) \Phi^{-1} \tilde{n}_{j,t-1}^\text{sat},$$

$$\sigma_{j,t-1}^2 (Z) = K(Z, Z) - \Psi,$$

where

$$\Psi = K(Z, \tilde{Z}) \left[ K(\tilde{Z}, \tilde{Z}) + \sigma_{\tilde{Z}}^2 I \right]^{-1} K(\tilde{Z}, \tilde{Z})^T,$$

$$\tilde{Z} = [\tilde{x}_{j,t-1}, \tilde{u}_{j,t-1}],$$

and $K$ is the kernel matrix. The mean $\mu_{j,t-1}^\text{cost}$ and the variance $\sigma_{j,t-1}^2$ of the maximum costs are estimated in the same way. The parameters of the GPs, i.e., $\mu_{j,t-1}^\text{imp}, \sigma_{j,t-1}^\text{imp}^2, \mu_{j,t-1}^\text{cost}, \sigma_{j,t-1}^\text{cost}^2$, are computed by means of the Gaussian conditioning’s rule as described in Section 3.4.

The click-through rate $w_j$ and the value per click $v_j$ are estimated using the historical data. In particular, given the past values of the value per click
6.2. IDL Algorithm

\( \tilde{v}_{j,k} \) and of the click-through rate \( \tilde{w}_{j,k} \) with \( k < t \), and assuming that both \( w_j \) and \( v_j \) are normally distributed, we apply the Central Limit Theorem. The Central Limit Theorem states that the sampling distribution of the mean can be approximated by a normal distribution with mean \( \mu \) and standard deviation \( \frac{\sigma}{\sqrt{t-1}} \). In our case we have:

\[
\mu_{j,t-1}^w := \frac{1}{t-1} \sum_{k=1}^{t-1} \tilde{w}_{j,k},
\]

\[
\sigma_{j,t-1}^w := \frac{1}{(t-1)(t-2)} \sum_{k=1}^{t-1} (\tilde{w}_{j,k} - \mu_{j,t-1}^w)^2,
\]

and

\[
\mu_{j,t-1}^v := \frac{1}{t-1} \sum_{k=1}^{t-1} \tilde{v}_{j,k},
\]

\[
\sigma_{j,t-1}^v := \frac{1}{(t-1)(t-2)} \sum_{k=1}^{t-1} (\tilde{v}_{j,k} - \mu_{j,t-1}^v)^2,
\]

thus, the sample mean of the click-through rate \( w_j \) and the value per click \( v_j \) can be approximated by:

\[
\hat{w}_{j,t} \sim N(\mu_{j,t-1}^w, \sigma_{j,t-1}^w),
\]

\[
\hat{v}_{j,t} \sim N(\mu_{j,t-1}^v, \sigma_{j,t-1}^v).
\]

Finally, the influence index \( u_j \) is estimated as follows:

\[
\hat{u}_j := \frac{1}{K_{GR}} \sum_{i=1}^{j-1} \sum_{h=t-1}^{K_{GR}} \tilde{d}_{i,j} \hat{n}_j(x_{i,h}, y_{i,h}, \hat{u}_{i,h}),
\]

where we use \( K_{GR} \), obtained from the Granger Causality test, as an estimate of the actual lag \( K \).

Gaussian Processes provide probability distributions over functions \( n_j^{sat}(\cdot, \cdot) \) and \( c_j^{sat}(\cdot, \cdot) \), and \( \hat{w}_{j,t} \) and \( \hat{v}_{j,t} \) provide probability distributions over the sample mean of \( w_j \) and \( v_j \). In what follows we simply use the estimated means \( \mu_{j,t-1}^{imp}(X) \), \( \mu_{j,t-1}^{cont}(Z) \), \( \mu_{j, t-1}^w \), and \( \mu_{j, t-1}^v \), since the IDL algorithm use an offline approach.

In the optimization phase the metric used to find the optimal solution is the total revenue, more precisely the number of conversions. Given a
bid $x$, a budget $y$ and an influence index $u$, the number of conversions of a campaign $C_j$ can be estimated as follows:

$$
c_{o_j} := \begin{cases} 
    \hat{v}_j \hat{w}_j \hat{n}_j(x, y, u) & \text{if } y_j \leq y \leq \bar{y}_j, \\
    0 & \text{if } y < y_j \lor y > \bar{y}_j,
\end{cases}
$$

where $y_j$ and $\bar{y}_j$ are respectively the minimum and the maximum budget for sub-campaign $C_j$.

Furthermore, in the optimization phase we decouple the bid optimization from the budget optimization. In particular, given a budget $y$ and an influence index $u$, we find the optimal bid $\chi(y, u)$ as the bid that maximizes the number of impressions:

$$
\chi(y, u) := \arg \max_x \hat{n}_j(x, y, u).
$$

Thus, we can rewrite the number of conversions of sub-campaign $C_j$ as:

$$
c_{o_j} := \begin{cases} 
    \hat{v}_j \hat{w}_j \hat{n}_j(\chi(y, u), y, u) & \text{if } y_j \leq y \leq \bar{y}_j, \\
    0 & \text{if } y < y_j \lor y > \bar{y}_j.
\end{cases}
$$

### Optimization

The optimization problem is solved in a dynamic programming way, assuming that sub-campaigns $C_1, \ldots, C_N$ are topologically ordered. This requirement is due to the fact that the budget optimization of sub-campaign $C_j$ needs to know the impressions generated by dependent sub-campaigns with a given budget allocation. The crucial point of the proposed solution is that at the $j$-th iteration is not possible to find an optimal budget allocation for a given budget $b$ if the sub-campaigns $C_1, \ldots, C_j$ influence following sub-campaigns; thus we need to store multiple solutions. The OPT algorithm (Algorithm 3) uses three structures $\Pi$, $L$ and $M$ defined as follows:

- $\Pi_{j,i,h}$ is a vector $[\pi_{0,i}, \ldots, \pi_{j,i}]$ that specifies a partial budget allocation with cumulative budget of $b_i$ among the sub-campaigns $C_1, \ldots, C_j$, where $\sum_{n=0}^{j} \pi_{n,i} = b_i$ and $\pi_{n,i} = 0 \forall n \in \{j + 1, \ldots, N\}$;
- $L_{j,i,h}$ is the revenue provided by the partial budget allocation $\Pi_{j,i,h}$;
- $M_{j,i,h}$ is a vector that specifies the value of the influence index of the sub-campaigns $C_{j+1}, \ldots, C_N$ provided by the sub-campaigns $C_1, \ldots, C_j$ with the partial allocation $\Pi_{j,i,h}$. 
Algorithm 3 OPT(\(\hat{n}, \hat{v}, \hat{w}, \hat{D}\))

**Input:** interdependence matrix \(\hat{D}\), trained models \(\hat{n}_j(\cdot, \cdot, \cdot)\), \(\hat{w}_j\), \(\hat{v}_j\), discretization of the total budget \(\{b_1, \ldots, b_N\}\)

**Output:** optimal bid/budget allocation \((\hat{x}^*, \hat{y}^*, \hat{u}^*)\)

1: for \(i \in \{1, \ldots, B\}\) do
2: \(\Pi_{1,i,1} \leftarrow (b_i, 0_{N-1})\)
3: \(L_{1,i,1} \leftarrow \hat{v}_1 \hat{w}_1 \hat{n}_1(\chi_1, b_i, 0)\)
4: \(M_{1,i,1} \leftarrow (1 - \hat{v}_1 \hat{w}_1)n_1(\chi_1, b_i, 0) \hat{d}_1\)
5: for \(j \in \{2, \ldots, N\}\) do
6: for \(i \in \{1, \ldots, B\}\) do
7: \(c \leftarrow 1\)
8: for \(k \in \{1, \ldots, i\}\) do
9: \(m = |\{\Pi_{j-1,k,h}\}|\)
10: for \(h \in \{1, \ldots, m\}\) do
11: \(l \leftarrow \hat{n}_j(\chi_j, b_j - b_k, M_{j-1,k,h}(j))\)
12: \(\Pi_c \leftarrow (0_{j-1}(b_j - b_k) 0_{N-j}) + \Pi_{j-1,k,h}\)
13: \(L_c \leftarrow \hat{v}_j \hat{w}_j l + L_{j-1,k,h}\)
14: \(M_c \leftarrow (1 - \hat{v}_j \hat{w}_j) l \hat{d}_j + M_{j-1,k,h}\)
15: \(c \leftarrow c + 1\)
16: \(\Pi_{j,i,c}, L_{j,i,c}, M_{j,i,c} \leftarrow \text{Discard}(\Pi_{j,i,c}, L_{j,i,c}, M_{j,i,c})\)
17: for \(j \in \{1, \ldots, N\}\) do
18: \(\hat{y}_j^* \leftarrow \text{max}, \Pi_{1,i,1}(j)\)
19: \(\hat{u}_j^* \text{ computed as in Equation } 6.13, j ≠ \hat{y}_j^*\)
20: \(\hat{x}_j^* \leftarrow \text{arg max}_x \hat{n}_j(x, \hat{y}_j^*, \hat{u}_j^*)\)
21: return \((\hat{x}^*, \hat{y}^*, \hat{u}^*)\)

The third index \(h\) in the structures mentioned above is necessary since the algorithm may need to store multiple partial budget allocations for each \(j\) and \(i\).

