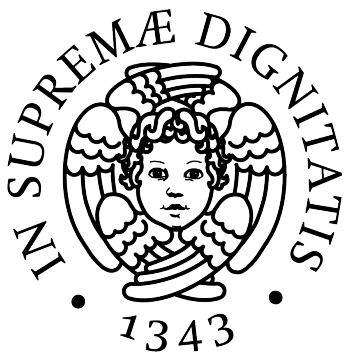

UNIVERSITÀ DI PISA

Dipartimento di Filologia, Letteratura e Linguistica

Corso di Laurea Magistrale in Informatica Umanistica

TESI DI LAUREA

A Budget Allocation Framework for Online Advertising, using Bayesian Inference

Relatore:
Prof. Nicola Ciaramella

Candidato:
Sirio Papa

ANNO ACCADEMICO 2020/2021

Infinite gratitude to my sweet Elisabetta, Nello, and especially Anna Elisa.
Always by my side.

Una gratitudine infinita ai miei dolci Elisabetta, Nello e soprattutto Anna Elisa.
Sempre al mio fianco.

Contents

List of Figures **5**

Introduction **8**

1 Why designing a Framework for Budget Allocation **12**

1.1 A Fast-Growing Market 13

1.2 Emerging Markets, Stable Markets, Hungry Markets 15

1.3 Trends, Technologies and Emerging Formats 17

1.4 GDPR and Privacy Concerns 20

1.5 Summary 21

2 All you need is Bayes **24**

2.1 Sample Space and Random Variables 25

2.2 Probability Distributions, Mass, and Density 27

2.3 Joint, Marginal and Conditional Probabilities 30

2.4 Posterior, Likelihood and Prior 35

2.5 The Coin Flipping Example 40

2.6 The Bayesian Process 42

2.7 Monte Carlo Markov Chains and Metropolis-Hastings 43

2.8 Bayes and Cognitive Biases 46

2.9 Summary 48

3 A Bayesian Framework for Online Advertising Budget Allocation **51**

3.1 Premises and Scope 52

3.2 The Framework in a Nutshell 55

3.3 The Payout Curve 57
3.4 The Optimal Point And Marginalization Problems 62
3.5 Temporal Information 70
3.6 Can the Framework Fail? 72
3.7 Summary 73

4 Implementation Choices and Development **76**
4.1 The Data 76
4.2 The Model 79
4.3 Sampling Process 83
4.4 Results and Debugging 84
4.5 Curve Extraction 88
4.6 The Optimal Point 91
4.7 Cost and Budget 94
4.8 Summary 95

Conclusions **98**

Bibliography **103**

Aknowledgements **107**

List of Figures

1.1 Global advertising expenditure by medium - US$ million at current prices (Zenith, 2021) 14
1.2 Global advertising revenues by medium - US$ million at current prices (GroupM, 2021) 15
1.3 Top 20 countries by advertisers' major media expenditure - US$ million at current prices (Zenith, 2021) 16
1.4 Global online advertising expenditure by device (Zenith, 2021) 17
1.5 Global online advertising expenditure by format (Zenith, 2021) 18
1.6 Social vs. Non-Social adspend share in online advertising (Statista, 2021) 19
1.7 Desktop and mobile social media adspend with respect to the total adspend (Statista, 2021) 20

2.1 Probability density of a discretized variable generated from a $N(\mu$=170cm, σ=15). The red line is set to μ. 29
2.2 Joint probability distribution of a bivariate $Normal$ distribution $\mu = 10, \sigma = 1.5$. 32
2.3 3D view of a joint probability distribution of a bivariate normal distribution $\mu = 10, \sigma = 1.5$. 33
2.4 Differences between three $Student - t$ distributions with different degree of freedom and a normal distribution with same μ and σ. 39

3.1 The shapes of two power-law with different exponent 59
3.2 Example of optimal point with C=10 (value of a conversion) 64
3.3 How the optimal point changes with the marginal approach over the constant λ 65

3.4 Plot of $f'(x)$. When $f'(x) = 1$, we spot the optimal point that is reported in $f(x)$ in Figure 3.3 66
3.5 How the optimal point changes with the average approach over the constant λ 67
3.6 Positioning of the optimal point over the variation of α with marginal approach 68
3.7 Positioning of the optimal point over the variation of α with conversion average value approach 69

4.1 Campaign Return vs. Cost daily data plot 77
4.2 Number of Outliers vs. Number of Neighbors in LOF 78
4.3 Dataset shape after outliers detection 78
4.4 $Half\ Normal$ distribution with different σ values 81
4.5 Beta distribution with different values of α and β 82
4.6 ArviZ trace plot. On the right there are the distribution, and the random walk on the left. 85
4.7 ArviZ trace plot. On the right there are the distribution, and the random walk on the left. 86
4.8 ArviZ pair plot. The scatter plots represent the draws with respect to a pair of parameters, while the density functions are the marginal probabilities of the single parameters. 87
4.9 ArviZ autocorrelation plot. 88
4.10 Curve random extrapolation of 300 draws 90
4.11 Optimal Point for Marginal and Average CPA approach. 91

Introduction

The current thesis focuses on the presentation of a work that lies between the experiment and the project. We are going to present a framework thought for the optimization and automation of budget allocation for advertising online campaigns. The core of the framework is the extrapolation of a regression curve that describes the campaign payout. This task is performed through Bayesian inference and this topic will be a pattern throughout the whole essay. The starting point is a 2020 paper by Han and Gabor[1] in which they report a work of optimization in online advertising campaigns done at Lyft, a major American company in the private transportation and logistics industry. The purpose of the framework is to maximize the number of conversions, using a marginalistic approach, having fixed the value of a single conversion. The choice of a Bayesian inference solution to the budget allocation problem comes from many reasons, but in particular, we wanted to build a solution able to work even in the presence of little data, or high variance, thanks to the exploratory approach of the algorithms.

The argumentation begins with a brief analysis of the advertising market, with particular attention to the online segment. We firstly take into account the global condition of the sector, but the focus is more concentrated on the peculiarities of the European market. The most meaningful and recent trends in online advertising will also be analyzed to try to give a glimpse of what could be the future developments of the digital market. Finally, it will be concluded with a brief reference to the developments that the Web Analytics market (a sector that directly and indirectly impacts also the digital advertising market) could undergo following the rulings of the European courts that follow the European regulation in the field of privacy and management of personal data of European citizens (GDPR). This chapter

[1] Han and Gabor, "Contextual Bandits for Advertising Budget Allocation".

aims to make the reader conscious of the rate at which the online advertising market is growing and how much companies are investing in this direction in the last years.

Since the dimension of the digital advertising market is constantly growing, it is considered important to have a tool that can help make informed and consistent decisions for advertising spending, especially for those who decide to invest significant amounts of money. It is also believed that tools like this will become increasingly necessary in a market where competition is galloping and spending is growing. It is reasonable to believe that it is also important to develop independent tools and don't rely only on products offered by the large advertising platforms, which are usually complete Black Boxes, in order to be able to have a clearer vision of the processes.

The second chapter of the thesis will deal with the bases of inferential statistics that are necessary to arrive at the full formulation of the Bayes theorem. A treatment was chosen that could start from the basics so as to be able to construct the discourse by providing all the pieces necessary to have, in the end, the means to fully understand the potential and implications of the Bayesian approach. The issue, however, will be addressed in a critical and reasoned way: the strengths and criticalities of both the Bayesian and the frequentist approaches will be highlighted, a perspective will be given on the algorithms used, the philosophy behind our Bayesian choice, the multiple implications of Bayes theorem also in everyday life.

The third chapter will deal with the theoretical structure of the framework. We will explain what is the purpose and the premises from which we need to start the construction of the entire system. We will deepen the question of the extraction of the curve, trying to fully understand the behavior of our function according to the context. Subsequently, we will deal with another focal aspect of the framework: how we can identify the optimal point that corresponds to the budget to invest, following two different approaches. Then we will give some possible solutions and points of reflection about the management of temporal information by the framework, and finally, we will try to bring out those contexts in which the framework could behave differently than expected.

Finally, the fourth chapter will include the part relating to the practical implementation

of the framework. We will start by discussing how we should process the available data before feeding it to the machine. Below we will describe, report, and explain the supporting structure of the model thanks to which Bayesian inference will be performed, but also how PyMC3[2], the powerful python library we are going to use, can, with a line of code, perform such complex calculations. The statistical choices that led to the construction of the model will be justified and then we will analyze the results trying to understand if the output provided is in line with expectations or not. Then the process of extracting the curve will be displayed and below we will show how in practice the optimal point of the curve is spotted. Finally, we will explain what the relationship between cost and budget is, how they should be intended, the common misconceptions, and how their relationship should be managed, with respect to the framework.

In the final discussions, however, we will explain how the framework was not able to optimize a budget allocation process in a specific case. We will have the opportunity to understand what are the optimal conditions under which the framework can express itself at its best, but in any case, a practical and documentary value will be recognized for the operations carried out.

[2]Salvatier, Wiecki, and Fonnesbeck, "Probabilistic programming in Python using PyMC3".

Chapter 1

Why designing a Framework for Budget Allocation

In this chapter, we will briefly analyze the current state of the Advertising market worldwide with a focus on the Internet Advertising market to highlight how much this sector is an increasingly fundamental part of the economic growth of many businesses. The fastest growing Countries will be highlighted to point out where this market is running the most. Furthermore, we will focus on the technologies and formats that currently constitute the most powerful drivers of investments and returns in the Online Advertising market to give as complete a vision as possible, albeit superficial and rapid, on the main trends of the market and also on their prospects. Reference will be made in several places to the impact that the pandemic has had on this sector to demonstrate that $i)$ the recovery has been rapid and more vigorous than might have been expected, despite the total eclipse of some important economic sectors that made massive use of advertising $ii)$ Online Advertising now as never before can be a vector of economic growth for many businesses (small and large) even in times of crisis, given its versatility, plasticity, and precision. It will then conclude with a short paragraph about the developments that Web Marketing, and consequently online advertising, could have after some European judgments that follow the implementation of the GDPR.

A reflection of this type was considered important at the beginning of this essay to underline how important it can be to develop an intelligent framework for budget allocation in online

advertising, given that expenditure in this sector is expected to reach 550 billion dollars, by 2024 globally. Since this market is responsible for the 0.77% of global GDP, and it is forecast to reach 0.8% by 2024, it is clear that business intelligence has to provide new tools to respond to new needs.[1] In our case, the main purpose is to make data-driven decisions for the allocation of budget, day by day, in a set of online advertising campaigns, but digital marketing, in general, is an industry quite interested in Artificial Intelligence, automation, and smart solutions. Spending money in a better and more organized way than a competitor can be the difference between wasting large sums of money, or maximizing your returns. This is true for any type of business, but it can be particularly significant for large companies where optimizing a process by a few percentage points can make a big difference in the long run (such as saving on advertising costs).

1.1 A Fast-Growing Market

The online advertising market has been growing more and more, over the years and, especially during the COVID-19 pandemic, it has demonstrated to be one of the most important boosters for businesses of any kind and size. The digital advertising sector has been proven to be strongly resilient even during the most severe 21st-century outbreak and it has been a fundamental tool for businesses to keep operating and delivering goods and services during the first strict lockdown of 2020 spring when no one was prepared for the exceptional nature of the situation. For many small and medium businesses, the online market has been a new means to avoid irreparable losses and those who had previous experiences with the sector have been capable to adapt more easily to lockdowns.

Despite the pandemic, the global internet advertising market has kept growing rapidly with an annual rate of almost 10%, while the advertising market, in general, suffered a plunge of -3.9%.[2] The data reported by Zenith show that online advertising is the only adv media that faced increased revenues, pushing the global advertising expenditure in 2020 to over 600 billion dollars and it is forecast to reach 870 billion by '24.[3]

[1] Zenith, *Advertising Expenditure Forecasts. December 2021*, p. 6.
[2] Ibid., p. 5.
[3] Ibid., p. 5.

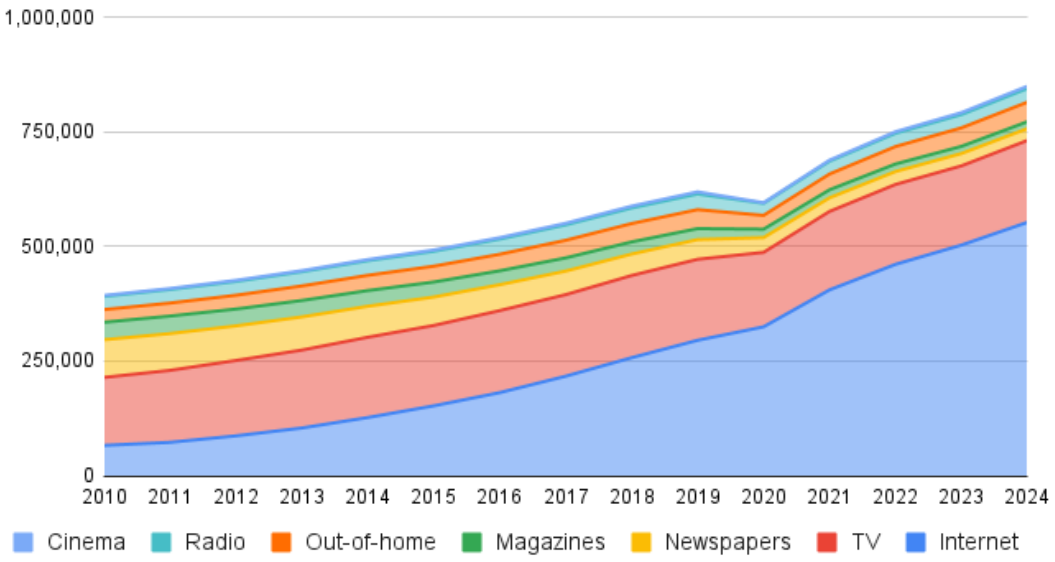

Figure 1.1: Global advertising expenditure by medium - US$ million at current prices (Zenith, 2021)

It is important to highlight that the only advertising medium that is constantly growing through the last 12 years is online, and following the projections proposed by Zenith, it appears not only that the pandemic has had a marginal impact, but that it has accelerated the growth process (Figure 1.2). Furthermore, Businesses' investments in online advertising are not just pioneer investments in a fast-growing industry but are actually supported by strong revenues. In addition, online advertising has a significant advantage over other media, since it is technically possible to keep track of the effectiveness of advertising campaigns, and most importantly it is possible to target specific personas based on interests, habits, and so on. Although digital marketing can actually generate sales not only online, but also push sales in physical stores (related to this, the problem of attribution models is still an open issue). It is clear that with these tools it is easier to make a data-driven cost-benefit analysis. Revenues reports seem to confirm this power. In 2020 global advertising revenue is estimated to be almost 635 billion dollars (-1.5% to 2019), with a 104.1% revenues-expenditure ratio, but if we isolate the online advertising area it is reported a 116% revenues-expenditure ratio. Moreover, for the first time in 2022, online advertising will cover more than 60% of total adspend and is estimated to reach 65% by 2024.[4]

[4]Zenith, *Advertising Expenditure Forecasts. December 2021*, p. 7.