At first, the algorithm initializes the values of the structures for \(j = 1\) (Lines \([1, 4]\), corresponding to the allocations of the partial budget to sub-campaign \(C_1\). For each budget \(b_i\), we allocate it to \(C_1\), formally, \(\Pi_{1,i,1} = (b_i, 0_{N-1})\), where \(0_{N-1}\) denotes a null vector of size \(N-1\). The sub-campaign \(C_1\), being the first in the topological order induced by \(\hat{D}\), is not subject to any interdependence from other sub-campaigns. Therefore, the computation of the revenue \(\{L_{1,i,1}\}_i\) and the influence index vector \(\{M_{1,i,1}\}_i\) is performed using the previously estimated models. The vector \(M_{1,i,1}\) is computed as \(M_{1,i,1} = \hat{n}_1(\chi_1, b_i, 0) \hat{d}_1\), where \(\hat{d}_i\) is the \(i\)-th row of the adjacency matrix \(\hat{D}\). This means that \(M_{1,i,1}(j)\), i.e., the \(j\)-th element of \(M_{1,i,1}\) is equal to \(\hat{n}_1(\chi_1, b_i, 0)\) if the sub-campaign \(C_1\) influences the sub-campaign \(C_j\) and zero.
For all the $j \in \{2, \ldots, N\}$, the algorithm computes the elements of the three structures $\Phi$, $L$, and $M$ using the values previously computed at the $j-1$-th step, in a dynamic programming fashion. For each budget $b_i$ and for each daily budget $b_k \leq b_i$, we compute the revenue and the influence index provided by the allocation of a daily budget $b_i - b_k$ to the sub-campaign $C_j$ and the remaining daily budget $b_k$ to the sub-campaigns $C_1, \ldots, C_{j-1}$. We do this by enumerating all partial allocations $\Pi_{j-1,k,1}, \Pi_{j-1,k,2}, \ldots$ of the first $j-1$ sub-campaigns, and then we allocate a daily budget $b_i - b_k$ to sub-campaign $C_j$ and, finally, we evaluate the total revenue $\bar{L}_c$ and the vector of the influence indices $\bar{M}_c$ provided by the partial allocations obtained, denoted with $\bar{\Pi}_c$ (Lines 9–15). After that, the algorithm discards some solutions (Lines 16–17).

**Algorithm 4 DiscardParetoDominatedSolutions**

```plaintext
1: for $h \in \{1, \ldots, |\{\bar{L}_c\}|\} \ do$
2: if $\not\exists k \mid \bar{L}_h < \bar{L}_k \land \forall p \in \{j+1, \ldots, N\} |\bar{M}_h(p) < \bar{M}_k(p) \ then$
3: $\Pi_{j,i,c} \leftarrow \bar{\Pi}_h$
4: $L_{j,i,c} \leftarrow \bar{L}_h$
5: $M_{j,i,c} \leftarrow \bar{M}_h$
6: $c \leftarrow c + 1$
```

**Algorithm 5 DiscardSolutions**

```plaintext
1: for $h \in \{1, \ldots, |\{\bar{L}_c\}|\} \ do$
2: if $\not\exists k \mid \bar{L}_h < \bar{L}_k \land \forall p \in \{j+1, \ldots, N\} |\bar{M}_h(p) = \bar{M}_k(p) \ then$
3: $\Pi_{j,i,c} \leftarrow \bar{\Pi}_h$
4: $L_{j,i,c} \leftarrow \bar{L}_h$
5: $M_{j,i,c} \leftarrow \bar{M}_h$
6: $c \leftarrow c + 1$
```

Different approaches for the **Discard** subroutine are considered depending on the presence or absence of monotonicity in the number of impressions with respect to the influence index. In the first case, we enumerate all the partial allocations $\Pi_{j,k,1}, \Pi_{j,k,2}, \ldots$ of the first $j$ sub-campaigns, and then we discard all the solutions that are Pareto dominated as shown in (Algorithm 4), where the optimality criteria are the revenue and the influence indices of sub-campaigns $C_{j+1}, \ldots, C_N$. For instance, given two partial budget allocations $\Pi_{j,i,h_1}$ and $\Pi_{j,i,h_2}$ the former solution is discarded if the latter has higher revenue and number of impressions. On the other hand, if the
former solution has a larger revenue and a smaller number of impressions and the latter vice versa, it is not possible to decide which one is the optimal before evaluating their influence on the sub-campaigns $C_{j+1}, \ldots, C_N$ and therefore we need to store both. In the second case, we discard solutions only when there are solutions with the same influence indices and different revenues as shown in (Algorithm 5). This approach does not make any assumption on the impact of the influence index on the number of impressions, indeed it is a more general approach. The main drawback is that we need to store more solutions than the first approach.

Finally, the algorithm returns the optimal allocations (Lines 18–22): the optimal budgets $\hat{y}_j^\ast$ are the elements of $\max_i \Pi_{N,i}(j)$; the optimal influence indices $\hat{u}_j^\ast$ are computed using Equation (6.13); the optimal bids $\hat{x}_j^\ast$ are computed using the number of impression models $\hat{n}_j(\cdot, \cdot, \cdot)$.

The complexity of the IDL is $O\left(\sum B^\Sigma_i \tilde{d}_{ij} + 2\right) \leq O\left(NB^2 B_{\max} \Sigma_i \tilde{d}_{ij} + 2\right)$ and strictly depend on the maximum indegree of the interdependence graph corresponding to $\tilde{D}$. Notice that only the pairs of sub-campaigns with the most significant interdependence is a crucial issue from a computational point of view since, taking into account all the possible pairs of sub-campaigns, the complexity is bounded by $NB^2 \frac{B^{N+1} - 1}{B - 1}$, which is intractable when $N$ is large as it happens in real-world applications.

### 6.3 Theoretical Properties of the IDL Algorithm

We analyze the properties of our problem and those of the IDL algorithm. Initially, we analyze the suboptimality of any algorithm ignoring the sub-campaigns inter-dependencies w.r.t. our algorithm, i.e., when the learner uses an adjacency matrix $\tilde{D} = 0$, and the real one $D$ is non-null. The following theorem shows that ignoring the sub-campaigns inter-dependencies might be arbitrarily suboptimal.

**Theorem 1.** Given the problem of optimizing an advertising campaign $C$, employing a model $\hat{n}_j(x, y)$ for the number of impressions that ignores the sub-campaigns interdependence may result in an arbitrary large loss in terms of revenue, defined as:

$$R_t = \sum_{j=1}^{N} v_j w_j n_j(x_{j,t}, y_{j,t}, u_{j,t}).$$  \hspace{1cm} (6.15)
Proof. Consider a campaign $C$ with $N = 2$ sub-campaigns, adjacency matrix $D = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, click-through rates $w_1 = w_2 = 1$, values per click $v_1 = 0$ and $v_2 = 1$, lag $K = 1$, and impression functions identified by two GPs having the following mean value:

$$n_1(x, t, u) = H y_1, t,$$
$$n_2(x, t, u) = \Upsilon u_2, t y_2 = \Upsilon n_1(x_{t-1}, y_{t-1}, u_{t-1}) y_{2, t},$$

where $\Upsilon \in [0, 1]$, $H \in \mathbb{R}^+$. Moreover, assume to have a total budget of $Y$ and to sample the solution space uniformly during training. Asymptotically (when we have an infinite number of samples or $\tau \to \infty$), a model that provides a stationary allocation and knows the sub-campaign interdependencies would compute the revenue:

$$R_t = \sum_{j=1}^{N} v_j w_j n_j(x, t, y, u) = \Upsilon n_1(x, y_1, u) y_2,$$

where we drop the temporal indices since the proposed solution is stationary and $y_1 + y_2 = Y$. The budget allocation maximizing the revenue is $y_2 = \frac{Y}{2}$, with a revenue $R_t = \Upsilon H Y^2$.

On the other hand, a model ignoring the sub-campaigns interdependence estimates the impressions of the sub-campaign $C_2$ as:

$$\tilde{n}_2(x, t, u) = \int_{0}^{H Y} n_2(x, t, y, u) du = \int_{0}^{H Y} \Upsilon u y_2, t du = \frac{\Upsilon H^2 Y^2}{2} y_2, t,$$

and the revenue one maximizes becomes:

$$\tilde{R}_t = \sum_{j=1}^{N} v_j w_j n_j(x, t, y, u) = \tilde{n}_2(x, t, y_2, t) = \frac{\Upsilon H^2 Y^2}{2} y_2.$$

The budget allocation maximising $\tilde{R}_t$ is $y_2 = Y$, with a real revenue $R_t = 0$. Hence, the difference of revenue of the two algorithms is $\Upsilon H Y^2 - 0$, which is arbitrarily large as $H$ goes to $\infty$. 

When the model is flexible enough to model the actual process properly, we can bound its error, formally, defined as follows:
Definition 6.3.1. Given a dataset $Z$, the total (estimation) error is:

$$E_\tau := \sum_{j=1}^{N} \left[ v_j w_j n_j (x_j^*, y_j^*, u_j^*) - \hat{v}_j \hat{w}_j \hat{n}_j (\hat{x}_j^*, \hat{y}_j^*, \hat{u}_j^*) \right],$$

where the tuples $(\hat{x}_j^*, \hat{y}_j^*, \hat{u}_j^*)$ are elements of the (stationary) output $(\hat{x}^*, \hat{y}^*, \hat{u}^*)$ of the IDL algorithm using the estimates of the parameters, and $(x_j^*, y_j^*, u_j^*)$ are elements of the (stationary) output of the IDL algorithm using the real parameters.

We can show the following:

Theorem 2. When the expected number of impressions $n_j (\cdot, \cdot, \cdot)$ of every subcampaign $C_j$ is distributed as a Gaussian Process, the total error between the real revenue and the estimated one using the output of the IDL algorithm is upper bounded, with a probability of at least $1 - \delta$, as follows:

$$E_\tau \leq 2N v^{(max)} \sqrt{\frac{1}{2\tau} \log \frac{6N}{\delta}} + N v^{(max)} \sigma^{(max)} \sqrt{2 \log \frac{3N}{2\delta}},$$

where $n^{(max)} := \max_j \max_{(x,y,u)} n_j (x, y, u)$ is the maximum number of estimated expected impressions over all the sub-campaigns, $\sigma^{(max)} := \max_j \max_{(x,y,u)} \sigma_j (x, y, u)$ is the maximum estimated standard deviation, and $v^{(max)}$ is the maximum value per click.

We remark that Rasmussen and Williams [2006] show that, in a generic GP, $\sigma^{(max)} \to 0$ as $\tau \to \infty$. Therefore, the total error $E_\tau$ decreases as the number of samples $\tau$ in the training set $Z$ increases.

Proof. Since estimates for click-through rate $\hat{w}_j$ and value per click $\hat{v}_j$ are sum of i.i.d. random variables with finite support $[0, 1]$ and $[0, v^{(max)}]$, respectively, we can apply the Hoeffding’s bound [Hoeffding 1963] and state that, with a probability of at least $1 - \delta$, we have:

$$w_j - \hat{w}_j \leq \sqrt{\frac{1}{2\tau} \log \frac{1}{\delta}}, \quad \text{(6.16)}$$

$$v_j - \hat{v}_j \leq v^{(max)} \sqrt{\frac{1}{2\tau} \log \frac{1}{\delta}}, \quad \text{(6.17)}$$

where $\tau$ is the number of samples we use to compute the estimates. By assumption, the number of impressions are generated by a GP and we have
that for each input \((x, y, u)\) in the GP domain we have:

\[
\frac{n_j(x, y, u) - \hat{n}_j(x, y, u)}{\hat{\sigma}_{j, \tau}(x, y, u)} \sim \mathcal{N}(0, 1),
\]

where \(\hat{\sigma}_{j, \tau}(x, y, u)\) is the standard deviation computed by the GP at the point \((x, y, u)\) by relying on \(\tau\) samples in the training set \(Z\). This implies that, with a probability of at least \(1 - \delta\), it holds:

\[
n_j(x, y, u) \leq \hat{n}_j(x, y, u) + \hat{\sigma}_{j, \tau}(x, y, u) \sqrt{2 \log \frac{1}{2\delta}}, \tag{6.18}
\]

where the last inequality is due to the fact that, for a Gaussian random variable \(X \sim \mathcal{N}(0, 1)\), it holds \(\forall \ x > 0, \ P(X > x) \leq \frac{1}{2} e^{-\frac{x^2}{2}}\).

Focus on \(E_{\tau}\). Sum and subtract from it the following quantities: \(\hat{\omega}_j v_j n_j(x_j^*, y_j^*, u_j^*)\), \(\hat{\omega}_j \hat{v}_j \hat{n}_j(x_j^*, y_j^*, u_j^*)\), and \(\hat{\omega}_j \hat{v}_j \hat{n}_j(\hat{x}_j^*, \hat{y}_j^*, \hat{u}_j^*)\) for each \(j \in \{1, \ldots, N\}\). The total error can be decomposed as:

\[
E_{\tau} = \sum_{j=1}^{N} \left( (w_j - \hat{w}_j) v_j n_j(x_j^*, y_j^*, u_j^*) + \hat{w}_j (v_j - \hat{v}_j) n_j(x_j^*, y_j^*, u_j^*) + \hat{w}_j \hat{v}_j (n_j(x_j^*, y_j^*, u_j^*) - \hat{n}_j(x_j^*, y_j^*, u_j^*)) \right) + \sum_{j=1}^{N} \hat{w}_j \hat{v}_j \hat{n}_j(\hat{x}_j^*, \hat{y}_j^*, \hat{u}_j^*). \]

If we focus on \(E_{1j}\), with probability at least \(1 - \delta\), it holds:

\[
E_{1j} = (w_j - \hat{w}_j) v_j n_j(x_j^*, y_j^*, u_j^*) \\
\leq v^{(\max)} \sqrt{\frac{1}{2\tau} \log \frac{2}{\delta}} \left( \hat{n}_j(x_j^*, y_j^*, u_j^*) + \hat{\sigma}_{j, \tau}(x_j^*, y_j^*, u_j^*) \sqrt{2 \log \frac{1}{2\delta}} \right) \\
\leq v^{(\max)} \sqrt{\frac{2}{\tau} \log \frac{2}{\delta}} \left( \hat{n}_j^{(\max)} + \hat{\sigma}_{j, \tau}^{(\max)} \sqrt{2 \log \frac{1}{2\delta}} \right),
\]

by relying on the inequalities in Equations (6.16) and (6.18), subsequently using a union bound over these two events, then defining \(\hat{n}_j^{(\max)} := \max_{(x,y,u)} \hat{n}_j(x, y, u)\) and \(\hat{\sigma}_{j, \tau}^{(\max)} := \max_{(x,y,u)} \hat{\sigma}_{j, \tau}(x, y, u)\), and finally since \(v_j \leq v^{(\max)}\).

Similarly, we derive the following bound holding with probability at least \(1 - \delta\) for \(E_{2j}\):

\[
E_{2j} \leq v^{(\max)} \sqrt{\frac{1}{2\tau} \log \frac{2}{\delta}} \left( \hat{n}_j^{(\max)} + \hat{\sigma}_{j, \tau}^{(\max)} \sqrt{2 \log \frac{1}{2\delta}} \right),
\]
by relying on the inequality in Equation 6.17, and the fact that \( w_j \leq 1 \).

Let us focus on \( E_{3j} \). Using the inequality in Equation 6.18, we have that with probability at least \( 1 - \delta \):

\[
E_{3j} = \hat{w}_j \hat{v}_j \left[ n_j(x_j^*, y_j^*, u_j^*) - \hat{n}_j(x_j^*, y_j^*, u_j^*) \right] \\
\leq v^{(\text{max})} \hat{\sigma}_{j,\tau}(x_j^*, y_j^*, u_j^*) \sqrt{2 \log \frac{1}{2\delta}} \\
\leq v^{(\text{max})} \hat{\sigma}_{j,\tau}^{(\text{max})} \sqrt{2 \log \frac{1}{2\delta}},
\]

where we used \( \hat{w}_j \leq 1 \), and \( \hat{v}_j \leq v^{(\text{max})} \).

Finally, let us focus on \( E_4 \). The vector \((\hat{x}^*, \hat{y}^*, \hat{u}^*)\) is the optimal solution of the optimization problem stated in Equations 6.2a–6.2d. Therefore, by definition, we have that for each \((x, y, u)\) satisfying the constraints in Equations 6.2a–6.2d, the following holds:

\[
\sum_{j=1}^{N} \hat{w}_j \hat{v}_j \hat{n}_j(x_j, y_j, u_j) - \sum_{j=1}^{N} \hat{w}_j \hat{v}_j \hat{n}_j(\hat{x}_j^*, \hat{y}_j^*, \hat{u}_j^*) \leq 0,
\]

which holds also for \((x^*, y^*, u^*)\) and, therefore, \( E_4 \) is negative.

Recalling that \( \hat{n}^{(\text{max})} := \max_j \hat{n}_j^{(\text{max})} \) and \( \hat{\sigma}_\tau^{(\text{max})} := \max_j \hat{\sigma}_{j,\tau}^{(\text{max})} \), it holds, with probability at least \( 1 - \delta \):

\[
E_\tau = \sum_{j=1}^{N} (E_{1j} + E_{2j} + E_{3j}) + E_4 \\
\leq 2 N v^{(\text{max})} \left( \sqrt{\frac{1}{2\tau} \log \frac{6N}{\delta}} (\hat{n}^{(\text{max})} + \hat{\sigma}_\tau^{(\text{max})}) \sqrt{2 \log \frac{3N}{2\delta}} \right) + \\
N v^{(\text{max})} \hat{\sigma}_\tau^{(\text{max})} \sqrt{2 \log \frac{3N}{2\delta}},
\]

which concludes the proof.

Our analysis has, so far, focused on the static properties of our problem. However, the scenario we are studying is a dynamical system due to the potentially delayed effects induced by the sub-campaigns interdependence. Therefore, it is crucial to show that, whenever a stationary allocation is used, the dynamics always reach a steady state in finite time and how their length is upper bounded. We can show the following:

**Theorem 3.** Using the stationary allocation \((\hat{x}^*, \hat{y}^*, \hat{u}^*)\) we reach a steady state after at most \( K \Gamma + 1 \) days, where \( K \) is the maximum lag of the influence index \( u_{j,t} \) and \( \Gamma \) is the length of the longest path of the graph \( \mathcal{G} \).
6.3. Theoretical Properties of the IDL Algorithm

The above theorem states that the more complex the process (e.g., presenting a cascade of interdependences), the more we have to wait to completely remove the effects of a suboptimal allocation.

Proof. Consider an adjacency matrix $\hat{D}$ and assume that the lag $K$ is the same for all the performance indices $u_{j,t}$. Moreover, consider the sub-campaigns $C_{i_1}, \ldots, C_{i_r}$ that make up the longest path on the graph $G$. To achieve the steady state revenue provided by the allocation $(\hat{x}_{i_1}^*, \hat{y}_{i_1}^*, \hat{u}_{i_1}^*)$, an algorithm needs that all the incoming-neighbour sub-campaigns have reached the optimal allocation for $K$ consecutive days, so as to provide exactly $\hat{u}_{i_r}^*$ as influence index. In particular, this also happens for the sub-campaign $C_{i_r}$.