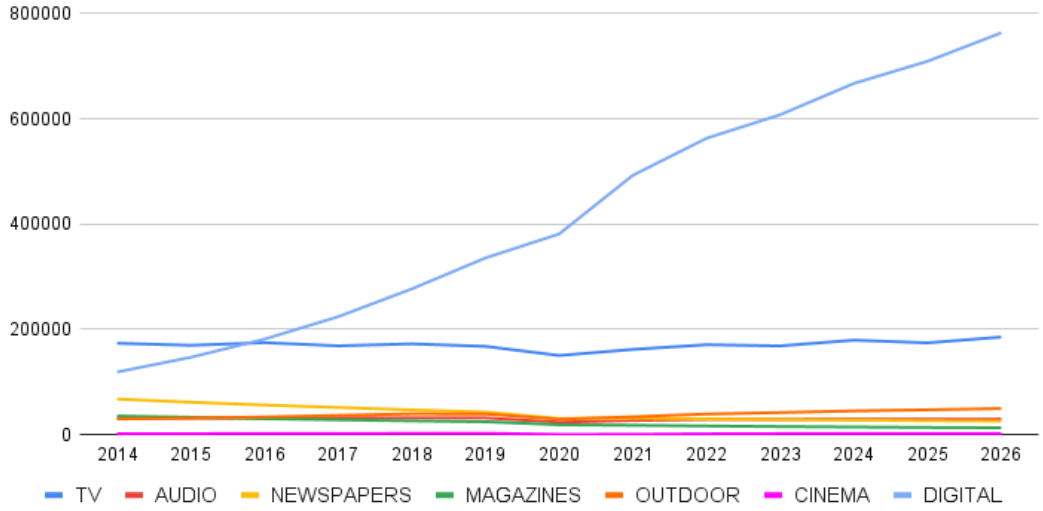

Figure 1.2: Global advertising revenues by medium - US$ million at current prices (GroupM, 2021)

1.2 Emerging Markets, Stable Markets, Hungry Markets

In the US, the year-on-year (YoY) 2019-2020 revenue growth deriving from advertising activities has increased by 12.2%, just a slight deceleration if compared with the 2018-2020 YoY 15.9% growth rate considering that "many organizations faced considerable revenue shortfalls in 2020 (e.g., airlines, hotels) and slashed marketing budgets and short-term advertising".[5] Another important piece of data that shows the resilience and capability of the American online advertising market is that " Ad revenues in Q4 [of 2020] were the highest to date, reaching$45.6 billion".[6] In other words, in the middle of the pandemic, the US had the best online ad revenue quarter of its history. According to this, in the past year, the US (+28.4% YoY '20-'21 growth in revenues of total advertising sector) is confirmed to be the more competitive and mature advertising market by far if we consider the total market volume On the other hand China (+18.8%) and especially the UK (+35.7%) show to be fastest-growing markets upon developed economies. According to the GroupM report, this growth is since "The U.S., the U.K., and China have all spurred substantial numbers of small businesses to increase their use of massive digital platforms. They have

[5] PwC, *Internet Advertising Revenue Report*, p. 3.

[6] Ibid., p. 11.

also been home to large numbers of suddenly big and rapidly growing venture-funded or early-stage publicly listed businesses".[7] The UK is the fastest-growing market among the biggest economies and this can be attributed to an advertising environment where well-funded startups are competing with each other to gain a prime position in their sector.[8] Of course, this rate is not sustainable in the long run, especially for one of the most developed countries in the World, so a progressive deceleration is inevitable. However, the UK makes up about 33% of the European digital ads market alone, and this is way beyond the runner-up (Germany) which accounts for just 15% of the market (with a much greater population and GDP). If the UK stands in the fourth position in the advertising expenditure rank, it overcomes Japan in the online advertising revenues rank.[9]

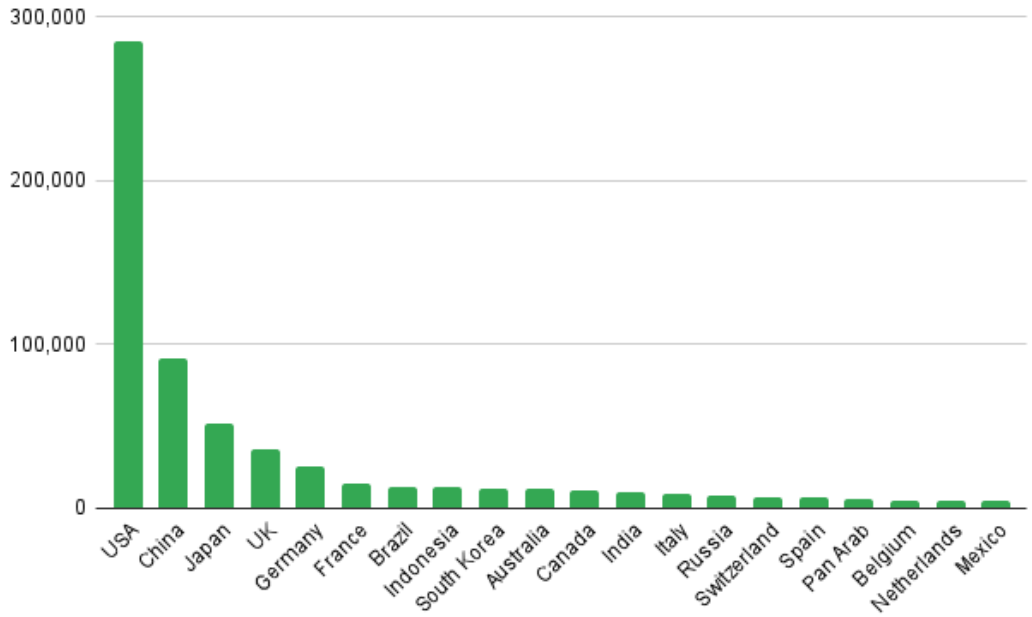

Figure 1.3: Top 20 countries by advertisers' major media expenditure - US$ million at current prices (Zenith, 2021)

In the European digital advertising market, there is another player that between 2019 and 2020, despite the pandemic, marked a vigorous growth. Turkey could still have room to grow, given the country's population size and considering that the per capita spending

[7] GroupM, *This Year Next Year. Global End-Of-Year Forecast. December 2021*, p. 14.
[8] Ibid., p. 17.
[9] Ibid., p. 13.

in online advertising is still substantially low compared to the European average. Turkey is up 33% YoY '19-'20 and it is by far the fastest-growing European country during the pandemic period. This trend seems to be exacerbated even more in 2021 when Turkey is forecast to grow up to 90%.[10] Ukraine (+19.2%) had the second-highest growth[11], while the average growth of European countries was +9.1%, and the European market in total grew by 6.3%.[12] At this point, should not be surprising that, despite the relatively small size of its market, in 2020 Turkey contributed up to 12% to the European growth, just behind Germany (23%) and UK (26%).[13]

1.3 Trends, Technologies and Emerging Formats

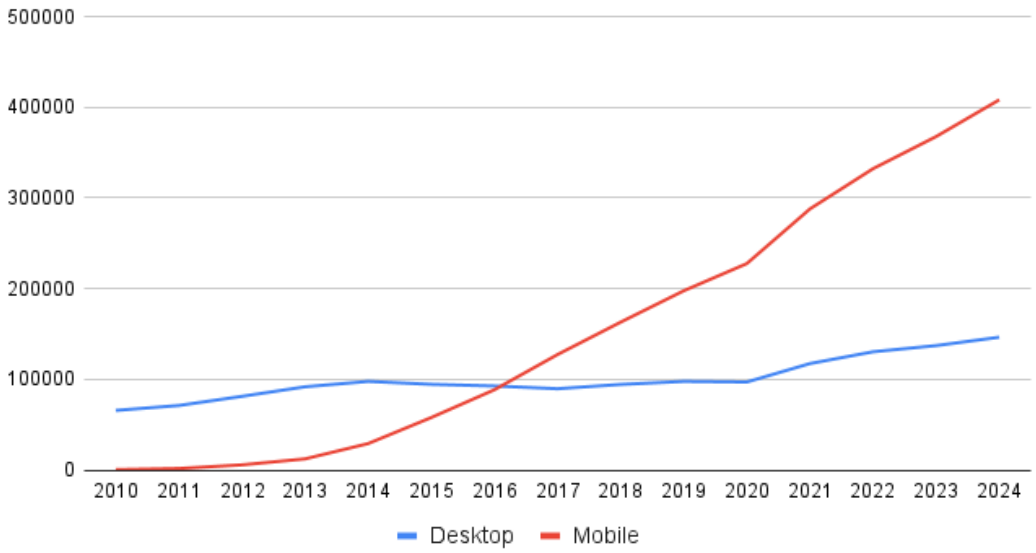

Figure 1.4: Global online advertising expenditure by device (Zenith, 2021)

Interesting points of reflection can be found if we dive more deeply into the trends of digital marketing and online advertising. If we take a look at the adspend for the device category,

[10] Zenith, *Advertising Expenditure Forecasts. December 2021*, p. 12.

[11] According to the IAB. Europe report, all the fastest growing countries in Europe in 2020 are Central-Eastern countries: Serbia, Czech Republic, Romania, Bulgaria.

[12] IAB.Europe, *Adex Benchmark Study 2020*, pp. 6–16.

[13] Ibid., p. 13.

it is clear that the impressive growth of online advertising is mainly driven by the mobile market, that since 2017 globally outperformed desktop advertising and now makes up more than 70% of the total and is projected to reach globally over 400 billion dollars of investments by 2024.[14] On the other hand, with a fast look at the fastest growing formats, the display advertising drives way more expenses than the other competitors, and the greatest part of this prevalence is due to the social media marketing being an almost total prerogative of the Display Adv format.[15] Online search advertising is still a flourishing sector, but the acceleration of display advertising is way more dramatic (figure 1.5).

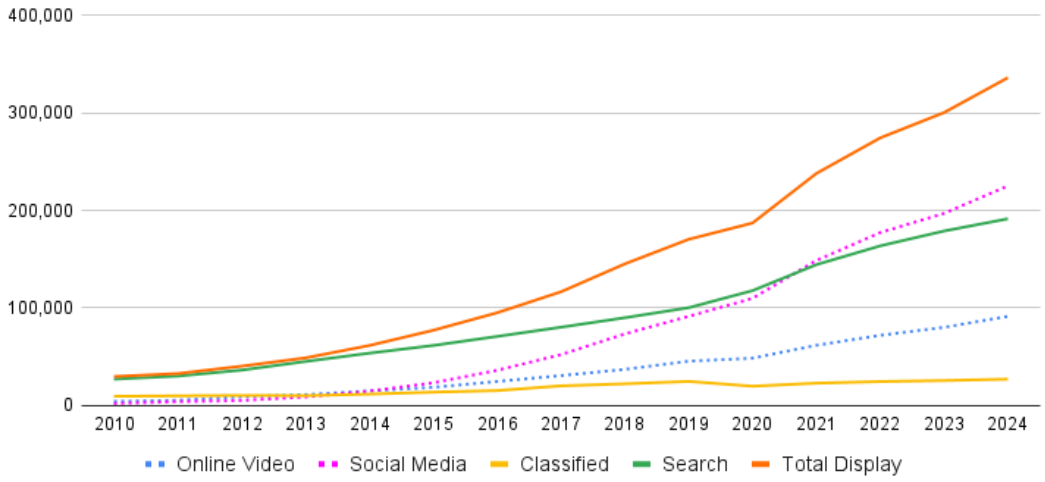

Figure 1.5: Global online advertising expenditure by format (Zenith, 2021)

It's no coincidence that the percentage of online advertising spend that comes directly from Social Media investments is growing consistently: while it made up just 22% in 2017, by 2021 it had reached 33% of the total, with a projection of 37% by 2026, which would equate to a total spend of over $250 billion[16].

This great growth takes place at the same time as an internal revolution in the world of social media, which are increasingly losing their Californian center of gravity and moving more and more towards China. As we have seen, the People's Republic has a healthy and growing advertising market, associated with the birth and development of new social media

[14] Zenith, *Advertising Expenditure Forecasts. December 2021*, p. 8.

[15] Ibid., p. 13.