By induction, this reasoning can be applied for all nodes $C_{i_k}$ in the longest path up to $C_{i_2}$. Therefore, every time we traverse a node, we require $K$ days to conclude the transient for that node, for a total of $K \Gamma$ days. Instead, the node $C_{i_1}$, being at the beginning of the longest path, has no incoming neighbours and, therefore, the allocation prescribed by the optimal solution is $(\hat{x}_{i_1}^*, \hat{y}_{i_1}^*, 0)$. This implies that the allocation is achieved on the same day that the stationary allocation is used, leading to a total number of $K \Gamma + 1$ days to reach the desired allocation on the longest path $C_{i_1}, \ldots, C_{i_r}$.

The same reasoning can be replicated on any other path, but since their length is shorter or equal to the longest one, the maximum number of days required to reach the allocation $(\hat{x}^*, \hat{y}^*, \hat{u}^*)$ takes no longer than $K \Gamma + 1$ days. □
Chapter 7

Experimental Evaluation

We experimentally evaluate the IDL algorithm in a real-world setting. We compare the revenue $R_t$ produced by IDL and AdComB-TS, the algorithm proposed by Nuara et al. neglecting any sub-campaign interdependence. However, due to the impossibility to directly compare the performance of the two algorithms online (e.g., by using an A/B testing system) we also evaluate the IDL algorithm in synthetic settings, generated by using a realistic simulator.

7.1 Synthetic Settings

We evaluate the performance of the IDL algorithm in three synthetic settings, generated by a realistic simulator, comparing the revenue $R_t$ produced by the following algorithms: IDL, DA-IDL (Dependency Aware-IDL), a variation of the IDL algorithm a priori knowing the dependency matrix $D$, and AdComB-Mean, an off-line version of the algorithm proposed by Nuara et al.. In the IDL algorithm we use a significance level both for the ADF test and for the Granger Causality test of 5%, i.e., $\alpha_{ADF} = 5\%$ and $\alpha_{GR} = 5\%$.

7.1.1 Synthetic Data Generation

In what follows, we describe in details how the simulated environment generates the data. Figure 7.1 shows which information agent and environment exchange and their interaction. Given an advertising campaign $\mathcal{C} = \{C_1, \ldots, C_N\}$, with $N \in \mathbb{N}$, where $C_j$ is the $j$-th sub-campaign, a dependency matrix $D \subseteq \{0, 1\}^{N \times N}$, where $d_{ij} = 1$ if sub-campaign $C_j$ “influence” sub-campaign $C_i$, our simulator generates for each $j$-th sub-campaign and for each day $t \in [1, \ldots, T]$ a vector of observations $[n_{j,t}, c_{j,t}, c_{o,j,t}, c_{j,t}]^\top$, ...
7.1. Synthetic Settings

Figure 7.1: Interaction between agent and environment.

where \( n_{j,t} \) is the number of impressions, \( cl_{j,t} \) is the number of clicks, \( co_{j,t} \) is the number of conversions, and \( c_{j,t} \) is the total cost.

At day \( t \), each sub-campaign \( C_j \) is characterized by the set of the users \( S_{j,t} = s_{j,t} \cup \left( \bigcup_{j \neq j} s_{i,t} \right) \) that could potentially visualize the ad of the sub-campaign \( C_j \). More precisely, we distinguish the set of the users \( s_{j,t} \), that would visualize the ad of \( C_j \) without having previously visualized the ads of the other interdependent sub-campaigns, from the set of the users \( s_{ij,t} \), that would visualize the ad of \( C_j \) only after having visualized the ad of \( C_i \). Notice that \( s_{ij,t} \) is non-empty only if the sub-campaigns \( C_i \) and \( C_j \) are interdependent and, more precisely, if \( d_{ij} \neq 0 \).

The number of users \( |s_{j,t}| \) is sampled from \( \mathcal{N}(\mu_j, \sigma_j^2) \), i.e., a Gaussian distribution with mean \( \mu_j \) and variance \( \sigma_j^2 \). Each user in \( s_{j,t} \) is characterized by a click probability \( p_j^{(cl)} \) and a conversion probability \( p_j^{(co)} \) specific for the sub-campaign \( C_j \). Conversely, the number of users \( |s_{ij,t}| \) is modeled through a linear combination of the number of daily impressions \( n_{i,t-1}, \ldots, n_{i,t-K} \) (whose generation is described in what follows), where \( K \) represents the maximum delay in the interdependence dynamics. Formally, we have that

\[
 s_{ij,t} := p_{ij}^{(res)} \sum_{k=1}^{K} \beta_k n_{i,t-k}, \text{ where } \beta_k \in [0, 1] \text{ are randomly sampled coefficients and } p_{ij}^{(res)} \text{ is the probability that a user having visualized ad of } C_i \text{ is a potential user that may visualize } C_j. \text{ Each user in } s_{ij,t} \text{ is characterized by}
\]
a click probability $p^{(cl)}_{ij}$ and a conversion probability $p^{(co)}_{ij}$.

At each day $t$, setting the bid/budget pairs on each sub-campaign allows the advertiser to take part to $A_j \leq |S_{j,t}|$ auctions based on the Vickrey-Clarke-Groves mechanism [Mas-Colell et al., 1995], in which $\gamma_j$ available ad slots are allocated to a subset of $\delta_j$ advertisers ($\gamma_j \leq \delta_j$). More specifically, each advertiser submits her bid $b_h$ and those with the first $\gamma_j$ highest values $b_h \rho_h$ are allocated in the $\gamma_j$ slots, where $\rho_h$ is the probability that $h$-th ad is clicked given it has been observed. The bids $b_h$ of the other ads participating in the auctions are drawn from a truncated Normal distribution $\mathcal{N}(\mu(b), \sigma(b))$, and the click probabilities $\rho_h$ are uniformly sampled in $[0, 1]$.

In the case the advertiser wins the $m$-th auction, the ad gets an impression ($n_{m,j,t} = 1$, otherwise $n_{m,j,t} = 0$). The ad can be visualized by an user $S_{j,t}$ according to the probability of being observed $p^{(obs)}(l)$. After the impression, the user can click on the ad and generate a conversion according to the click $p^{(cl)}_{ij}$ and conversion $p^{(co)}_{ij}$ probabilities if the user belongs to $s_{j,t}$, and according to the click $p^{(cl)}_{ij}$ and conversion $p^{(co)}_{ij}$ probabilities if the user belongs to $s_{ij,t}$. A click on the ad of $C_j$ provided by the user corresponding to the $m$-th auction is denoted by $cl_{m,j,t} = 1$ ($cl_{m,j,t} = 0$ otherwise), and imposes a payment of $CPC_{m,j,t}$, as specified by the VCG auction (see [Mas-Colell et al., 1995], [Gatti et al., 2015a] for details). The auctions are generated until the daily budget $y_{j,t}$ allocated on the sub-campaign $C_j$ is totally spent, i.e., the total number of auctions $A_j$ is s.t. $A_j = \sum_{m=1}^{A_j} CPC_{m,j,t} = y_{j,t}$ or until $A_j = |S_{j,t}|$. Finally, in the case a click happen, the $m$-th user may convert ($co_{m,j,t} = 1$) or not ($co_{m,j,t} = 0$). The daily impressions, the daily clicks, the daily conversions (assuming unitary value per conversion), and the daily costs are computed as $n_{j,t} = \sum_{j=1}^{A_j} n_{m,j,t}$, $cl_{j,t} = \sum_{m=1}^{A_j} cl_{m,j,t}$, $co_{j,t} = \sum_{m=1}^{A_j} co_{m,j,t}$, $c_{j,t} = \sum_{m=1}^{A_j} CPC_{m,j,t}$, respectively.

7.1.2 Experiments

The following experiments consist of two phases: training phase and testing phase. During the exploration phase, we adopt a particular training procedure to mimic the behavior of a real advertiser and be as explorative as possible, to properly learn the sub-campaign inter-dependencies. More specifically, the exploration phase has been conducted in this way: in the first half of the exploration phase ($t \in \{1, \ldots, \frac{T}{2}\}$), we progressively allow the budget to be allocated to a larger set of sub-campaigns. At first the budget is allocated on the sub-campaign $C_N$ (with the bid set at random), then we randomly split the budget on $C_N$ and $C_{N-1}$, and so on. In the
remaining days (in \{\tau_2, \ldots, \tau\}) we allocate the budget randomly over all the
sub-campaigns. For the sake of notation, we assume that the order of the
sub-campaigns’ indices is one of the topological orders.

Once we collected a dataset \(Z := \{z_{j,t}\}\) of \(\tau\) samples that provide, for
each day \(t \in \{1, \ldots, \tau\}\) and each sub-campaign \(C_j \in \mathcal{C}\) with \(j \in \{1, \ldots, N\}\)
the used bid \(\tilde{x}_{j,t}\), and daily budget \(\tilde{y}_{j,t}\), the received impressions \(\tilde{n}_{j,t}\), clicks
\(\tilde{c}_{j,t}\), values of the conversions \(\tilde{c}_{o_{j,t}}\), and costs \(\tilde{c}_{j,t}\), we begin the exploitation
phase. During this phase, we exploit the collected information to find out
the optimal bid/budget allocation. The IDL algorithm firstly estimate the
inter-dependencies among sub-campaigns and after that proceeds with the
optimization phase. Instead, the DA-IDL algorithm find the optimal policy
with the full knowledge of the inter-dependencies matrix \(D\).