[16] Statista.com, *Digital Advertising - Worldwide.*

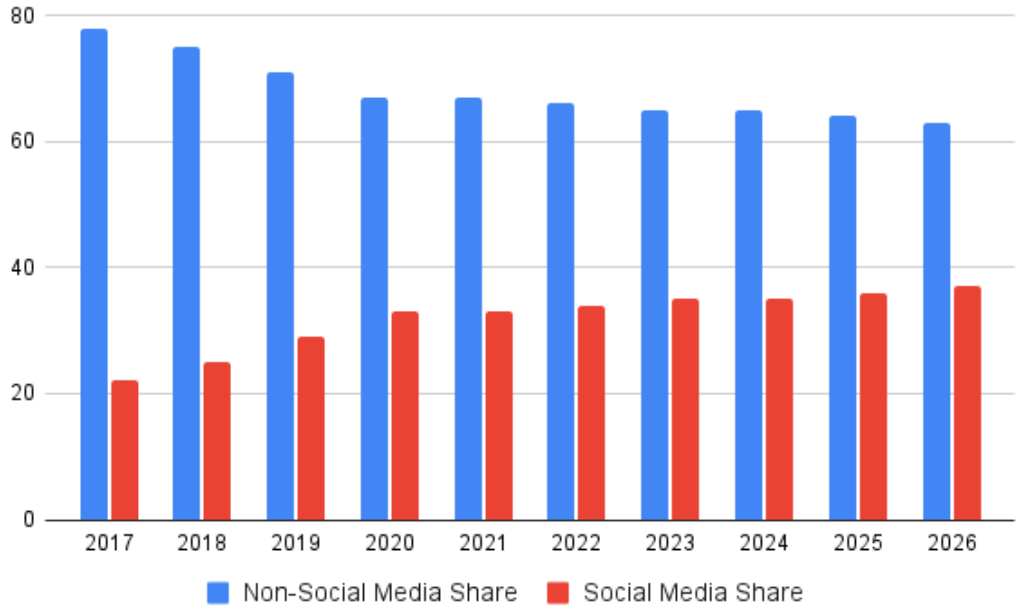

Figure 1.6: Social vs. Non-Social adspend share in online advertising (Statista, 2021)

with enormous international potential (TikTok first of all, but also WeChat, and Weibo, but the latter is still a Chinese national answer to Twitter). However, it is necessary to add for completeness that Western social media have a great difficulty to making inroads into the Chinese market due to severe restrictions (Google and Facebook are not present in the country) and the strict control of the government on user data. The latest American giant to give up on the Chinese market is LinkedIn, the world's largest job-search social media owned by Microsoft, which in October 2021 decided to leave the country, although other Microsoft services will continue to be available for the time being[17]. Interestingly, but not surprisingly, the trend already observed about the relationship between mobile and desktop does not change even when looking at social media: social desktop adspend remain pretty stable through the years, whereas investments on mobile grow constantly[18].

[17] LinkedIn.com, *LinkedIn is Leaving China!*

[18] Statista.com, *Digital Advertising - Worldwide.*

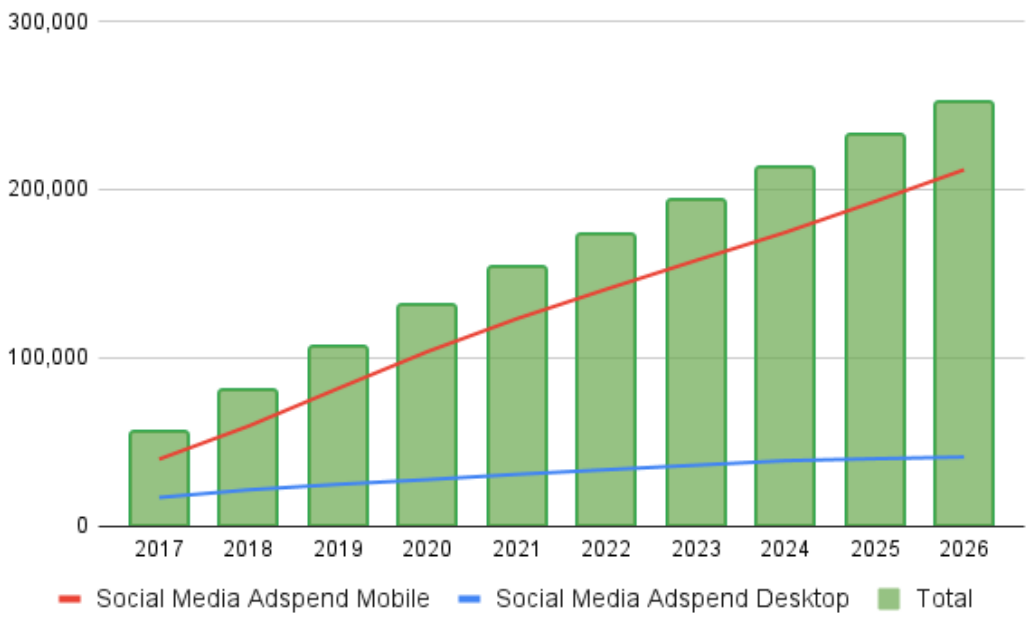

Figure 1.7: Desktop and mobile social media adspend with respect to the total adspend (Statista, 2021)

1.4 GDPR and Privacy Concerns

Given that a major revolution in the relationship between citizens and data collection has been underway in Europe for four years now, it is relevant to conclude with a small parenthesis on the European situation. As of May 25, 2018, the General Data Protection Regulation (GDPR) on privacy, data management, user rights came into force, and "it imposes obligations onto organizations anywhere, so long as they target or collect data related to people in the EU". The GDPR sets strict obligations on companies working with data of European citizens and is based on seven key principles: lawfulness, fairness and transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality, and accountability[19]. It is not the intent of this essay to go into the details of the regulation, but it is necessary to highlight that important changes have prospected.

Nevertheless, despite the recent pronouncement of several courts in European nations (France and Austria have issued very harsh opinions on American Big Corps collecting data on European websites) against how European users' data is handled, authorities and large digital

[19]GDPR.eu, *What is GDPR, the EU's new data protection law?*

companies are likely exploring all the possibilities to make their operations GDPR compliant. On the other hand, it is also possible (how France threatens), though rather unlikely, that the largest Web Analytics and Web Hosting service providers, such as Google, Facebook, Amazon, Microsoft, Apple, and many others, will be banned in Europe, at least until a solution that satisfies both parties can be found[20]. Facebook already threatened Europe about the possibility of interrupting its services due to strict privacy regulation[21], and Google started to discuss and test new solutions to create a more private, open, and accessible web[22]. Even if those efforts were recognized as a good start, they are still not deemed sufficient, at least by the french court[23]. Currently, Google's efforts seem to be mainly focused on overcoming cookie technology, aiming to eliminate third-party cookies, although initial proposals to achieve their elimination by 2022 have been delayed by at least a year[24]. In any case, the main French concerns come not from the granularity or quality of the data collected, but from where the data is physically stored and processed. The bone of contention is Chapter V of the GDPR, which discusses "Transfers of personal data to third countries or international organizations." The judgment of the French court suggests that the data of European citizens stored or processed in the United States are not fully protected, given the difference that persists between the U.S. government and European institutions on the subject of privacy[25]. The issue is still widely debated, and surely there will soon be numerous news that will inevitably impact the world of online advertising, especially in Europe.

1.5 Summary

In this chapter, the situation of the international advertising market has been analyzed, with a particular focus on online advertising, which is the direct scope of the subjects discussed

[20] Politico.eu, *French privacy regulator rules against use of Google Analytics*.

[21] Bloomberg.com, *Meta Renews Warning to EU It Will Be Forced to Pull Facebook*.

[22] Privacysandbox.com, *Building a more private, open web*.

[23] CNIL.fr, *Use of Google Analytics and data transfers to the United States: the CNIL orders a website manager/operator to comply*.

[24] Bloomberg.com, *Google Delays Phaseout of Advertising Cookies Until 2023*.

[25] Eur-lex.europa.eu, *CHAPTER V*.

in this essay. Initially, the most important macro-trends were highlighted: which media are concentrating their investments and revenues on them, noting the ever-increasing presence of the online, and which macro-areas are more developed from this point of view (first and foremost North America, led by the immense US market). It was noted that the pandemic has had little impact on this sector and, indeed, may even have been an accelerator for some processes already underway, especially in the more advanced markets. We have seen how the USA, UK, and China are the economically mature countries in greatest expansion from the point of view of analysis, although they are already largely developed, and we have mentioned the causes of this great growth. Next, we reported on the European case that sees Turkey and several Eastern European countries growing at important rates that momentarily don't stop, even though the English supremacy in Europe is still far from being close. Then we delved into digital marketing trends, observing how mobile devices have long been, and increasingly are, the biggest drivers of audience and advertising investment. In addition, it ha reported that display advertising is proving to incorporate much of the volume of the online advertising market, thanks mainly to social media (old and new). Finally, we concluded the chapter with a mention of the fact that the web data collection market in Europe is going through a period of transition that could make the management of European citizens' data by foreign companies operating on the web more complex and controlled.

Chapter 2

All you need is Bayes

This chapter will introduce the use of one of the basic theorems in statistics. To get to deal with the Bayes theorem a short path will be covered through some fundamental concepts of inferential statistics so that we can arrive at the formulation of the theorem with all the necessary tools to fully understand it. Sadly, the Bayes theorem is one of the most forgotten and mistrusted decision-making build blocks, despite having lots of profound implications in everyday life. It was conceived almost three hundred years ago, yet, it seems to have regained gloss only in recent decades, both for a growing interest in its advanced applications in Data Mining and Machine Learning, but also for the increased capability of computers. This is because, as we will see, Bayes Theorem presents critical issues for the complexity of the calculations that can generate, especially in models in which a large number of parameters are involved. Fortunately, this is not the case we are dealing with, so we can avoid worrying too much about this eventuality. However, it is necessary to take into account the limitations and the traps that we can run into to understand how far we can go with Bayesian Inference, which is nothing more than the practical application of Bayes' theorem for problems related to data, models, and parameters [1].

[1] The theoretical part of this chapter is supported by Martin, *Bayesian Analysis with Python: Introduction to Statistical Modeling and Probabilistic Programming Using PyMC3 and ArviZ*, Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, Wasserman, *All of Statistics: A Concise Course in Statistical Inference* and Everitt and Skrondal, *The Cambridge Dictionary of Statistics*

2.1 Sample Space and Random Variables

To understand the importance of the Bayes theorem, it is important to start with fairly basic concepts and then gradually add complexity and abstraction. Bayes theorem is a quite straightforward mathematical formulation, yet it hides some pitfalls. However, the most important thing to know is that it, despite its extreme simplicity, is the backbone of all Bayesian inference: there is not much else to know. Probably, all the complexity of the theorem lies in finding the right tricks to avoid running into unmanageable calculations, but the basic theorem remains the core of the process. For personal preference of the author, but also to get to understand the topics without a useless, for this essay, and excessive formalization of concepts, we will make use as limited as possible of mathematical notation. We will try instead to use many examples to get to the heart of the issues, glossing over the details, but we will formalize only those elements that will be essential for us to get to the end, without losing pieces.

The first concept that makes sense to introduce is that of $sample\ space$. The sample space is the complete set of all possible outcomes of an experiment. If I flip a coin once, the possible outcomes are head (H) or tail (T), because if I flip a coin once, I can only get heads or tails. If I flip a coin twice, on the other hand, the sample space becomes a list like the following $\{$H-H, H-T, T-T, T-H$\}$, depending on the order in which heads and tails appear in the sequence of flips (H-H means that in a sequence of two flips, the head came up in both flips). We can also think of sample space with physical quantities such as a person's height. In this case, the sample space is composed of all real numbers in an interval that can go from 0 (certainly the height of a person can not be negative) to $+\infty$. Obviously, the height of a person can not be infinite, as not even 100 meters, but there are no particular contraindications to consider the sample space much larger than it is in reality.[2] Connected to the concept of sample space are the definitions of $outcome$ (one of the possible results of an experiment - the set of all possible outcomes makes up the sample space) and $event$ (a subset of outcomes, or, depending on the point of view, a portion of the sample space).

[2]Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, p. 3.

At this point, it should be clear that a practical way to think about sample space can be to visualize it as an experiment that is repeated an infinite number of times, each time with an outcome. The set of all distinct outcomes of this infinite experiment corresponds to the sample space.

How can we practically deal with a sample space? How can we translate it into a manageable object? It is time to introduce the notion of a random variable. A random variable is the corresponding numerical value of the outcomes of an experiment. Why is this important? The outcomes of many events may not be numerical, such as the outcomes of flipping a coin (which are head or tail, instead of a number), or the outcomes of the weather forecast (tomorrow can be rainy, cloudy, or sunny, but it can't be 1, or π). For that reason it is important the concept of a random variable: assign a real number to each possible outcome of the sample space. Sticking to the previous example, if a coin is tossed n times, let X (random variables are usually expressed with the capital letters, whereas the value they assume is expressed with the same lowercase letter) be the number of heads in the sequence of outcomes that the experiment generated. So, it should be obvious now that the value of a random variable may change over the experiment repetition.

Another very neat definition of a random variable is given by The Cambridge Dictionary of Statistics which describes it as "A variable, the values of which occur according to some specified probability distribution".[3] This definition may be hermetic if you do not already possess the concept of probability distribution (which will be discussed soon), but on the other hand, in my opinion, it may be easier sometimes to start from the idea of a probability distribution, which seems to be (at least for the author) a more concrete and visible element, and then define the concept of a random variable which seems more vague and elusive. Now, random variables can be divided into two categories, with respect to the sample space they are mapping: discrete random variables and continuous random variables. A discrete random variable can assume a finite number of values (the number of heads can be 1, 2, or 10, but can't be 2.345, or $\sqrt{2}$), and a continuous random variable can assume an infinite number of values (the height of a person, theoretically, can assume every possible

[3] Everitt and Skrondal, *The Cambridge Dictionary of Statistics*, p. 356.

positive real number).

At this point, after the explanation of the concept of a random variable, it may still be unclear why this step is necessary for the path that is to lead us to the Bayes theorem. However, this rudimentary introduction to some elements of statistics is important in order to be able to abstract and generalize the concept of outcomes and events and then bring them to our specific case. In addition, these elements are important to introduce the notions of probability distribution, probability mass function, and probability density function. These conceptual steps are crucial because, subsequently, the parameters of the model will be treated as random variables, as well as their specific probability distributions, and not as fixed and immutable values. This change of perspective can be disorienting at first sight, but we will try to make it as painless as possible.