**Experimental Setting 1**

![Interdependence graph \(G\) for the Experimental Setting 1.](image)

In the first experimental setting there are \(N = 4\) sub-campaigns, with
delayed dynamics of \(K = 3\) days, whose interdependence graph is shown in
Figure 7.2. The longest path of the interdependence graph \(G\) is \(\Gamma = 1\). \(C_1\)
and \(C_2\) are on the display advertising channel and are targeted to a wide
range of daily users, thus generating a large number of daily auctions, but
their conversion probability is low. \(C_3\) and \(C_4\) are on the search advertising
channel, generating a small number of daily auctions, but their conversion
probability is high. We report in Tables 7.1–7.3 the values of the parameters
used in the synthetic Setting 1. We use \(Y = 500\) and \(B = 100\) daily budget
values evenly spaced in the range \([0, 500]\). The GPs used to estimate the
impressions model of the sub-campaigns adopt a squared exponential kernel
in which the kernel parameters are chosen as recommended by [Rasmussen
and Williams, 2006]. We evaluate the performance of the algorithms with
different numbers of samples \(\tau \in \{64, 80, 96\}\) in the training set \(Z\).
7.1. Synthetic Settings

Table 7.1: Parameters values for Setting 1.

<table>
<thead>
<tr>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_j$</td>
<td>5000</td>
<td>5000</td>
<td>200</td>
</tr>
<tr>
<td>$\sigma_j$</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>$\gamma_j$</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>$\delta_j$</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>$\mu^{(b)}_j$</td>
<td>0.89</td>
<td>1.19</td>
<td>1.59</td>
</tr>
<tr>
<td>$\sigma^{(b)}_j$</td>
<td>0.32</td>
<td>0.12</td>
<td>0.2</td>
</tr>
<tr>
<td>$p^{(obs)}(1)$</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>$p^{(obs)}(2)$</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>$p^{(obs)}(3)$</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>$p^{(obs)}(4)$</td>
<td>0.6</td>
<td>0.6</td>
<td>-</td>
</tr>
<tr>
<td>$p^{(obs)}(5)$</td>
<td>0.5</td>
<td>0.5</td>
<td>-</td>
</tr>
<tr>
<td>$p^{(ci)}_j$</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>$p^{(co)}_j$</td>
<td>0.001</td>
<td>0.05</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 7.2: Click probability $p^{(cl)}_{ij}$ and conversion probability $p^{(co)}_{ij}$ on $C_j$ given an impression generated by an interdependent sub-campaign for Setting 1.

(a) $p^{(cl)}_{11}$ $p^{(cl)}_{12}$ $p^{(cl)}_{13}$ $p^{(cl)}_{14}$

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>-</th>
<th>-</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_2$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.45</td>
</tr>
<tr>
<td>$C_3$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_4$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(b) $p^{(co)}_{11}$ $p^{(co)}_{12}$ $p^{(co)}_{13}$ $p^{(co)}_{14}$

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>-</th>
<th>-</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_2$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.35</td>
</tr>
<tr>
<td>$C_3$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$C_4$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 7.3: Research probability $p^{(res)}_{ij}$ on $C_j$ given an impression generated by an interdependent sub-campaign for Setting 1.

| $p^{(res)}_{11}$ $p^{(res)}_{12}$ $p^{(res)}_{13}$ $p^{(res)}_{14}$ |
|----------------|----------------|----------------|
| $C_1$ | -    | -    | 0.5  |
| $C_2$ | -    | -    | -    | 0.5  |
| $C_3$ | -    | -    | -    | -    |
| $C_4$ | -    | -    | -    | -    |

Results  In Figure 7.3a, we report the average (over 100 repetitions) revenue $R_t$ produced by the algorithms with a training of $\tau = 96$ samples. From $t = 96$ on, the optimal stationary solution is used. The average revenue of each algorithm peaks at $t = 97$ and, for $t > 97$, decreases by converging to a steady state within $K \Gamma + 1 = 4$ days. The peak is generated by the
7.1. Synthetic Settings

presence of a large number of residual users who have observed display ads during training and who, after \( t = 96 \), observe search ads. These residual users decrease for \( t > 96 \) until they reach a steady state. Thus, (temporary) peaks may be achieved with non-stationary policies.

The DA-IDL algorithm exhibits the best performance, exploiting the a priori knowledge of the adjacency graph \( D \). The gap between the revenue produced by the IDL and DA-IDL algorithms, due to the estimation error introduced on \( \hat{D} \), is sufficiently small, showing that the Granger Causality test used by the IDL algorithm works well in practice. Instead, the revenue produced by the AdComB-Mean algorithm, neglecting the interdependence among sub-campaigns, is much smaller than that produced by the other two algorithms. This is due to the very different budget allocations chosen by the three algorithms: the IDL and DA-IDL algorithms optimally balance the budget on all the sub-campaigns, while the AdComB-Mean algorithm greedily invests the budget only in the search sub-campaigns \( C_3 \) and \( C_4 \). Interestingly, the performance of the AdComB-Mean algorithm is quite similar to that of the uniformly random allocation used during training.

In Figure 7.3b, we report the average revenue \( R_t \) at the steady-state (averaged over the 100 independent repetitions) and the 95% confidence intervals as the number of samples \( \tau \) used for training increases. All algorithms always perform better than the uniformly random allocation. The performance of both the IDL and DA-IDL algorithms is significantly better than the one provided by AdComB-Mean (confidence intervals do not overlap). The use of more training samples provides an improvement in terms of steady-state revenue for the IDL and DA-IDL algorithms. On the other hand, the performance of the AdComB-Mean algorithm does not benefit from having more samples, which is probably due to the presence of a model bias induced by the fact that it neglects the sub-campaign interdependence.

In Tables 7.4–7.5 we report the number of times a specific dependency is detected by the IDL algorithm for different values of \( \tau \) over 100 repetitions. In bold are highlighted the true dependencies, while the others are false positive dependencies detected by IDL. Increasing the number of training samples provides better results in terms of true positive, while the number of false positive seems not to be change. The number of false positive is not consistent with \( \alpha_{GR} \) since there may be non-linear dependencies. The dependency between sub-campaigns \( C_2 \) and \( C_4 \) has been identified fewer times than the dependency between sub-campaigns \( C_1 \) and \( C_3 \). This is due to the fact that \( C_2 \) has a lower impact on \( C_4 \) than \( C_1 \) on \( C_3 \) since it wins less auctions. Indeed, the mean bid of the other ads participating in the auctions of sub-campaign \( C_2 \) \( \mu_2^{(b)} \) is greater than the one of \( C_1 \) \( \mu_1^{(b)} \).
Figure 7.3: Figure (a) shows the revenue $R_t$ over time for the Setting 1. In Figure (b) the revenue $R_t$ is reported in steady state conditions for different training sizes $\tau$ in Setting 1. The revenue of the random allocation is reported with a dotted magenta line and the vertical lines represent the 95% confidence intervals for the algorithms revenue.

Table 7.4: Percentages of rejecting the non-causality hypothesis for Setting 1 with $\tau = 64$, and $\tau = 80$. 

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>0</td>
<td>0</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>$C_2$</td>
<td>5</td>
<td>0</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>$C_3$</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$C_4$</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
<th>$C_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>0</td>
<td>2</td>
<td>96</td>
<td>2</td>
</tr>
<tr>
<td>$C_2$</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>$C_3$</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$C_4$</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 7.5: Percentages of rejecting the non-causality hypothesis for Setting 1 with \( \tau = 96 \).

<table>
<thead>
<tr>
<th></th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>0</td>
<td>3</td>
<td>95</td>
<td>0</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 7.4: Interdependence graph \( G \) for the Experimental Setting 2.

Experimental Setting 2

In the second experimental setting there are \( N = 3 \) sub-campaigns, with delayed dynamics of \( K = 3 \) days, whose interdependence graph is shown in Figure 7.4. The longest path of the interdependence graph \( G \) is \( \Gamma = 2 \). \( C_1 \) is a display sub-campaigns directed to a wide audience and have a low cost per impression, but a low conversion rate. \( C_2 \) is a social sub-campaign, whose number of impressions is influenced by the influence index of the display sub-campaigns. \( C_3 \) is on the search advertising channel, generating a small number of daily auctions, but their conversion probability is high. We report in Tables 7.6–7.8 the values of the parameters used in the synthetic setting 3. We use \( Y = 300 \) and \( B = 30 \) daily budget values evenly spaced in the range \([0, 300]\). We evaluate the performance of the algorithms with different numbers of samples \( \tau \in \{64, 80, 96\} \) in the training set \( Z \).
Table 7.6: Parameters of the Settings 2.

<table>
<thead>
<tr>
<th></th>
<th>C₁</th>
<th>C₂</th>
<th>C₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>µ₁</td>
<td>10000</td>
<td>10000</td>
<td>100</td>
</tr>
<tr>
<td>σ₁</td>
<td>100</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>γ₁</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>δ₁</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>µ(²)</td>
<td>0.10</td>
<td>1.0</td>
<td>1.59</td>
</tr>
<tr>
<td>σ(²)</td>
<td>0.32</td>
<td>0.12</td>
<td>0.2</td>
</tr>
<tr>
<td>p(¹)(1)</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>p(¹)(2)</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>p(¹)(3)</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>p(²)</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>p(⁴)</td>
<td>0.001</td>
<td>0.06</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table 7.7: Click probability p⁽ᶜ⁾_{ij} and conversion probability p⁽ᶜ⁾_{ij} on Cⱼ given an impression generated by an interdependent sub-campaign for Setting 2.