2.2 Probability Distributions, Mass, and Density

The probability distribution of a discrete random variable is just the list of each possible distinct outcome related to their correspondent probabilities. This definition is possible because in a random variable the space (the sample space, indeed) can be binned into a finite number of outcomes that are separated one to another (they are mutually exclusive). Sometimes, for simplicity, it is also possible to $discretize$ a continuous random variable, since it is much easier to work with a defined number of possible outcomes instead of an infinite one. This is one of the reasons why Bayes Theorem has been ostracized for many years: it can be very hard work in spaces that are composed of an infinite number of possible values.

Since we are dealing with discrete probability, it is possible to calculate the probability of each distinct outcome. For continuous random variables, it is possible to discretize the space and work with intervals, instead of single outcomes. The result will be a trade-off between the need of maintaining the original granularity of the space, and the capability to deal with a finite number of outcomes. In a discrete random variable, the probability of a specific outcome or the probability of an outcome inside a specific range (or bin) in continuous probabilities is called probability mass. The probability mass is the proportion

of the outcomes into a discrete space, and it is computed as the number of outcomes in the range over the total number of outcomes. For example, in the Figure 2.1, a probability distribution has been artificially generated (contemporary computers always reason in discrete terms, which is why it becomes complicated to compute complex integrals) starting from a normal distribution as a mean over 170 and with a standard deviation of 15. In the example of the image 1000000 values have been generated (which we remember to be a finite number anyway, even if quite large) and discretized in 30 intervals. The parameters chosen are rather random, but this distribution would likely approximate the height is distributed in the population, in centimeters. Now, if we wanted to calculate the probability of being tall between 155 and 185 centimeters, we would have to calculate the probability mass function of this interval. In this sense, the properties of the normal distribution also come to our aid: roughly 68% of the values fall within one standard deviation from the mean. Since the mean is 170, in the interval 155-185 we should find approximately 68% of the values. With a simple ratio, it is possible to find that in the distribution of the image in example 68.23% of the values, which is equivalent to saying that in our hypothetical probability distribution of the population height variable there is a 68% probability of being tall between 155cm and 185cm. One more important concept about the probability mass is that the sum of all the probabilities across the intervals must be equal to 1. This is quite intuitive because if we sum all the outcomes contained in each bin, we will get the total number of outcomes of the distribution. The formalization of this condition is that:

$$\sum_{x \in X} p(x) = 1 \tag{2.1}$$

Since in continuous distributions, the actual probability of a single outcome is almost zero, and the width of the intervals to discretize this function is arbitrary, there is a risk of creating discrete distributions from continuous distributions that fail to convey the specificity of the distribution. For this reason, for continuous distributions, it is preferable to speak of probability density. The density is nothing more than the mass, divided by the volume it occupies. Since we are reasoning in the context of a single random variable, the volume becomes the width of the interval, so the probability density is the probability of an outcome

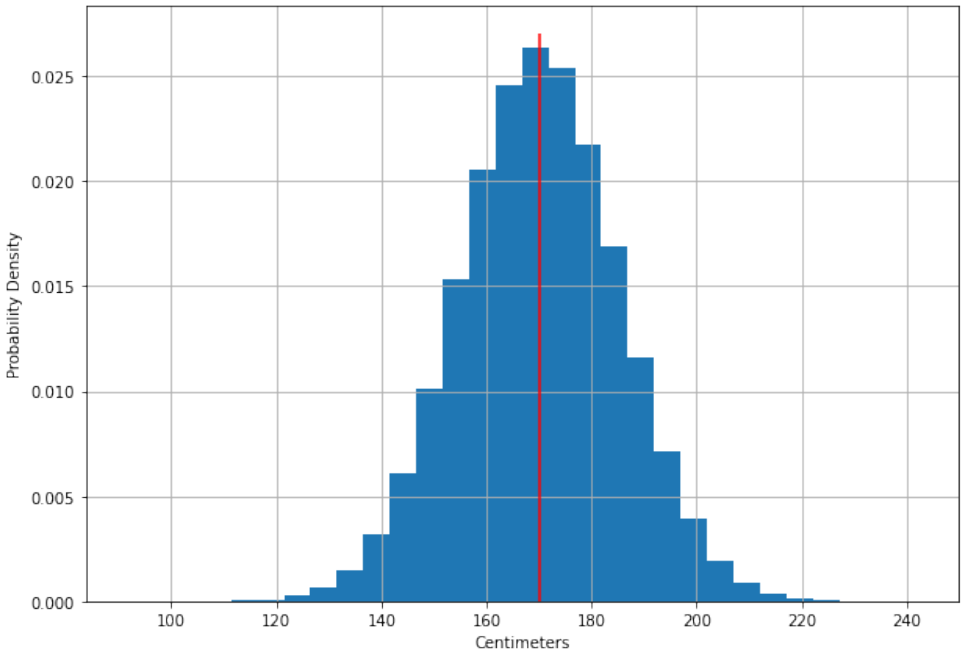

Figure 2.1: Probability density of a discretized variable generated from a $N(\mu$=170cm, σ=15). The red line is set to μ.

to fall in an interval normalized by the thickness of that interval. To summarize:

$$Probability\ mass\ of\ an\ interval = \frac{Number\ of\ outcomes\ in\ the\ interval}{Total\ number\ of\ outcomes}$$

$$Probability\ density\ of\ an\ interval = \frac{Probability\ mass\ of\ the\ interval}{Interval\ width}$$

In opposition to probability mass, individual probability densities can be greater than 1, this is because density is a measure relative to the specific interval. However, if we are dealing with a continuous distribution, the intervals are infinitesimal measures. This means that if we need to calculate the area under the curve described by the probability distribution we need to calculate the integral of the probability densities over all the infinitesimal intervals of the function. At this point, it is easy to show that the result of this integral is equal to 1. Mathematically this property is expressed as:

$$\int p(x)\, dx = 1 \tag{2.2}$$

Why should we care about probability density? If we want to skip the complications of dis-

cretizing a continuous variable, we can calculate the probability of belonging to a certain interval of the distribution by calculating the area subtended by that interval. In this way, we can avoid the arbitrariness of binning, but from a computational point of view, things become much more complicated: computing the space of a continuous interval means exploring an infinite number of points. Moreover, we usually have to work in a context in which we are not managing one variable at the time, but even tens or hundreds of variables (which can be equivalent to parameters in a complex mathematical model). In such an environment, the computational complexity becomes explosive, and the calculation becomes messy to manage.

2.3 Joint, Marginal and Conditional Probabilities

Let's consider two random variables with their respective outcomes. For sake of simplicity, let's imagine two discrete random variables, but the same reasoning applies to continuous ones. We may be interested in the relation between the outcomes of two separate random variables. If the two random variables belong to the same sample space (so in some way they can "communicate"), each outcome of a random variable corresponds to another possible outcome of the other random variable. The joint probability distribution is the set of all possible combinations of outcomes between the two random variables. If we want to calculate the probability that a specific outcome of a random variable occurs together with another specific outcome of the other random variable, we are computing the joint probability of those two specific outcomes. Formally, the joint probability of two discrete random variables is expressed as follows:

$$p(X = x, Y = y)$$

The same formula can be written as $p(x, y)$ with x and y in lowercase, indicating a specific outcome of the random variables. From now on in the essay, this more agile version is preferred. The comma tells us that the two outcomes have to happen at the same time, and that's really the focus of joint probability. The joint probability of two events A and B can

be written also as $p(AB)$ or as $p(A \cap B)$ which is the intersection between A and B. It may help if we consider the sample space as a set of all possible outcomes and we want to find the subsection that is common to two sample spaces.

Let us now consider that we only want to know the probability of a specific outcome of a random variable $p(X = x)$: leave out completely the random variable Y and make it somehow "collapse". This probability is called marginal probability and is calculated by fixing the outcome of which we want to consider and summing up the joint probabilities with all possible outcomes of the other (or other, in case I was on more than two dimensions) random variable. The mathematical formalization of this concept for a discrete random variable is:

$$p(x) = \sum_{y \in Y} p(x, y) \tag{2.3}$$

The image 2.2 represents a bivariate normal distribution composed by two normal distributions with $\mu = 10$ and $\sigma = 1.5$. The contour plot in the middle is the joint probability distribution of the two bivariate normals. The circles become more frequent where the joint probability density increases fast, and the small circle in the middle is the contour line that delimits the area around the mean of the two random variables (the densest part). The two pdfs at the right and at the top of the joint plot are the marginal probabilities of the bivariate distribution. In the current example, we are dealing with normal distributions that are continuous per definition, so in the process of marginalization, the summation of all possible outcomes becomes an integral, as follow:

$$p(x) = \int p(x, y) dy \tag{2.4}$$

It may help to visualize in 3D the plot of the joint probability distribution of the two normal distributions seen above. The joint probability of the figure 2.2 has a top view, but to fully appreciate the composition and the variation of the probability density (vertical axis) of a joint probability it may be useful from a 3D perspective. The grid in the plot can help to understand how the probability density of a random variable can change fixing an outcome

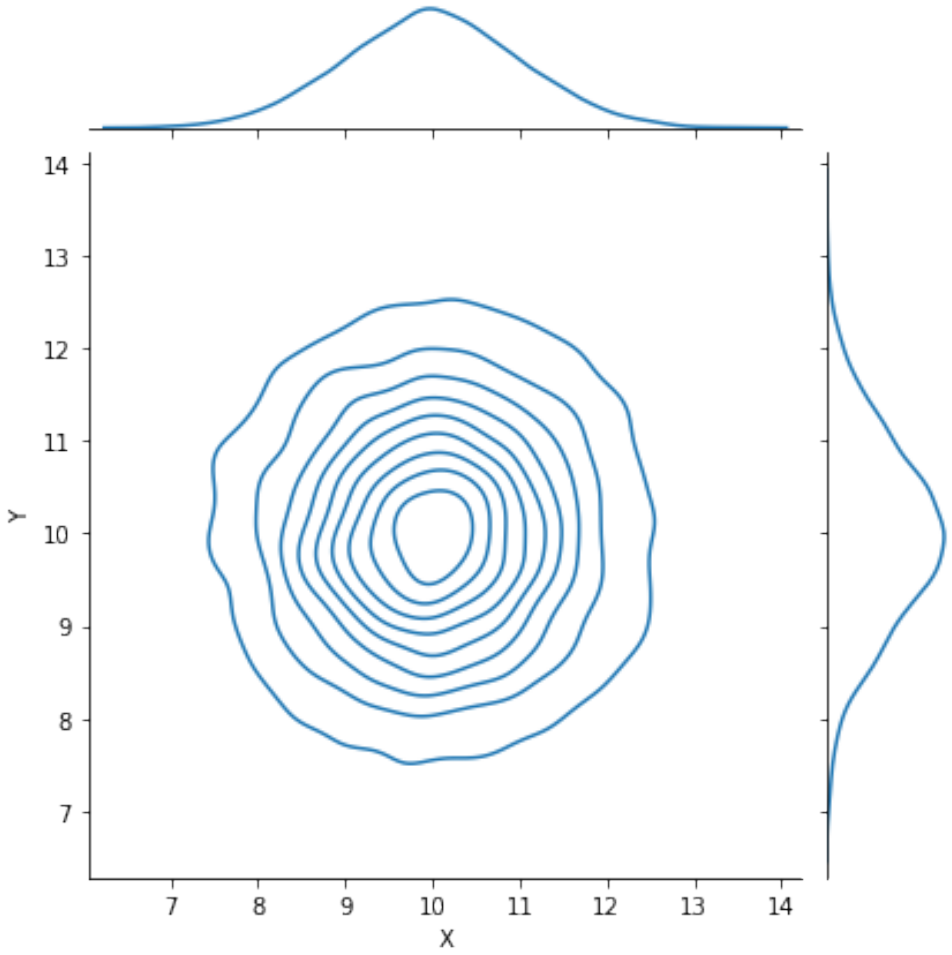

Figure 2.2: Joint probability distribution of a bivariate $Normal$ distribution $\mu = 10, \sigma = 1.5$.

(or an interval) of the other random variable. Indeed, coming back to the discrete variable scenario, we may be interested in knowing the probability of an outcome of one variable, fixed (given for certain) the outcome of the other variable. For example, we might want to know what is the probability of drawing a king from a deck of cards, given that the cards are spade suit. In practice, we want to calculate the probability of the event (Card Figure=King), conditioned by the event that (Card Suit=spades). Mathematically speaking the conditional probability of an event A $given$ event B is defined by the equation:

$$p(A|B) = \frac{p(A,B)}{p(B)} \tag{2.5}$$

However, may happen to be in the presence of two independent events A and B. In the

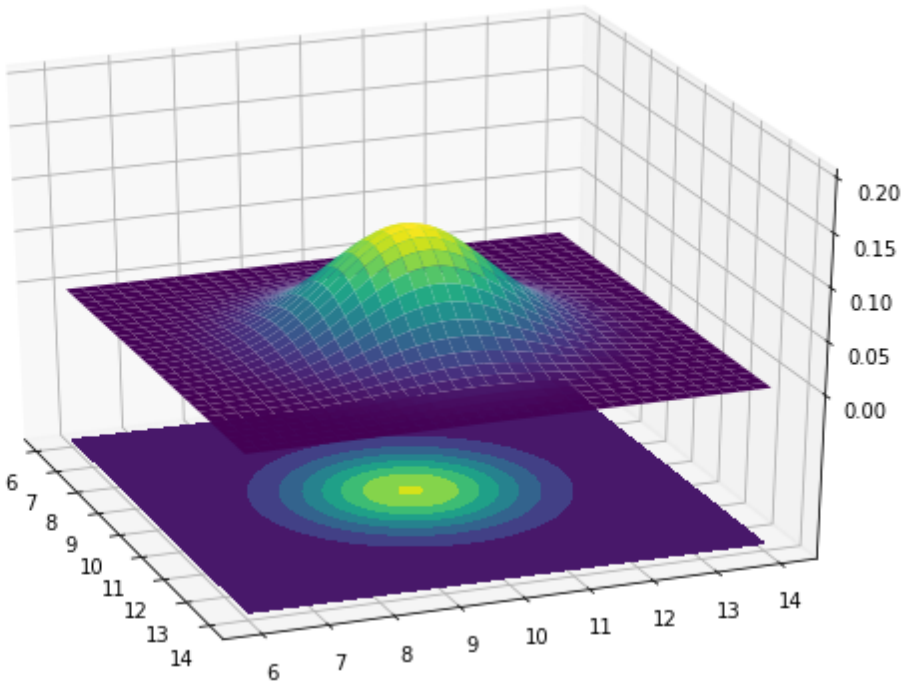

Figure 2.3: 3D view of a joint probability distribution of a bivariate normal distribution $\mu = 10, \sigma = 1.5$.