(a)  

<table>
<thead>
<tr>
<th></th>
<th>p⁽ᶜ⁾₁</th>
<th>p⁽ᶜ⁾₂</th>
<th>p⁽ᶜ⁾₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>-</td>
<td>0.35</td>
<td>-</td>
</tr>
<tr>
<td>C₂</td>
<td>-</td>
<td>-</td>
<td>0.4</td>
</tr>
<tr>
<td>C₃</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(b)  

<table>
<thead>
<tr>
<th></th>
<th>p⁽ᶜ⁾₁</th>
<th>p⁽ᶜ⁾₂</th>
<th>p⁽ᶜ⁾₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>-</td>
<td>0.25</td>
<td>-</td>
</tr>
<tr>
<td>C₂</td>
<td>-</td>
<td>-</td>
<td>0.35</td>
</tr>
<tr>
<td>C₃</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 7.8: Research probability p⁽ʳ⁾_{ij} on Cⱼ given an impression generated by an interdependent sub-campaign for Setting 2.

<table>
<thead>
<tr>
<th></th>
<th>p⁽ʳ⁾₁</th>
<th>p⁽ʳ⁾₂</th>
<th>p⁽ʳ⁾₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>-</td>
<td>0.7</td>
<td>-</td>
</tr>
<tr>
<td>C₂</td>
<td>-</td>
<td>-</td>
<td>0.6</td>
</tr>
<tr>
<td>C₃</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Results  In Figure 7.5a, we report the average (over 100 repetitions) revenue of the algorithms. Similarly to Settings 1 the performance of IDL and DA-IDL are significantly higher with respect to AdComB-Mean. Indeed, AdComB-Mean algorithm greedily invests the budget only in the search sub-campaign C₃ while IDL and DA-IDL algorithms balance the budget on all the sub-campaigns.
7.1. Synthetic Settings

Figure 7.5: Figure (a) shows the revenue $R_t$ over time for the Setting 2. In Figure (b) the revenue $R_t$ is reported in steady state conditions for different training sizes $\tau$ in Setting 2. The revenue of the random allocation is reported with a dotted magenta line and the vertical lines represent the 95% confidence intervals for the algorithms' revenue.

Table 7.9: Percentages of rejecting the non-causality hypothesis for Setting 2 with $\tau = 64, \tau = 78, \tau = 96$. 

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_1$</td>
<td>$C_2$</td>
<td>$C_3$</td>
<td>$C_1$</td>
</tr>
<tr>
<td>$C_1$</td>
<td>0</td>
<td>83</td>
<td>0</td>
</tr>
<tr>
<td>$C_2$</td>
<td>1</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>$C_3$</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Experimental Setting 3

In the third experimental setting there are \( N = 5 \) sub-campaigns, whose interdependence graph is shown in Figure 7.6. The longest path of the interdependence graph \( \mathcal{G} \) is \( \Gamma = 2 \). \( C_1 \), \( C_2 \), and \( C_3 \) are display sub-campaigns directed to a wide audience and have a low cost per impression, but a low conversion rate. \( C_4 \) is a social sub-campaign, whose number of impressions is influenced by the influence index of the display sub-campaigns. Finally, \( C_5 \) is a search sub-campaign, whose impressions depend on the influence index of \( C_1 \) and \( C_4 \). The interdependence among the sub-campaigns occurs within \( K = 3 \) days and it is modeled as in Setting 1. We report in Tables 7.10–7.13 the values of the parameters used in the synthetic Setting 3. We set a cumulative budget of \( Y = 500 \) and we selected the budget discretization from the interval \([0, 500]\) with \( B = 100 \). The number of samples for training is \( \tau \in \{100, 150, 200\} \).

<table>
<thead>
<tr>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
<th>( C_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.35</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.35</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.45</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 7.11: Click probability \( p_{ij}^{(cl)} \) on \( C_j \) given an impression generated by an interdependent sub-campaign for Setting 3.
7.1. Synthetic Settings

Table 7.10: Parameters of the Settings 3.

<table>
<thead>
<tr>
<th>(C_1)</th>
<th>(C_2)</th>
<th>(C_3)</th>
<th>(C_4)</th>
<th>(C_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu_j)</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>700</td>
</tr>
<tr>
<td>(\sigma_j)</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>(\gamma_j)</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>(\delta_j)</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>(\mu^{(o)})</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>1.0</td>
</tr>
<tr>
<td>(\sigma^{(o)})</td>
<td>0.032</td>
<td>0.032</td>
<td>0.012</td>
<td>0.2</td>
</tr>
<tr>
<td>(p^{(obs)}(1))</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>(p^{(obs)}(2))</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>(p^{(obs)}(3))</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>(p^{(obs)}(4))</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>-</td>
</tr>
<tr>
<td>(p^{(obs)}(5))</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-</td>
</tr>
<tr>
<td>(p^{(co)}_j)</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>(p^{(co)}_j)</td>
<td>0.001</td>
<td>0.001</td>
<td>0.011</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 7.12: Conversion probability \(p^{(co)}_{ij}\) on \(C_j\) given an impression generated by an interdependent sub-campaign for Setting 3.

<table>
<thead>
<tr>
<th>(C_1)</th>
<th>(C_2)</th>
<th>(C_3)</th>
<th>(C_4)</th>
<th>(C_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p^{(co)}_{11})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.4</td>
</tr>
<tr>
<td>(p^{(co)}_{12})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.4</td>
</tr>
<tr>
<td>(p^{(co)}_{13})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.4</td>
</tr>
<tr>
<td>(p^{(co)}_{14})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 7.13: Research probability \(p^{(res)}_{ij}\) on \(C_j\) given an impression generated by an interdependent sub-campaign for Setting 3.

<table>
<thead>
<tr>
<th>(C_1)</th>
<th>(C_2)</th>
<th>(C_3)</th>
<th>(C_4)</th>
<th>(C_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p^{(res)}_{11})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.45</td>
</tr>
<tr>
<td>(p^{(res)}_{12})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.45</td>
</tr>
<tr>
<td>(p^{(res)}_{13})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.45</td>
</tr>
<tr>
<td>(p^{(res)}_{14})</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Results In Figure 7.7a, we report the average (over 100 repetitions) revenue of the algorithms. The performance of AdComB-Mean is worse than the one of the uniformly random allocation and gets worse as \(\tau\) increases. This is an empirical confirmation of the statement of Theorem 7 showing that a solution that is optimal without interdependence might perform arbitrarily bad. Conversely, the performance of IDL and DA-IDL are significantly larger.
than that of the uniformly random allocation and increase as the number of samples increases. The most interesting behaviour is the one regarding the transition phase we experience at the end of the training phase. Differently from the Setting 1, we have that the more complex structure of the interdependence induces a longer transition phase (evidenced by drawing vertical black lines). The length of this phase supports the claim of Theorem 3 stating that the transition is at most \( K\Gamma + 1 = 6 \) days long.

Tables [7.14][7.15] show the number of times each dependency has been detected over 100 repetition with different values of \( \tau \) by the IDL algorithm. Even though dependency between sub-campaigns \( C_4 \) and \( C_5 \) has been identified few times, the IDL algorithm performance converge to that of the DA-IDL algorithm, as shown in Figure [7.7b] Taking into account the relation \( C_4 - C_5 \) does not provide better performance. Thus, sub-campaign \( C_4 \) does not significantly affect sub-campaign \( C_5 \).

Table 7.14: Percentages of rejecting the non-causality hypothesis for Setting 3 with \( \tau = 100 \).

<table>
<thead>
<tr>
<th></th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
<th>( C_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>66</td>
<td>2</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>87</td>
<td>99</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.15: Percentages of rejecting the non-causality hypothesis for Setting 3 with \( \tau = 150 \) and \( \tau = 200 \).

(a)  

<table>
<thead>
<tr>
<th></th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
<th>( C_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>99</td>
<td>0</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>98</td>
<td>1</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

(b)  

<table>
<thead>
<tr>
<th></th>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
<th>( C_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>99</td>
<td>0</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>98</td>
<td>1</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 7.7: Figure (a) shows the revenue $R_t$ over time for the Setting 2. In Figure (b) the revenue $R_t$ is reported in steady state conditions for different training sizes $\tau$ in Setting 3. The revenue of the random allocation is reported with a dotted magenta line and the vertical lines represent the 95% confidence intervals for the algorithms revenue.

7.2 Real Data Experiments

7.2.1 Causality detection

In the following sections we shows the results of the Granger Causality test (see Section 3.3 for more details) applied to real-world Internet advertising campaigns.
7.2. Real Data Experiments

Figure 7.8: Graphs representing the interdependencies of real-world Internet advertising sub-campaigns inferred by Granger Causality test from real data. The numbers on the edges are the p-values (in terms of %) of the Granger Causality test; display ads are depicted in blue, social ads in yellow, and search ads in red.

Campaign 1

In order to test the Granger Causality, we collected data for 8 months, from 01/01/2018 to 01/08/2018, from an Internet advertising campaign for a financial service of an insurance company. Data correspond to $N = 12$ sub-campaigns, on Google AdWords (search channel), Facebook (social channel), and Google display with a cumulative budget of $Y = 600$ Euros. The results obtained from Granger Causality test are shown in Figure 7.8, where the most significant elements of $(\tilde{n}_{j,t}, \tilde{c}_{j,t}, \tilde{c}_{o,j,t}, \tilde{c}_{l,j,t})$ are represented as nodes of different colours according to their specific channel: display ads are depicted in blue, social ads in yellow, and search ads in red. The detected interdependencies (with a p-value less than 5%) are represented as directed edges. In particular, Figure 7.8(a) shows the results when all the sub-campaigns data are aggregated by channels, while Figure 7.8(b) focuses only on sub-campaigns targeting retired people. These results confirm the presence of the interdependence between display and search advertising as previously observed in the literature. They also show that social and search advertising are interdependent and that the inter-dependencies may be targeting specific. Moreover, the interdependence between clicks and impressions of the social channels and the impressions of the search one in this specific scenario seems to be more relevant than others, since they appear in both graphs. Furthermore, the Granger Causality test detects that interdependence dynamics between sub-campaigns are delayed up to 2 days.
7.2. Real Data Experiments

Figure 7.9: Graphs representing the inter-dependencies of real-world Internet advertising sub-campaigns inferred by Granger Causality test from real data. The numbers on the edges are the p-values (in terms of %) of the Granger Causality test; social ads in yellow, and search ads in red (for branded sub-campaigns) and orange (for other search sub-campaigns). \( \bar{y}_{j,t} \) of different sub-campaigns \( C_j \).

Campaign 2

We test for Granger causality with a second dataset with data collected for 3 months (from 20/7/2018 to 20/10/2018) from an Internet advertising campaign of another financial product with about \( Y = 1100 \) Euros. There are \( N = 14 \) sub-campaigns belonging to social and search advertising channels. The resulting graph is depicted in Figure 7.9 (with a p-value less than 5%). As in the previous dataset, many sub-campaigns are subject to inter-dependence. In particular, in this case, the interdependence phenomenon is only among impressions, suggesting that these can be the most significant in practice. Moreover, differently from the previous case, we distinguish search sub-campaigns into two sub-classes which are at different depths in the marketing funnel: unbranded (orange nodes) or branded (red nodes). Finally, the delay \( k \) of the interdependence dynamics is up to 3.

Campaign 3

We experiment the Granger Causality also with a third dataset from 17/04/2018 until 10/11/2018, for a total of 7 months. The Internet advertising campaign involved in this experiment is related to an airport parking service. There are \( N = 4 \) sub-campaigns, belonging to search and social channels, with a total budget \( Y = 5300 \) Euros. In Figure 7.10 the resulting inter-dependencies graph is represented, with a p-value less than 5%. Also in this campaign, we distinguish search sub-campaigns into two sub-classes which are at different depths in the marketing funnel: unbranded (orange nodes) or branded (red
7.2. Real Data Experiments

7.2.1 Nodes. In this case, the ad of the no-branding sub-campaign is shown when a user is looking for a competitor on the search channel. It can be considered as an unbranded sub-campaign, since it could be an intermediate research before reaching the focus on the campaign’s brand.

7.2.2 IDL Experiment

In the following experiment, we use real-world data from an Italian internet advertising company described above (Campaign 2) to train our model. We recall that the length of the dataset is \( \tau = 93 \) days (from 20/7/2018 to 20/10/2018), the advertising campaign \( \mathcal{C} = \{C_1, \ldots, C_N\} \) is composed of \( N = 14 \) sub-campaigns belonging to both social and search advertising channels. The corresponding estimated interdependence is provided in Figure 7.9. From 25/10/2018 to 24/11/2018, the campaign optimization has been performed by the IDL algorithm, with the exception of the week from 6/11/2018 to 13/11/2018, when the previous algorithm has been used.

At day \( t \), we are asked to set for each sub-campaign \( C_j \) a bid \( x_{j,t} \in [\underline{x}_j, \overline{x}_j] \), and a daily budget \( y_{j,t} \in [\underline{y}_j, \overline{y}_j] \), subject to that the daily cumulative budget of all the sub-campaigns cannot exceed \( Y \in \mathbb{R}^+ \). At day \( t + 1 \), we observe the performance of the campaign \( \mathcal{C} \) at the previous day \( t \), which specifies, for every \( C_j \), the tuple \((\tilde{n}_{j,t}, \tilde{c}_{j,t}, \tilde{c}_{j,t}, \tilde{c}_{j,t})\), where \( \tilde{n}_{j,t} \) denotes the

**Figure 7.10:** Graphs representing the inter-dependencies of real-world Internet advertising sub-campaigns inferred by Granger Causality test from real data. The numbers on the edges are the p-values (in terms of \( \% \)) of the Granger Causality test; social ads in yellow, and search ads in red (for branded sub-campaigns) and orange (for other search sub-campaigns). \( \tilde{n}_{j,t} \) of different sub-campaigns \( C_j \).
number of impressions, $\tilde{c}_{l,j,t}$ denotes the number of received clicks, $\tilde{c}_{o,j,t}$ denotes the cumulative value of the conversions, and $\tilde{c}_{j,t}$ denotes the amount of money spent for it. The money spent in one day for a sub-campaign may be different from the daily budget previously allocated.

We extend the previous studies on the sub-campaigns interdependence described in Chapter 4 applying the Granger Causality test [Granger, 1969] to real-world Internet advertising campaigns optimized by an Italian web media agency using the AdComb-TS algorithm [Nuara et al., 2017]. The algorithm, being online, produces policies explorative enough to make the test significant. This method provides a formal procedure to statistically test if the values of a time series are significantly improving the prediction of another time series. In such a case the first time series is said to Granger cause the second one. To the best of our knowledge, this is the only test able to provide a statistical significance of a direct relationship between time series.

The interdependence suggested by the Granger Causality Test are confirmed by estimations provided by the GPs. Indeed, in Figures 7.11a–7.11b we show the expected value of the prediction by the GPs of the number of potential impressions, i.e., the number of impression with unlimited budget, for the sub-campaigns $C_5$ and $C_6$. In both cases we report the number of potential impressions with a bid value set to the most frequent choice during the training set, in the first case $x = 5$ while in the second case $x = 1$. In Figure 7.11b the number of potential impressions increase as the value of the influence index increases, suggesting that a positive correlation between $C_5$ and $C_6$ exist. However, in Figure 7.11a the number of potential impressions does not monotonically increase with respect to the influence index. Thus, we cannot exclude the presence of negative interdependence in the optimization phase. Indeed, to compute the optimal allocation with the IDL algorithm we discard solutions with the Algorithm 5. In this way we are able to provide the optimal allocation even if generic interdependence among sub-campaigns are present.

In Figures 7.12a–7.12c the daily cost, the daily cost-per-lead (CPL), i.e., the daily cost divided by the total daily conversions, and the total daily conversions are reported considering only the sub-campaigns involved in the sub-graph $S_1 = \{C_2, C_5, C_5, C_6\}$. We can identify two periods of the IDL algorithm (24/10 – 6/11 and 13/11 – 24/11), and we can observe that the higher investment in the last period leads to lower CPL and higher
conversions. Thus, to exploit at their best the interdependencies, higher investments may be necessary.

Figures 7.14a–7.14c show that, given the same investment, the two algorithms have almost the same performances in terms of conversions and CPL during the observed period. This is a result of the minor impact of the sub-campaigns involved in the sub-graph \( S_1 \) on the overall conversions.

In Figure 7.13a and Figure 7.15a are respectively reported the daily conversions generated by \( S_1 \) and by \( C \) given the total daily cost incurred. Figure 7.13b and Figure 7.15b show the curves fitted with the data of Figure 7.13a and Figure 7.15a. Due to the absence of enough data, the obtained results do not provide statistical evidence that the AdComb-TS algorithm is better than the IDL algorithm, and viceversa. Furthermore, it has not been possible to directly compare the performance of the two algorithms online, since the company did not have an effective system to perform A/B testing for this setting.
7.2. Real Data Experiments

Figure 7.11: GPs estimation of the potential number of impressions $\hat{n}_{5}^{sat}(5, u)$ and $\hat{n}_{6}^{sat}(1, u)$ depending on the influence index $u$. 
7.2. Real Data Experiments

Figure 7.12: In (a), (b), and (c) are reported the daily cost, the daily CPL, and the daily conversions of the sub-campaigns of the sub-graph $S_1 = \{C_2, C_3, C_5, C_6\}$. In blue are reported the values of the AdComb-TS algorithm, and in red the values of the IDL algorithm.
Figure 7.13: Figure (a) shows the relationship between the daily conversions and the daily cost of the sub-campaigns of the sub-graph $S_1 = \{C_2, C_3, C_5, C_6\}$. In Figure (b) are reported the curves fitted with the data of Figure (a).
Figure 7.14: In (a), (b), and (c) are reported the daily cost, the daily CPL, and the daily conversions of the sub-campaigns of the overall campaign. In blue are reported the values of the AdComb-TS algorithm, and in red the values of the IDL algorithm.
Figure 7.15: Figure (a) shows the relationship between the daily conversions and the daily cost of the overall campaign. In Figure (b) are reported the curves fitted with the data of Figure (a).
Chapter 8

Conclusions and Future Works

8.1 Conclusions

In this work, we formalize, for the first time, the problem of optimizing an Internet advertising campaign exploiting sub-campaigns interdependence. We consider interdependence among sub-campaigns belonging to different channels, like search, social and display. The performances of sub-campaigns of different channels cannot be directly compared. Indeed, a display and social sub-campaigns provide lower conversions than search sub-campaigns, but they contribute to those conversions.

To confirm the presence of interdependence among sub-campaigns, we apply the Granger Causality test on data coming from real Internet Advertising campaigns. In particular, the results present in the literature confirm the presence of the interdependence between display and search advertising, and between social and search advertising.

We design the IDL algorithm that, given a set of past observations, models these interdependencies and returns an optimal allocation of the bid/daily budget on the sub-campaigns maximizing the revenue. The IDL algorithm consists of two phases: the Interdependence Graph Learning Phase and the Estimation and Optimization Phase. Identifying the most significant interdependencies is crucial since the complexity of the optimization procedure strictly depends on the number of dependencies.

We analyze the properties of the IDL algorithm, providing a bound on the total error and showing that employing a model that ignores the dependencies provide a loss in the revenues. We empirically show that the IDL algorithm provides revenues on synthetic settings significantly better than
8.2 Future Works

In our model we included only interdependence among impressions, since they turn out to be the most significant. However, we could extend this model by including also the interdependence between clicks and conversions. In our model we assume linear interdependencies among sub-campaigns. However, in a real-world setting it could be not always verified. As a future work, it is interesting to extend the Granger Causality with a non-linear model and compare the results with the linear case. Moreover, it is possible to improve the accuracy of the model by further investigating the regularities on the number of impressions with respect to the influence index.