case of independence of events only, the formulation of the joint probability of two events A and B is the following $p(A, B) = p(A)p(B)$. For example, if I open a book twice, the probability of getting an odd page number on the first try and the second, are independent, since the two outcomes are not influencing each other. That means that in the case of $independent\ events$ the equation for the conditional probability becomes:

$$p(A|B) = p(A)$$

$$p(B|A) = p(B)$$

Let's put aside independent events and pretend we are dealing with non independent events. Rearranging the equation of conditional probability, we get that $p(A, B) = p(A|B)p(B)$ but, at the same, since $p(A, B) = p(B, A)$ (because we are calculating the probabilities of

each possible pair of outcomes) we also get that $p(A, B) = p(B|A)p(A)$. Now, we just put together the two equations that are equal and we get that $p(A|B)p(B) = p(B|A)p(A)$. Before you know it, we have derived the famous Bayes theorem:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2.6}$$

We have preferred to refer to events, rather than outcomes, but the argument does not change much in the case of discrete random variables. The formulation of the Bayes theorem can be formalized as:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{2.7}$$

We know the definition of conditional probability, so we can add one more piece to the formula of the theorem. We have already seen how for a discrete random variable the marginal probability is computed as $p(y) = \sum_i p(y, x_i)$, and from the formula above $p(y, x) = p(y|x)p(x)$, it is possible to extend the marginalization process as $p(y) = \sum_i p(y|x_i)p(x_i)$ and the extended formula of Bayes theorem for discrete random variables becomes:

$$p(x|y) = \frac{p(y|x)p(x)}{\sum_i p(y|x_i)p(x_i)} \tag{2.8}$$

The formulation for a continuous random variable is quite the same, but it is important to treat it briefly. To imagine the conditional probability in a continuous variable, we should get back to Figure 2.3. In conditional probability, we want to condition the probability of a variable $given$ an outcome of the other variable. Since we are dealing with continuous variables, we can't speak of outcomes anymore, but imagine an infinitesimal interval as an outcome in a discrete variable. In substance, it is like if we cut the plot of the joint distribution along a function (that corresponds to an infinitely small interval) of the variable X. Doing so we define a pdf that will still have the profile of a normal distribution and that will express the density of the variable Y, conditioned to a function of X along which we have "cut" the joint probability. Once we get this pdf, it is easy to explore all the possible densities of the conditioned X variable along the new pdf. The main point here is that the

conditional probability of a continuous random variable is not an outcome, but a density function. The formalization of this peculiarity is the following:

$$f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx} \tag{2.9}$$

2.4 Posterior, Likelihood and Prior

Bayes' theorem has been seen in its most general form, but it is worth analyzing the elements that make up the formula to better understand how it has to be interpreted, and how to fully understand its potential. Let's initially talk about the small bar that defines conditional probability. We have seen that its meaning is to impose that the element to the left of the bar is conditional on the value to its left. We have seen that $p(A|B)$ stands for the probability of A given B: what is the probability that it will rain GIVEN THAT it will be cloudy tomorrow? Well, we saw that $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$ and that therefore $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$ and that $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$. The direct consequence of this formulation is that $p(y|x) \neq p(x|y)$, or at least the two conditioned probabilities, in general, are not equal. But what does it mean in practice? If we have to estimate the probability that tomorrow it is going to rain, given that will be cloudy, it is reasonable to think a probability around 60%: $p(rainy\ day|cloudy\ day) = 60\%$. Now, consider the other conditional probability: $p(cloudy\ day|rainy\ day)$. We can reasonably say that the probability that it is cloudy on a rainy day is much higher than 60%, we could say it is almost 100%: even if I'm not an expert of meteorology I'm quite confident to say that rain usually comes from clouds and rainbows are not so common events. This trivial example shows how the two conditional probabilities are not equal at all, even though sometimes we are prone to think it.

At this point, we found out that the two conditioned probabilities in the Bayes theorem are not the same. Pushing one step forward it is possible to define the conditional probability at the left of the equation as $posterior\ probability$, and it is actually the probability we are interested in, the information that we want to infer from the data, and thanks to our

prior knowledge. Data and prior knowledge: it is time to abstract Bayes' theorem from the condition from which we started (talking about events), to the formulation in which we are interested. Let's get to the heart of Bayesian inference.

We have discussed random variables which are nothing more than the numerical mapping (or mathematical formalization) of all possible outcomes of an experiment. Now, let's imagine that we want to study a phenomenon: we have to formalize the mathematical model that underlies the phenomenon structure. We need to know the structure of this mathematical model that is given by its parameters. But how can we formalize a phenomenon? First, we have data available. We know their values (we hope they are many) and we can treat them as outcomes of a random variable. We call the random variable which refers to the data as D. But how can we introduce this random variable in the model? We must assume that behind the data there is a process, more or less ordered, that regulates them. In an extreme case, if the outcomes are randomly shaped, we can assume that all events in this process happen with the same probability: this is licit, but somehow we have to assume that there is $something$ behind it. This assumption is the distribution of the data $p(D), knowing\ nothing\ about\ the\ parameters$! Knowing nothing about the parameters means that we have to take into account all the possible values that parameters can assume. $P(D)$ is called $marginal\ likelihood$ or sometimes $evidence$. However, this element is not crucial in the theorem, because it does not depend on the parameters. In the Bayes theorem, the problem of data distribution is not approached head-on, so, in practice, we don't need a solid definition of how data are described by a pdf, but the idea that behind data there is a generative process is still valid. Now, we can set the data distribution aside for the moment, but I just wanted to point out that the marginal likelihood is the probability (or the density function in the context of continuous random variables) of the data taking into consideration all the possible values the parameters can assume. In fact, in bayesian inference, we normally want to know the features that define the model (its parameters), given the data (conditioning the probability of the parameters on the data values we know), known as posterior probability. So, we are mainly interested in the parameters GIVEN the data, not in the processes behind the data alone. Since the parameters are usually defined with the Greek letter θ, the posterior probability from this perspective of Bayes theorem

centered on the inference of parameters given the data is: $p(\theta|D)$

We miss talking about the other conditional probability in the equation, the one that is normally called $likelihood$: $p(D|\theta)$. The likelihood is something that needs to be defined in the design phase of the model and it can be seen as a section of joint probability. The joint probability between two random variables can take a rather peculiar form since it is the intersection of two probabilities. Therefore, sometimes the definition of the likelihood (at least in the author's brief experience) can be quite fuzzy. In some specific cases, we have distributions that are just right for us (like the binomial in the case of the coin toss that we will see later in an example), other times we have to be guided by our experience and intuition, hoping to get lucky. Sometimes we can understand to be wrong. Fortunately, when the mistake is gross it is usually quite evident. Other times, however, we can cope with a mistake and have acceptable results, if the priority is particularly precise. Sometimes, assuming that the likelihood follows a normal distribution may be acceptable, other times a t-student distribution may be more appropriate, due to its long tails. Once again: there is no single possible solution, also because usually the reality is not what we expect it to be.

At this point, the logic of the theorem should be clear enough: the likelihood is the probability of observing the data we have, conditioned on the possible values the parameters can assume. In a sense, it is the plausibility of the data as the parameters vary. In fact, in the absolute easiest scenario, the one in which we have to extract the posterior probability of a single parameter, it is useful to remember that we have to treat the data and the parameter as two random variables. Let's recall what is seen in Figure 2.3, and assume that it is the joint probability of data and parameters $p(D, \theta)$. The likelihood has the task of defining the probability of the data, given a possible value (or an infinitesimal interval of values) of the parameter. It is clear that the closer we get to the most realistic value of the parameter, the higher the likelihood will be. But now the problem is right here: how do we say what is the $most\ realistic$ value for the parameter if the parameter is defined by ourselves? In practice, we only have the data and we are trying to design a descriptive model behind the data, so we have no "empirical" information about the parameter. This is where our experience, deductive ability, and a good amount of luck must come to the rescue.

Some assumptions are needed about how the distribution of the parameter can be described. We need to establish its prior probability, that is, its theoretical distribution independent of the data. We must define how the random variable is distributed. There are cases in which we are luckier: we already know what boundaries this value can take. There are other times when we do not know anything or we know very little about the behavior of a certain parameter. In such a case it is convenient to choose a prior probability more "flat" and less informative. As mentioned before it is also possible to say: the prior probability of a parameter is uniform on all its support, so let's choose a uniform prior distribution! This is possible, but rarely advisable, for at least two cons: $i)$ In such a scenario we give all the power to define the posterior probability to the likelihood, and if the likelihood is not particularly good, we fall into surely wrong inference $ii)$ If we do not know anything about the parameter, it will also be difficult to establish its support. In fact, even the uniform distribution has parameters (the boundaries) that define the maximum and minimum value that can take. Generally, if you don't know anything about the parameter you should choose long-tail distributions, in the sense that even if they have a higher density in some points of the distribution, they decrease very slowly. In Bayesian inference is a good habit to avoid the zero probability situation: even if an event is extremely rare, never assign a zero probability to it. For example, if we know that the parameter can have both positive and negative values, but we don't know much else, we generally opt for a normal distribution with a very high standard deviation. Of course, this means very little without a context, but it is just to keep in mind that usually there are distributions more usable than a flat uniform. Everything has to be adapted to the context, in Bayesian learning: there are no ready-made solutions.

We have already briefly discussed marginal likelihood, the $p(D)$ at the denominator of the Bayes theorem. We have mentioned that it is a secondary element of the theorem, but we have not explained why. Recall that the formula for computing marginal likelihood in a continuous random variable is: $\int f(D|\theta)f(\theta)d\theta$. Since we are integrating over θ, $p(D)$ is not dependent on θ. This is implies that the posterior probability $p(\theta|D)$ is $proportional$ to the $likelihood$ times the $prior$. $p(D)$ is sometimes also called $normalizing\ constant$ since it has not a prime role in the calculation. Formally this property is expressed as:

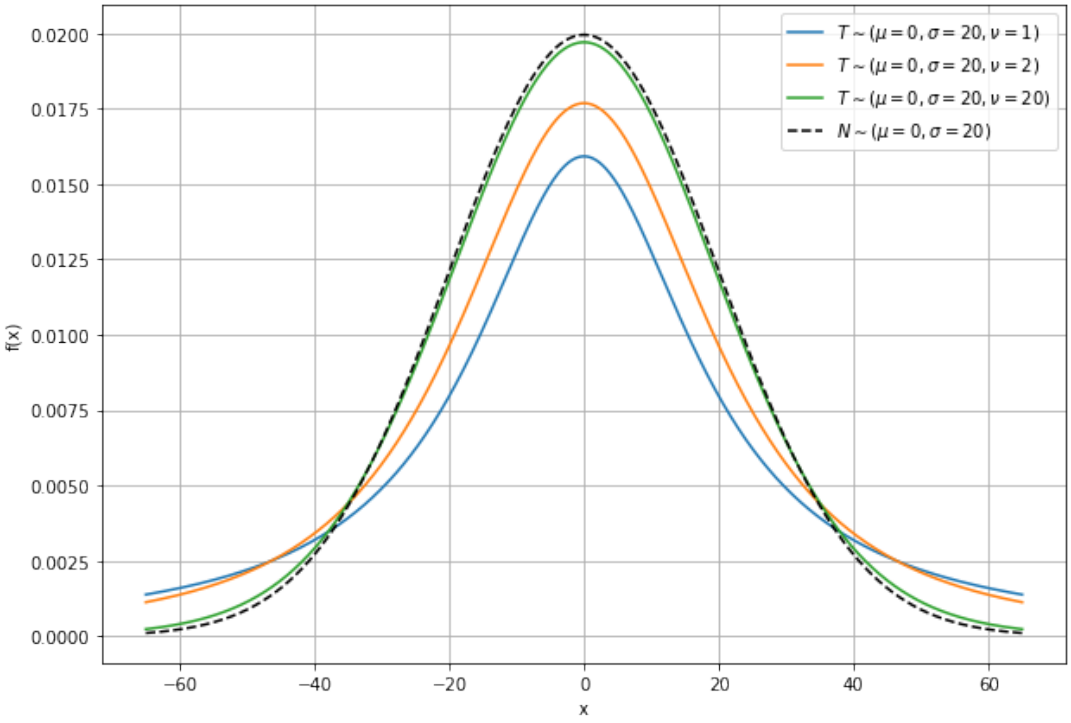

Figure 2.4: Differences between three $Student - t$ distributions with different degree of freedom and a normal distribution with same μ and σ.

$$p(\theta|D) \propto p(D|\theta)p(\theta) \tag{2.10}$$

This property helps us to imagine the conditioned probability as a "compromise" between the likelihood and the prior probability. The marginal likelihood, however, although not central to the calculation, is the element that complicates things. Calculating that integral can be an incredibly complex thing when we are in a context with tens, hundreds, or thousands of parameters. The space over which to have to integrate becomes enormous and the complexity is explosive. That's why more or less efficient methods have been invented to avoid calculating that integral, through approximation processes. These approximations are not always very successful, and it is still a rather prolific area of research, but thanks to these processes (sometimes called inference engines) we have been able to develop probabilistic programming languages such as PyMC3, the Python library we used for the example discussed in this essay.

2.5 The Coin Flipping Example

This section is aimed to report a very simple, but explanatory example[4]. It can give a simple indication of how appropriate likelihoods and prior probabilities can be chosen, and in general is a well-known example due to its simplicity. Consider the classic coin example where we are in a context where a coin is flipped n times and we have already obtained a set of outcomes (heads or crosses). Our goal is to know what is the $bias$ of the coin, given the outcomes of the tosses. The coin's bias means how much the coin is unbalanced with respect to a 50% probability: if the bias is 1. the coin toss will always give us a single outcome, if the bias is zero, it will always give us the opposite outcome, if the bias is 0.5 there will be maximum uncertainty between the two possible outcomes. In summary, posterior probability must determine what is the probability that the coin is as we expect it to be (in principle with a bias of 0.5), given the results we have observed so far. What do we do? First, we try to establish the likelihood. We are clearly facing a discrete scenario because the coin can be flipped 1, 2, 3, N times, but not π times. We know that likelihood should inform us about the plausibility of the data, given a parameter. We assume that each toss is independent of the other and that all tosses are part of the same distribution. Making these reasonable assumptions a good candidate may be the binomial distribution which is: $i)$ A discrete distribution $ii)$ It describes the distribution of the numbers of successes in a sequence of consecutive trials, each of which is a single (independent) Bernoulli experiment. This all seems pretty reasonable. This distribution is formalized as:

$$p(y|\theta, n) = \frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y} \tag{2.11}$$

In the notation used above y is the number of successes (the data, in our environment), n is the number of trials and θ is the probability of success in a single trial. This is the parameter that we want to model through the prior probability since we want to estimate

[4]This example is taken from the book of Osvaldo Martin about Bayesian Analysis with Python (Martin, *Bayesian Analysis with Python: Introduction to Statistical Modeling and Probabilistic Programming Using PyMC3 and ArviZ*, pp. 22–30)

which is the bias of the coin, once we have seen the data. In general, in Bayesian statistics every time we don't have precise information about a parameter, let's put a prior on it. How we can choose the prior? We know that this prior describes a probability, so it can embeds values between 0 and 1($p \in [0, 1]$). This time it can assume values continuously in the support range, so we need a continuous distribution. Seems that the $Beta$ distribution fits perfectly! The mathematical formulation of this distribution is the following:

$$p(\theta) = \frac{1}{B(\alpha - \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{2.12}$$

The first component of the Beta distribution is called the Beta function. It is a constant that is θ independent and has the purpose of keeping the final probability equal to 1. For that reason, it is not such useful to express the full formula, but for completeness, we report also the formulation of the Beta distribution with the Gamma function $\Gamma(z)$ which is strictly related to the Beta function, and sometimes it is preferred[5]:

$$\frac{1}{B(\alpha - \beta)} = \frac{\Gamma(\alpha + \beta)}{\Gamma\alpha + \Gamma\beta} = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}$$

At this point, we have all we need and it is useful to recall that the posterior distribution is proportional to the likelihood multiplied by the posterior: $p(\theta|D) \propto p(D|\theta)p(\theta)$. This property in our setting is expressed by:

$$p(\theta|y) \propto \frac{n!}{y!(n-y)!} \theta^y (1-\theta)^{n-y} \frac{1}{B(\alpha - \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{2.13}$$

Proportionality is maintained even if we remove the constant elements that are independent of θ, so a posterior probability would be like:

$$p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \tag{2.14}$$

At this point, if we pay attention, it is easy to spot that the posterior probability is proportional to something that has the shape of a Beta distribution, without the constant normalization. Since the constant does not depend on θ we can assert that the posterior distribution of our problem is proportional to a Beta distribution with the following formulation:

[5]Encyclopediaofmath.org, *Beta-function*.

$$p(\theta|y) \propto Beta(\alpha_{prior} + y, \beta_{prior} + n - y) \tag{2.15}$$

2.6 The Bayesian Process

After the general overview we gave in this Chapter, associated with practical examples, it should be quite clear the mechanism behind Bayesian inference. Let's recap what are the four basic steps of Bayesian inference (they are often summarized into three steps, but in this context, a more "algorithmic" formulation is preferred):

- Choose a likelihood $p(D|\theta)$ that reflects our beliefs about how the data can behave, given the parameter status θ
- Choose a prior probability $p(\theta)$ for all that elements that we want to estimate. This prior should reflect our beliefs about the parameter before any data enter the game
- Compute the posterior distribution $p(\theta|D)$ conditioning the model to the data we observed
- Ensure that the predictions built from the proposed model make sense with respect to the data, and possibly try new models that are more capable of describing the data generation process

However, it is worth making some final points about the Bayesian $philosophy$, its advantages, and where it is likely to fall. Probably, the most controversial element of Bayesian statistics is the concept of prior probability that, at times, can be rather vague and risky to adopt. In the examples that were brought up, it was fairly easy to make predictions about prior probabilities, but it may well be that this is not the case. Sometimes, we need very poorly informative distributions and this can be a problem for the posterior distribution. It may happen that we don't know anything about a prior, so choosing any prior, even the less informative on Earth, could be problematic because we are just walking in the dark. We have already pointed out that a less informative distribution gives more responsibility to the likelihood, and in some cases, it can lead to errors, or, in general, it nullifies the power of the Bayesian approach since it converges with a frequentist approach:

$p(\theta) \propto p(D|\theta)p(\theta) \propto p(D|\theta)$.[6] Moreover, mistakes may occur even when choosing the likelihood, and this probably generates even greater errors than a wrong prior: "Bayesians are slaves to the likelihood function. When the likelihood goes awry, so will Bayesian inference".[7]

It is important not to consider Bayesian statistical inference as the cure-all, since it is based on the principle of combining a certain degree of prior knowledge with the available data. If we have little data, and we are confident enough in our beliefs about the model structure, then it may make sense to use Bayesian inference to extract something meaningful from what is available. If, on the other hand, we have a lot of data, but little knowledge about prior probabilities, it is probably best to opt for a frequentist approach. If we have neither prior knowledge nor enough data, we need a lot of luck. In general, it is worth remembering that with large samples and under regularity conditions the results provided by Maximum Likelihood Expectation (frequentist approach) converge with the results of the Bayesian inference.

2.7 Monte Carlo Markov Chains and Metropolis-Hastings

We pointed out that the posterior probability can be difficult to compute for the integral at the denominator which may risk being intractable. For that reason, many clever methods that avoid the integral have been developed. We are going to treat the family of Monte Carlo Markov Chain algorithms, and in particular one of them, the Metropolis-Hastings algorithm[8], which is the standard and more general member of the family. We prefer this set of algorithms because the Python library we are going to use for the framework implementation is specialized in this kind of solution.

Let's first talk about the naming: Markov Chain Monte Carlo (MCMC). Monte Carlo methods are known to be inspired by the casinos, and the name is clearly related to this. This approach is the following: if I cannot know, or if it is difficult to compute, the probability

[6] Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, p. 182.

[7] Ibid., p. 189.

[8] Chib and Greenberg, "Understanding the Metropolis-Hastings Algorithm".

distribution of an event I can build this distribution by randomly sampling from the outcomes. In the long run, I will have a satisfactory approximation of the real distribution. Since the MCMC family deduces the posterior probability distribution by sampling from the posterior probability, we are using a Monte Carlo method. Moreover, the sampling is performed sequentially by a semi-random walk through the parameter space, where the position of the sampled draw depends only on the position of the previous draw. For that reason are using a Markov Chain: every link in the chain (a sampled draw), is bounded only to the previous link.

How can we sample in practice from a posterior distribution? The Metropolis-Hastings algorithm comes to help! Remember that in order to approximate the true posterior probability the sampling process is iterative, and usually, a great number of draws is required to adequately build a posterior distribution.

1. Initialize a random point x_0 in the parameter space (in Bayesian inference x_0 is a possible value of the parameter)

2. for $i = 0$ in $N - 1$:

   (a) Sample $u \sim Uniform(0, 1)$

   (b) Sample $x^* \sim q(x^*|x_i)$. $q(x^*|x_i)$ is called proposal distribution and it is just a perturbation of x_i following some distribution, such that $x^* = x_i + \epsilon$[9]. $q(x^*|x_i)$ is called proposal distribution and it is just a .

   (c) Compute the acceptance rate $A(x_i, x^*)$. The Acceptance rate is nothing more than the probability of acceptance of the proposed new step of the random walk.

$$A(x_i, x^*) = min\left(\frac{p(x^*)q(x_i|x^*)}{p(x_i)q(x^*|x_i)}, 1\right) \tag{2.16}$$

[9]This ϵ can be considered as some noise to add to x_i. It could be a value sampled from any distribution, but let's imagine it is form a $Normal \sim (0, \sigma)$. The greatest is the σ, the greatest will be the amplitude of the random walk: the walk will go more frequently at the tails of the distribution, and will visit less probable values of the parameter. For that reason for too great σ, many of the proposed steps will be rejected, but if σ is too low, the posterior distribution may not grasp a complex posterior shape, and the sampling process will be too approximated

(d) if $u < A(x_i, x^*)$ *then* $x_{(i+1)} = x^*$, else $x_{(i+1)} = x_i$

N is the number of draws we want to sample from the posterior distribution[10], while in the Equation 2.16 $p(x^*)$ and $p(x_i)$ are the posterior probabilities of the current value x_i and the proposed value x^*. They are reported is short here, but $p(x^*) = p(x^*|D)$, and $p(x_i) = p(x_i|D)$. However, since we are calculating a ratio of posteriors and the normalizing factor (the denominator of Bayes theorem) is the same for both of them (the data are the same, they have the same $p(D)$), we don't need to compute the marginal likelihood just by rearranging the equation:

$$\frac{p(x^*)}{p(x_i)} = \frac{\frac{p(D|x^*)p(x^*)}{p(D)}}{\frac{p(D|x_i)p(x_i)}{p(D)}} = \frac{p(D|x^*)p(x^*)}{p(D)} \cdot \frac{p(D)}{p(D|x_i)p(x_i)} = \frac{p(D|x^*)p(x^*)}{p(D|x_i)p(x_i)} \tag{2.17}$$

Working with ratios makes the trick and we can rid of the troublemaker $p(D)$. Moreover, the factors $q(x_i|x^*)$ and $q(x^*|x_i)$ are normalising factors that we need to correct the asymmetry of the proposal distribution (in the case it is an asymmetrical distribution).[11]

In summary, the Acceptance rate is the probability of moving from x_i (the current state, present link in the chain) to x^* (the proposed state, next link in the chain). If the proposed value is accepted the random walk moves toward it, otherwise it remains at the same value. Now just consider the ratio $\frac{x^*}{x_i}$, leaving apart the normalization factors. If $p(x^*) > p(x_i)$, then $\frac{x^*}{x_i} > 1$, and $min\left(\frac{x^*}{x_i}, 1\right) = 1$: the proposed position is always accepted. On the other hand, if $p(x^*) < p(x_i)$, the proposed position will be accepted only probabilistically (since u is randomly sampled from a uniform distribution).

Following this logic for the random walk, in the long run, the histogram produced by all the positioning records of the chain will produce an approximation of the parameter posterior probability. This example is made with just one parameter, but imagine that it can be repeated with n different parameters and the logic would be the same, but the space to explore would grow exponentially. For that reason, the Metropolis-Hastings algorithm

[10] The more iterations we have, the better, but the operation is quite expensive and there are normally time limits to be met.

[11] Gelman et al., *Bayesian Data Anlysis*, pp. 279–280.

is very efficient for models with a relatively small number of parameters, and many other smarter methods have been developed for more complex models, such as NUTS. This topic is beyond our scope, but we will briefly touch NUTS in the next chapter.

2.8 Bayes and Cognitive Biases

The Bayes theorem, in its simplicity and elegance, remains one of the most easily forgotten theorems. The reality is that in the decisions we have to make every day, applying mathematical theorems, even the most trivial, is not easy and is not spontaneous. In problems where Bayesian statistics is more hidden and less visible, even people with a solid statistical base can fall into the temptation of answering complex questions instinctively, following their intuition. Let's say that, in general, as much as possible, human beings try to replace rational statistical calculations with heuristics that can be useful to quickly solve everyday problems. Unfortunately, as already anticipated, often the likelihood and the posterior probability are confused. For example, it has been demonstrated a long time ago that by using the heuristic of representativeness (the probability of an event is established based on the salient characteristics of the groups it belongs to, without considering all other frequencies) we usually skip the difficult part of calculating probabilities.[12] For example, knowing that a person named John is shy, introverted, a bit clumsy, and eccentric, but also smart, hard-worker, and usually rational and methodic: do you think that John is more willing to be an economics student or a mathematics student? Well, if your answer is mathematics, you used the heuristic of representativeness to solve this small problem.

In reality, we don't have any information about the academic history of john, but we do know something about him that is not particularly related to his field of studies or his academic path. The same characteristics can be easily applied to a law student as long as a physician. So, given the data that we have, we should rely only on priors, and of course, the prior probability of being an economics student is much greater than being a math student. On the other hand, when we are prone to answer mathematics, our brain makes fit the description it is given with its representation of a math student and with the image

[12] Tversky and Kahneman, "Subjective Probability: A Judgement of Rapresentativeness", pp. 446–449.

of an economics student. The representation of a math student seems to match better, so we instinctively answer that the student is a mathematician, but that is not the most reliable answer. It follows that the prior probabilities are generally ignored, and this can lead to significant errors. Interestingly, it has been observed that usually, we return to use prior probabilities only when other references are lacking, i.e. we are not able to retrieve the $representativeness$ of a certain event.[13]

Going deeper, we notice that besides not being able to calculate the posterior probability by intuition, we are also very bad at putting the order in pretty easy concepts around the probability that are far behind the posterior probability, such as the joint probability. There are experiments that have shown that in some cases the probability $p(A, B)$ is considered greater than $p(A)$ or $p(B)$. This is impossible: an intersection of two sets can't be greater than one of the two sets, but at most can be equal. If there is a lack of intuition to deal with the basic rules of joint probability, it is hard to expect the correct formulation of the Bayes theorem, especially when it comes to conditional probabilities.[14]

Sometimes, however, some human errors involving the Bayes theorem or probabilities, in general, may be caused by external motivations that should not be underestimated: since the theorem is a framework for making decisions and reallocating probabilities between different possibilities the context can be critical in many cases. In fact, the process of belief updating, which is the process thanks to which we reallocate beliefs with the Bayes theorem, is slowed down by environments that are not trusted. Intuitively this can be relatable with everyone's experience: in a sense, priors probabilities (or pre-existing beliefs in general) that our brain applies, acquire more importance than likelihoods when we do not trust the process that generates the results.[15] This process is referred to as conservationism in belief revision. This may be more visible in laboratory studies where participants that have to prove their decision-making framework may be unconsciously conditioned by the

[13]Tversky and Kahneman, "Judgment under Uncertainty: Heuristics and Biases", pp. 1124–1130.