Finally, the proposed algorithm learns a stationary optimal policy in an offline fashion starting from a dataset of \( \tau \) samples. A future direction could be extend this work in an online framework examining the right balance between exploration and exploitation, and ensuring to find a stationary policy.
Bibliography


Jason Mandar. Understanding the mobile-only Internet user, 2017. [URL]{https://insight.globalwebindex.net/hubfs/Reports/Trends-17.pdf}


J Xu, K.-C. Lee, W Li, H Qi, and Q Lu. Smart pacing for effective online ad campaign optimization. In SIGKDD, pages 2217–2226, 2015.

Appendix A

Granger Causality Experiments

A.1 Synthetic Experiments

We report in this section some experiments on the Granger Causality detection, with synthetic data. We consider 100 realizations of different simulation systems for time series of length $n = 365$. The performance of the Granger causality is quantified by the percentage of statistically significant couplings in the 100 realizations.

A.1.1 System 1

System Definition  At first, we simulate data from a stationary, linear and uncoupled process. We used the following linear dynamic system to generate the data we used to test Granger causality:

$$
\begin{align*}
    x_1(t) &= 0.1 x_1(t-1) - 0.3 x_1(t-2) + \varepsilon_1(t) \\
    x_2(t) &= 0.7 x_2(t-1) + \varepsilon_2(t)
\end{align*}
$$

(A.1)

where $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 1$ is zero-mean Gaussian white noise.

Results  It is evident from Equation [A.1] that there is no dependence between the two variables. What we expect from this experiment is a low or null percentage, over the one hundred tests. The results obtained with the data coming from system in Equation [A.1] are presented in Table A.1.
Appendix A. Granger Causality Experiments

Table A.1: Percentages of rejecting the non-causality hypothesis for System 1.

<table>
<thead>
<tr>
<th>Coupling</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1 \rightarrow x_2$</td>
<td>8</td>
</tr>
<tr>
<td>$x_2 \rightarrow x_1$</td>
<td>6</td>
</tr>
</tbody>
</table>

A.1.2 System 2

**System Definition** In this experiment we use a non-stationary, linear and bivariate model, with bidirectional couplings $x_1 \leftrightarrow x_2$, to generate data. More specifically:

\[
\begin{align*}
  x_1(t) &= -0.7 + 0.7x_1(t-1) + 0.2x_2(t-2) + \epsilon_1(t) \\
  x_2(t) &= 1.3 + 0.2x_1(t-1) + 0.2x_2(t-2) + \epsilon_2(t)
\end{align*}
\]  

(A.2)

where $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 1$ is zero-mean Gaussian white noise.

**Results** Non-stationarity and bidirectional coupling are present in the system: the two variables are dependent from each other at time $t+2$. For this reason we expect an high percentage on both relations, confirmed by the experiment. Obtained results with the data coming from system in Equation A.2 are presented in Table A.2.

Table A.2: Percentages of rejecting the non-causality hypothesis for System 2.

<table>
<thead>
<tr>
<th>Coupling</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1 \rightarrow x_2$</td>
<td>99</td>
</tr>
<tr>
<td>$x_2 \rightarrow x_1$</td>
<td>98</td>
</tr>
</tbody>
</table>

A.1.3 System 3

**System Definition** In this experiment we use a non-stationary, linear, multivariate system, with unidirectional couplings $x_1 \rightarrow x_3$, $x_2 \rightarrow x_1$, $x_2 \rightarrow x_3$, $x_4 \rightarrow x_2$, generated by a $VAR(5)$ model. The specific system is:

\[
\begin{align*}
  x_1(t) &= 0.8x_1(t-1) + 0.65x_2(t-4) + \epsilon_1(t) \\
  x_2(t) &= 0.6x_2(t-1) + 0.6x_4(t-5) + \epsilon_2(t) \\
  x_3(t) &= 0.5x_3(t-3) - 0.6x_1(t-1) + 0.4x_4(t-4) + \epsilon_3(t) \\
  x_4(t) &= 1.2x_4(t-1) - 0.7x_4(t-2) + \epsilon_4(t)
\end{align*}
\]  

(A.3)

where $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 1$ is zero-mean Gaussian white noise.
A.1. Synthetic Experiments

Results  Since in this model we have unidirectional couplings $x_1 \rightarrow x_3$, $x_2 \rightarrow x_1$, $x_2 \rightarrow x_3$, $x_4 \rightarrow x_2$, we expect high percentage in correspondence of these relations. The results obtained with the data coming from system in Equation A.3 show that the Surplus lag approach indicate the true couplings, while no spurious couplings are observed (Table A.3).

Table A.3: Percentages of rejecting the non-causality hypothesis for System 3.

<table>
<thead>
<tr>
<th>Coupling</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1 \rightarrow x_2$</td>
<td>4</td>
</tr>
<tr>
<td>$x_1 \rightarrow x_3$</td>
<td>100</td>
</tr>
<tr>
<td>$x_1 \rightarrow x_4$</td>
<td>6</td>
</tr>
<tr>
<td>$x_2 \rightarrow x_1$</td>
<td>100</td>
</tr>
<tr>
<td>$x_2 \rightarrow x_3$</td>
<td>100</td>
</tr>
<tr>
<td>$x_2 \rightarrow x_4$</td>
<td>5</td>
</tr>
<tr>
<td>$x_3 \rightarrow x_1$</td>
<td>3</td>
</tr>
<tr>
<td>$x_3 \rightarrow x_2$</td>
<td>2</td>
</tr>
<tr>
<td>$x_3 \rightarrow x_4$</td>
<td>4</td>
</tr>
<tr>
<td>$x_4 \rightarrow x_1$</td>
<td>5</td>
</tr>
<tr>
<td>$x_4 \rightarrow x_2$</td>
<td>97</td>
</tr>
<tr>
<td>$x_4 \rightarrow x_3$</td>
<td>1</td>
</tr>
</tbody>
</table>

A.1.4 System 4

System Definition  We generate data from a non-stationary, linear, bivariate system, with unidirectional coupling $x_2 \rightarrow x_1$, with cointegrated variables. Formally:

$$
\begin{cases}
  x_1(t) = 0.7x_2(t-1) + \epsilon_1(t) \\
  x_2(t) = x_2(t-1) + \epsilon_2(t)
\end{cases}, \quad (A.4)
$$

where $\epsilon_1(t) \sim N(0, \sigma^2)$, $\sigma = 1$ is zero-mean Gaussian white noise.

System 4 is co-integrated since variable $x_2$ is non-stationary ($x_2 \sim I(1)$) and, consequently, also $x_1$ is non-stationary. Nonetheless, there exists a linear combination of them which is stationary:

$$
x_1(t) - 0.7x_2(t) = 0.7x_2(t-1) + \epsilon_1(t) - 0.7(x_2(t-1) + \epsilon_2(t))$

$$
= \epsilon_1(t) - 0.7\epsilon_2(t) \sim I(0)
$$
Table A.4: Percentages of rejecting the non-causality hypothesis for System 4.

<table>
<thead>
<tr>
<th>Coupling</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1 \rightarrow x_2$</td>
<td>5</td>
</tr>
<tr>
<td>$x_2 \rightarrow x_1$</td>
<td>100</td>
</tr>
</tbody>
</table>

Results  The results obtained with the data coming from system in Equation A.4 are as we expected: the unidirectional coupling $x_2 \rightarrow x_1$ is revealed, even if the variables are cointegrated. Results are presented in Table A.4.

A.1.5 System 5

System Definition  At last, we examine a non-stationary, linear, multivariate system, with unidirectional couplings $x_2 \rightarrow x_1$ and $x_3 \rightarrow x_1$, with co-integrated variables, defined as:

$$
\begin{align*}
    & x_1(t) = 0.4x_1(t - 1) + 0.4x_2(t - 1) + 0.5x_3(t - 1) + 0.2x_1(t - 2) \\
    & -0.2x_2(t - 2) - 0.2x_1(t - 3) + 0.15x_2(t - 3) + 0.1x_3(t - 3) + \epsilon_1(t) \\
    & x_2(t) = 0.6x_2(t - 1) + 0.2x_2(t - 2) + 0.2x_2(t - 3) + \epsilon_2(t) \\
    & x_3(t) = 0.4x_3(t - 1) + 0.3x_3(t - 2) + 0.3x_3(t - 3) + \epsilon_3(t)
\end{align*}
$$

(A.5)

where $\epsilon_i(t)$ is Gaussian white noise. The variables of System 5 are $I(1)$ and co-integrated with one co-integration relationship.

Results  The results obtained with the data coming from system in Equation A.5 are presented in Table A.5. Also in this case the obtained results are as we expected and the unidirectional couplings $x_2 \rightarrow x_1$ and $x_3 \rightarrow x_1$ are revealed.

Table A.5: Percentages of rejecting the non-causality hypothesis for System 5.

<table>
<thead>
<tr>
<th>Coupling</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1 \rightarrow x_2$</td>
<td>5</td>
</tr>
<tr>
<td>$x_1 \rightarrow x_3$</td>
<td>9</td>
</tr>
<tr>
<td>$x_2 \rightarrow x_1$</td>
<td>99</td>
</tr>
<tr>
<td>$x_2 \rightarrow x_3$</td>
<td>5</td>
</tr>
<tr>
<td>$x_3 \rightarrow x_1$</td>
<td>100</td>
</tr>
<tr>
<td>$x_3 \rightarrow x_2$</td>
<td>5</td>
</tr>
</tbody>
</table>