[14]Tversky and Kahneman, "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment", p. 313.

[15]Corner, Harris, and Hahn, "Conservatism in Belief Revision and Participant Skepticism", pp. 1625–1629.

context in which they are making decisions. In fact, they are usually wary about the exogenous factors they have to deal with in a laboratory and the belief-updating process requires more time and more ”evidence”.

One last point that seems relevant to report at the end of this chapter is that the human brain in many contexts does use the Bayes theorem to make decisions, but the theorem is applied unconsciously. Many cognitive processes seem to be explainable through rational and stochastic processes. Especially in learning tasks such as acquiring vision and speech stimulus and information probabilistic frameworks can explain, at least partially, how brains work and why brains are built the way they are.[16] Important developments in this direction have been made in the fields of information retrieval, computational linguistics, natural language processing, and machine learning in general. Once again, it seems that the Bayes theorem can help us understand one of the greatest and most fascinating mysteries of science, as well as one of the most complex observable systems in the universe: the human brain.

2.9 Summary

In this chapter, we have been able to touch briefly on some fundamental aspects of statistics thanks to which it is possible to treat Bayesian inference with more confidence. We started by explaining the concepts of sample space, event, outcome, and random variable to understand why it is often useful to think about probability in a spatial way to understand concepts that otherwise could be very abstract. Subsequently, we have dealt with the issues of probability mass function and probability density, explaining that in the first case we are in a context of discrete variables (or discretized), while in the second we are operating with continuous variables. Then, we tried to treat simply and practically the topics of the joint, marginal, and conditional probability that are the pieces that help us to derive the formulation of the Bayes theorem, but also help us to extrapolate many interesting properties. We have also dealt with prior, likelihood, and posterior probability explaining what are the roles of the various pieces of the Bayes theorem to make sure that we arrive at sensible

[16] Chater, Tenenbaum, and A., “Probabilistic models of cognition: Conceptual foundations”, p. 289.

and reasonable conclusions. The elements discussed in this chapter are the backbone of all Bayesian inference and it would have been impossible to move forward in the discussion, without having them clear. To corroborate the theoretical part we have also brought a classic example that deals with a single parameter: the coin toss. Then we wanted to make a summary of the Bayesian process by putting all the pieces together and linking them organically and also explaining what is the philosophy behind a Bayesian approach, and why it can be integrated, depending on the context, with the more traditional one. Finally, we wanted to conclude with a section on the cognitive biases that can affect the proper use of the Bayes theorem in the decisions we make every day, although it seems that there are unconscious mental learning processes, which would be explainable only through a probabilistic approach.

Chapter 3

A Bayesian Framework for Online Advertising Budget Allocation

The problem of budget allocation for online advertising campaigns has been faced by many with various and rather sophisticated solutions, especially in the last few years[1]. Machine learning algorithms and Artificial Intelligence support of any kind have become more and more central in many economic sectors and the advertising market is no exception. In our case, we will use a Bayesian approach to infer the structure of our model. This approach is useful in contexts of high uncertainty, thanks to its probabilistic nature and exploratory willingness. Therefore, the theoretical structure of the framework will be reported, with a discussion of its elements. We will talk about the premises that will clarify the work field, the objectives, the necessary assumptions, and the workflow of the model. Finally, we will discuss how to interpret the results of the model, trying to identify the reasons why this could behave in the way we do not expect, and also its functional limits. In addition, viable suggestions will be made throughout the discussion to improve the structure of the framework or to address specific problems that may occur in contexts where certain features may take on more importance.

[1]The problem has been discussed by Han and Arndt, "Budget Allocation as a Multi-Agent System of Contextual & Continuous Bandits", by Wang, Li, and Jia, "Optimal Advertising Budget Allocation across Markets with Different Goals and Various Constraints", by Nuara et al., "Online joint bid/daily budget optimization of Internet advertising campaigns", and by Kong et al., "A Combinational Optimization Approach for Advertising Budget Allocation"

3.1 Premises and Scope

It is necessary to start with the premises and establish our field of action. We are in a context where we can work with one or more advertising campaigns operating at the same time. Our specific case will involve just one campaign, but this is not particularly relevant. The important thing is that all campaigns must have the same objective, meaning that they all have the same conversion goal. Moreover, the monetary value of each conversion must be known. This is a fundamental point that we will explain in more detail shortly. The data we are talking about is the daily data about the expenditure and the conversions. If the value of a conversion is known, the daily revenue is also known, given the number of conversions. We purposely want to limit ourselves to the daily acquisition of data relating solely to the costs and conversions to work with the lightest context possible. We could easily operate in a context in which many other features are taken into consideration, there is no downside in doing so, but the considerations would be many more. For that reason, here we will limit ourselves to the simplest possible context because, in any case, the spending and conversion features are those that have the greatest impact on the model, and deserve to be analyzed separately. Operating with just these two features we keep the model as general as possible and somehow context-agnostic.

Another fundamental core point is that we are in a context in which each day (or any defined period) corresponds to a single round, and at each round, we are going to allocate a budget that will generate a return. The framework operates circularly: the data is acquired, the budget is allocated, new data is generated, the new data is acquired. This means that at the beginning of each subsequent round, one more data will be available: the previous round data. If in round n we will have all the data available up to round $n-1$, at round $n+1$ we will have the data available up to round n. Our first goal is to build a payout curve that can predict the progress and the relationship between costs and conversions for the campaign. The independent variable is the cost, while conversions are the dependent variable. The extraction of the curve is necessary to identify the point at which the value of the conversions equals the expenditure made to obtain those conversions. In summary, since the value of a conversion is known, we want to try to maximize the number of possible

conversions (and revenues). So the framework's general objective is to:

$$Maximize \sum_i f_i(x_i) \tag{3.1}$$

Each day, therefore, a new budget will be identified, and it will be allocated for the next round. Indeed, it is assumed a budget allocated greater than 0. The optimal point along the curve identified for each round will correspond, therefore, to a coordinate $[x,\ y]$ where y corresponds to the number of conversions generated by x which corresponds to the hypothetical expense. The identified expense will be converted into a budget and will be allocated.

You may have noticed that when we talk about the available data we need information about costs, but the framework aims to allocate a new daily budget for the next round. Costs and budget are not the same, even if sometimes they are mixed up. We are going to explain the difference in detail in Chapter 4, but keep in mind that the two concepts do not always overlap, and in general, is a good habit to keep them apart. The cost is what we spend for the interactions people had with our campaign, while the budget is the expenditure limit we don't have to overcome. The cost is something real, while the budget should be a theoretical evaluation. For that reason, the relationship between them has to be studied.

Now let's focus on the fact that the value of one conversion must be known. This element is central in the construction of the model. Since the value of a single conversion must be known, the revenues generated by a sum of conversions are also known. Knowing the revenue, our goal is to maintain a certain ratio between costs and revenues, and we can conclude that there is no maximum a priori that we must not exceed. Let's explain why. We have not yet discussed the structure we hypothesize for the curve to be extracted, but we certainly know that, generally, the more budget you invest the more you convert: in stable contexts, on average, if I invest x and get y, it is reasonable to believe that if I spend $2x$ I will get a return somehow greater than y[2]. Given that, we can assume that the curve is a $monotonic\ increasing\ function$ (if x grows, $f(x)$ grows too) and we can theorize

[2]In practice, this statement it's hardly verifiable since the context is always changing, and it is perfectly reasonable to see that for greater budgets we can observe minor conversions. However, the correlation between budget and conversions should still be positive, albeit with very large variances.

a curve structure like $y = \lambda x^{\alpha}$ with $\lambda > 0$, $x > 0$ (we want to allocate a budget always greater than 0) and $0 < \alpha < 1$, because this constraint let the curve respect the monotonic growth and it does not end up in an exponential shape, nor in a Pareto-like distribution [3]. However, in a borderline case, historical data might show us that for each unit of cost, the returns are greater even for big investments. Under such conditions, the extracted curve function would be something like $y = \lambda x^{\alpha}$, with a very large α. As we will see later in this chapter, small variations in the exponent, expecially for exponents close to 1, generate an enormous growth in the allocable budget: sometimes values out of scale for any realistic advertising campaign. Such a scenario seems absurd, but it may not be so far from reality and it is analytically possible, as we will see in the part of this chapter dedicated to the optimal point. Beyond the extreme scenarios, we intend to allocate extra budget until each extra unit of cost generates a return greater than or equal to the expense incurred.

In practice, we need a constraint under which we want to maximize the revenues. Well, a first reasonable formulation can be: at the next round I want to get a conversion more than the previous round, only if the cost of the new conversion is less than its monetary value. This is the first approach we want to propose, but we are going to discuss the pros and cons in a future section. Formally, we can express this condition in the context of multiple campaigns with:

$$\left(\frac{df_i(x_i)}{dx_i}\right)^{-1} \leq C,\ with\ x_i > 0 \tag{3.2}$$

Rearranging the equation:

$$\left(\frac{df_i(x_i)}{dx_i}\right)^{-1} \leq C = \frac{dx_i}{df_i(x_i)} \leq C = dx_i \leq C df_i(x_i) \tag{3.3}$$

So, finally we get the equation

$$C df_i(x_i) \geq dx_i,\ with\ x_i > 0 \tag{3.4}$$

This equation may appear hard to grasp, but let it be clearer. dx is the Cost Per Incremental Acquisition (CPIA) which is the cost for each new conversion I get. The CPIA is not

[3]The structure of the curve will be discussed in more detail in the next section, for now, we take these assumptions for granted.

constant over the curve, and it generally grows progressively. If I get 10 conversions in the $n-1$ shift, I intend to get 11 conversions in the n shift and I want to get 12 in the $n+1$ shift, the cost I have to bear to get 11 conversions instead of 10 will be different (probably smaller) from the cost of 11 conversions instead of 12. We will explain shortly why this is true, but as anticipated, it is unlikely to deal with a constantly linear relationship between costs and conversions, so payout curves reduce their slope as you push higher budgets: increasing the budget increases the CPIA. C is the value of a single conversion that we recall to be fundamental here. $df(x)$ is the number of corresponding conversions obtained by the CPIA. The equation is expressing that the cost of obtaining each new conversion must be lower than the value of the same conversion. In summary: we can spend until we lose money getting new conversions. Let's make a small example to make the constraint clearer. If the value of a conversion $C = 100$, and, in the next round, I want to get 1 more conversion than the previous round (so $df_i(x_i) = 1$), I have to make sure that the Cost Per Incremental Acquisition dx_i is less than 100, otherwise it is no longer convenient to try to get a new conversion and I have to stop at the previous conversion.

3.2 The Framework in a Nutshell

After talking about the premises and purposes, it is necessary to explain upfront and very roughly how the workflow of the framework works. Recall that the framework works iteratively in a set of rounds. A round is a custom-defined time frame: it can be a day, a week, 18 hours, ...

For $i = 0$ in $N - 1$, where N is the number of rounds:

1. **Data collection** First of all the data: we need daily historical data about spending and conversions. If we don't have enough historical data we can interpolate them artificially with any regressor or oversampling method we like. The data is composted, for each day by costs $\{x_1, x_2, ..., x_n\}$ and conversions obtained $\{y_1, y_2, ..., y_n\}$

2. **Bayesian inference** Once the data is acquired we generally should remove outliers with an appropriate algorithm. Moreover, it is necessary to study the relationship between cost and conversions: we have to fit a regression to the data. To do this, we

prefer a Bayesian probabilistic approach and we would compute through Bayesian inference the joint posterior probability (JPP) of the parameters of the model. The joint posterior probability is the joint probability of the posterior probabilities of the model parameters after Bayesian learning. To compute the JPP we have to choose the prior probabilities of the model parameters ($p(\theta)$), which in our case are the prior of the exponent α and the prior of the constant λ. Once we have the priors of the model, let's define also the likelihood $p(D|\theta)$ that generally follows the model and has an error normally distributed. In practice, $likelihood \sim N(f(x), \sigma^2)$. Consequently, we define also the prior over σ. Finally, Let's perform the inference through a sampling algorithm of the MCMC family and get a joint posterior probability. Performing the MCMC algorithm we get a sampled distribution from the JPP of the parameters

3. **Thompson-Sampling** Since the Bayesian inference does not give us one possible curve to adopt, but it gives us a posterior probability density function of the curve, we need a heuristic to choose the "best" curve. The Thompson-Sampling algorithm will do the job: an algorithm that will choose the curves proportionally to their probability of being the best (proportionally to their probability density). So we randomly take a point in the joint posterior probability obtained by the inference in the previous step. This point in the joint space of parameters corresponds to a vector of parameters of the model. Thanks to Thompson-Sampling we are going to explore different parameters at each turn, in proportion to their probability of being the best: more likely we are going to extract from the areas that are probabilistically denser and less likely from the areas that are probabilistically less dense. At this point, we have extracted the parameter vector of the model θ that we need to build the curve, for the current turn

4. **Optimal Point** Once obtained the curve we have to find the optimal point of the curve: we will find how much we should spend at most, such that a new conversion will be more expensive than the value of the conversion itself. With the marginalistic approach we have to compute the first derivative of the function in dx, and equate it to 1, since we whant to find the point along the curve where it assumes a linear shape

with slope 1. When the $f'(x) = 1$ we spot the exact cost where $df(x)C = dx$

5. **Cost-Budget relationship** Once we have identified the optimal cost, we must derive the relationship that exists between cost and budget and from there derive the optimal budget. We can do it again with a Bayesian approach: if we assume a linear relationship[4] $y = \alpha x + \beta$ with a normally distributed error ϵ, we have to put the priors over α, β, σ, define the $likelihood \sim N(\alpha x + \beta, \sigma^2)$, perform the sampling process and extract a curve that can derive the optimal budget from the optimal cost

6. **Budget Allocation** At this point, once we extracted the relationship between cost and budget we can compute the budget and allocate it to the ad campaign and wait for it to do its job in the advertising environment.

7. **Record the data** After the end of the round, new data $D = \{x_n, y_n\}$ will be produced that will have cost-conversion coordinates. The new data is added to the historical data.

3.3 The Payout Curve

It's time to talk about the function that links cost and conversions. To build the probabilistic model that we need to predict the performance of the campaign, we must first of all study the function underlying the model, its backbone. However, some assumptions are necessary, and those assumptions must take into consideration the elements we have already dealt with. Given the limits to which the curve is subject, it is necessary to compute the CPIA. According to that, the curve must be continuous and differentiable. The second characteristic: as we have already anticipated, the function will have monotonous growth

[4] In platforms where Smart Bidding solutions are implemented, generally the daily budget set is an average quantity for the daily budget in the long run. In a single day, the cost may be greater than the allocated daily budget. Since we change the budget at each round, we can see directly the regression toward the mean of the budget, so it can be reasonable to think of the relationship budget-cost as linear. Normally, we should just assume that $budget = cost + quantity$ and this is the linear regression with slope 1 in $y = \alpha x + \beta$, so our assumption of linearity can include the hypothesis that the budget is simply the cost, plus something.

(basically the more I invest the more I get), but it won't grow always with the same acceleration. The slope will progressively decrease. We must therefore imagine that there are so called points of diminishing returns: the curve, as you increase the budget, decreases its ability to generate revenues.

The diminishing return theory is not new at all in economics[5] but, why do we have to assume that the curve has diminishing returns points even in advertising? Let's imagine a certain audience made up of a pool of people with quite different characteristics. This pool is the target that I have identified for my advertising campaign. Among these people, some are more interested in converting, and some are less, but there are also those who in reality will never convert, even if they are part of the target cluster, which always contains also misclassified elements. Let us assume that with an expense of x the advertisement will be able to reach the entire target once. Let us also impose the condition that after a person has converted he can convert again, but less willingly each new conversion. If I spend x in the first round, all the most interested people that need to see the ad once to convert will convert. In the second round, to get the same result as the day before, the advertisement will have to convince people who are more reluctant to convert than the ones who converted in the first round, or it has to convince people who have already converted to convert again. In both cases, more effort will be required, which translates into greater cost. At every round, therefore, will become more and more expensive to obtain the same results as the previous round and, consequently, if I want to obtain even greater results, the expenditure will not increase linearly, but exponentially. That is a practical example that shows how a payout curve progressively reduces its ability to generate conversions as the budget grows.

At this point, how can we formally define a function with these characteristics? Something similar to a logarithmic function would be needed, but since negative conversions cannot be generated, we need a function defined for $y > 0$. It seems that the function that would do the trick is a power-law, with an exponent between 0 and 1. Fixing the exponent between 0 and 1, we prevent that the function "degenerates" to exponential function (exponent greater than 1), but also that it takes the form of a long-tailed power-law (exponent smaller than

[5] Britannica.com, *diminishing returns*.

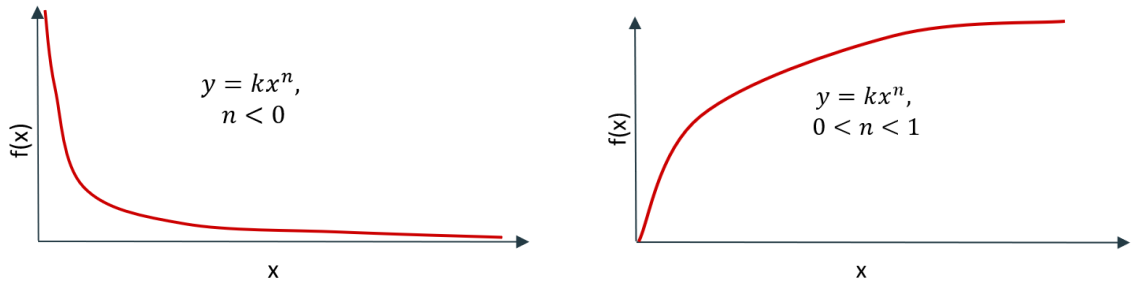

Figure 3.1: The shapes of two power-law with different exponent

0), a function very popular in many natural and social sciences[6](see figure 3.1). By fixing the exponent between 0 and 1 we also respect the premises of having sublinearity, a monotonical growth and a differentiable curve.

Extracting a power-law from data can be approached in several ways. The theorization of this function is not recent at all, but it has rediscovered its wide usability since it began to be used to explain social and natural phenomena that require a large amount of data (World Wide Web, Social Media connections, Proteins interactions, and so on). For this reason, in the last thirty years or so, scientists and academics have begun to talk about this function as they did never before, due to the fact that it seems to fit very well different kinds of phenomena[7]. However, some believe that this weight given to the power-law seems to be disproportionate and that many phenomena, that were believed to be scale-free, can be explained more precisely through other functions[8]. The debate is still open and it involves different schools of thought. For this reason, there was also abundant discussion about which method is the most appropriate for deriving the power-law from the data. We will focus on Bayesian inference, which we have already discussed in Chapter 1, but it may be useful to keep in mind that there are two other methods: Maximum Likelihood Estimation (often abbreviated as MLE, traditional frequentist approach) and linear graphical fitting on log-log scale data and the consequent transformation of the linear regression into power-law[9].

[6]Newman, "Power laws, Pareto distributions and Zipf's law".

[7]Barabási and R., "Emergence of Scaling in Random Networks".

[8]Broido and Clauset, "Scale-free networks are rare".

[9]Goldstein, Morris, and Yen, "Problems with fitting to the power-law distribution".

We have already anticipated that all the parts related to parameter estimation will be delegated to PyMC3, a library dedicated to probabilistic programming which is mainly based on algorithms belonging to the family of Markovian methods. So we will not go into the details of how these specific parameters are estimated, but it is necessary to describe them briefly. We have seen that the function will have a structure like $y = \lambda x^{\alpha}$ where λ is the constant, and α is the true center of gravity of the function, describing the slope of the curve. We said that n must be between 0 and 1 to maintain the structure needed to fit the data. k, on the other hand, is a parameter that is more contextualized to the available data and has no particular boundaries to adhere to. Furthermore, we can treat the parameter λ and α as dependent since the constant and the exponent are inversely proportional starting from the same data. This means that each extracted sample will have its parameters that will respect this ratio: each extracted sample will be the result of the joint posterior probability of the parameters, so this dependence will be respected. On average, therefore, if a sample will have an exponent with a relatively high value, it will also have a relatively low constant and vice versa.

We mentioned that the curve can be extracted following a frequentist approach or following a Bayesian approach. We have said that we will follow a Bayesian approach because this is more advantageous, especially when we have little data, but why? The output of a Bayesian regression does not correspond to one possible curve, but to the joint posterior probability of the parameters that make up the model. In this way, there will be more dense and less dense areas in the space of the posterior probability and each point of the joint posterior probability corresponds to a possible curve. In the Bayesian approach is that there will be more likely curves (the ones derived from the points where the joint posterior probability is denser) and less likely curves (the ones derived from the points where the joint posterior probability is less dense), but all simultaneously proportionally possible according to their density. The key point is that the posterior probability is the result of two components: the likelihood $p(D|\theta)$ and the prior probability $p(\theta)$. We want to take into account both of them. On the other hand, in the frequentist approach, this does not happen. The only factor that it can take into consideration is the likelihood $p(D|\theta)$ and maximize it. Generally, we assume that the error is normally distributed so the $likelihood \sim N(f(x), \sigma^2)$. Since it is

maximizing a probability, it will produce one possible output as result, and in the case when the data are sparse, poor, and little, this approach can generate big problems. This is not the purpose of this essay, but it can be demonstrated rather easily that, for example, in linear regression maximizing the likelihood (Maximum Likelihood Estimation) is equivalent to minimizing the cost function based on the Summed Squared estimate of Error.

At this point, we have added a few more elements to the framework. In addition, we recall that the framework operates iteratively and the extraction of the curve will take place at each round, having one more piece of data available. Doing the curve is refined progressively having more and more data available, round after round. However, since Bayesian inference will not return a single possible value that the parameters can acquire, but posterior distribution of possible values, we must briefly introduce another simple segment of the framework: the Thompson-Sampling algorithm[10].

It must be remembered that for the calculation of the posterior probability, to avoid the intractable integral to the denominator of Bayes theorem, we use a family of algorithms that aim to approximate the posterior probability distribution. The approximation is obtained generating a large number of samples of the joint posterior probability and hoping that indeed everything has gone right (no sub-optimal points, no divergences). These samples are a set of values that are more or less likely to aspire to be the $true$ extracted set of parameters, with respect to their density. If we are lucky, the extracted pdfs are narrow and tall (meaning that there is a lot of density of extracted samples in a small range of values), if we are less lucky the extracted distributions will be wide and short, indicating that no specific range of values is significantly denser than the others. Obviously, tall is the equivalent of $probabilistically\ denser$, whereas broad imply a higher variance in the the samples distribution, and vice versa. The power of this family of algorithms is that, if there are no convergence problems, in the long run more samples will be drawn from the more likely areas, while in the less likely areas there will be a lower concentration of samples. So, if we assume that sampling has led to consistent results, we must also assume that if a sample is randomly chosen from the set of joint posterior probability samples,

[10]Strens, "A Bayesian framework for reinforcement learning".

the probability that a this is taken from an interval is directly proportional to the posterior probability of that interval. This is the actual logic of the Thompson-Sampling algorithm that fits perfectly with posterior probability sampling algorithms. We highlight once again that a sample extracted from the set of posterior probabilities represents a point of the joint posterior probability which therefore maintains the information relating to the relationships existing between the various parameters. If the parameters are effectively independent everything goes in the right direction in any case, but most of the time these affect each other and therefore it is important not to take random values by drawing lots from the marginal posterior probabilities of the individual parameters, but to resort to a sample of the joint posterior probability. In such environment Thompson-Sampling seems to be an excellent trade-off between exploration and exploitation, since it is capable of exploring the posterior probability space of the parameters proportionally to the probability of being the actual parameter.

From the description that has been given, it should be clear that the algorithm is self-adaptive: new observations progressively modify the probability of picking a value belonging to an interval, with respect to the density of the interval. If the new observations fit the curve, this will imply that at the successive round the density of the interval chosen will be greater, and therefore it will be more probable that the choice of the previous round will be confirmed. On the contrary, if new observations will be far from the expectation, the density of the posterior probability will be lower and therefore the values that in the round $n-1$ had a lower probability of being chosen will have a higher probability in the round n.

3.4 The Optimal Point And Marginalization Problems

The constraints we have imposed on the framework require us to think in terms of derivatives and marginality. If we want to make sure that we spend for getting a new conversion as much as the economic value of that conversion, we need to find the point where each subsequent cost unit would generate a lower value in terms of conversions. This concept is deeply related to the previous assumption about the presence of diminishing returns points.

Although historically economists have struggled to mathematically formalize the problem of profit marginalization in a timely and precise manner, now this type of approach seems to be the most correct in our case as well[11]. We have defined that our goal should be to maximize conversions, however, the focus in this regard must go to the cost of each new conversion, instead of the average cost of a conversion. We've defined this as Cost Per Incremental Acquisition because every conversion doesn't cost the same: the first ones cost less, and the more you invest budget the more these will need more work to be obtained. If we didn't have an analytical solution (the calculation of the derivative), our method would be to start a sort of loop in which, at each round, starting from zero, we get one conversion more than the previous round (starting from zero) and compute the cost of the new conversion obtained ($[Cost\_Conversions\_Round_{i+1}] - [Cost\_Conversions\_Round_i]$). Since the value of each conversion is known, we should stop when the calculated difference is equal to the conversion value (or when we see that the calculated difference is greater than the conversion value, but on the next round we should stop at the previous conversion instead). Given the diminishing returns law, it is assumed that the next conversion, on average, will cost more than the previous one, and so inevitably the conversion following a conversion that cost as much as the value of conversion will cost more than the value of the conversion. Consequently, the marginality will diminish following a power-law, not linearly, therefore if at round $n-1$ the marginality was x and at round n was $\frac{x}{2}$, at round $n+1$ it will surely be lower than $\frac{x}{4}$. In short, finding the optimal point empirically would be technically possible, but learning would be extremely slow and the process exhausting. Here we wanted to report this example because it is probably easier to understand the problem of diminishing returns and decreasing marginality. However, there is an analytical solution that is extremely straightforward.

The analytical solution is provided by the computation of the derivative of the function. To find the optimal point it is necessary to identify the interval of the curve that assumes a slope of 45 ° (that corresponds to an angular coefficient equal to 1 in a linear context): that infinitesimal interval will be the space in which the curve describes a ratio of 1 between budget and conversions. This is the point where the curve assumes a bisecting slope. So,

[11] Brue, "The Law of Diminishing Returns".

the point is determined by the condition $dx = Cdf(x)$. Once the point is spotted, if we move to the right, investing larger budgets, the ratio would be greater than 1, resulting in lower returns than the budget invested. Moving to the left we would find a ratio of less than one as a certain budget expenditure will result in a greater amount of returns. In practice, to do so we must impose the derivative of the extracted function equal to 1, $f'(x) = 1$, since the derivative study how much a function changes with respect to the argument of the function, and we want to find the point where the change is the same for $f(x)$ and x. We equate the first derivative to 1 since this is the slope that the bisecting line has. By placing the first derivative of the extracted function equal to 1, we will identify a point in the x axes. That x coordinate along the original curve $f(x)$ will spot the optimal point.

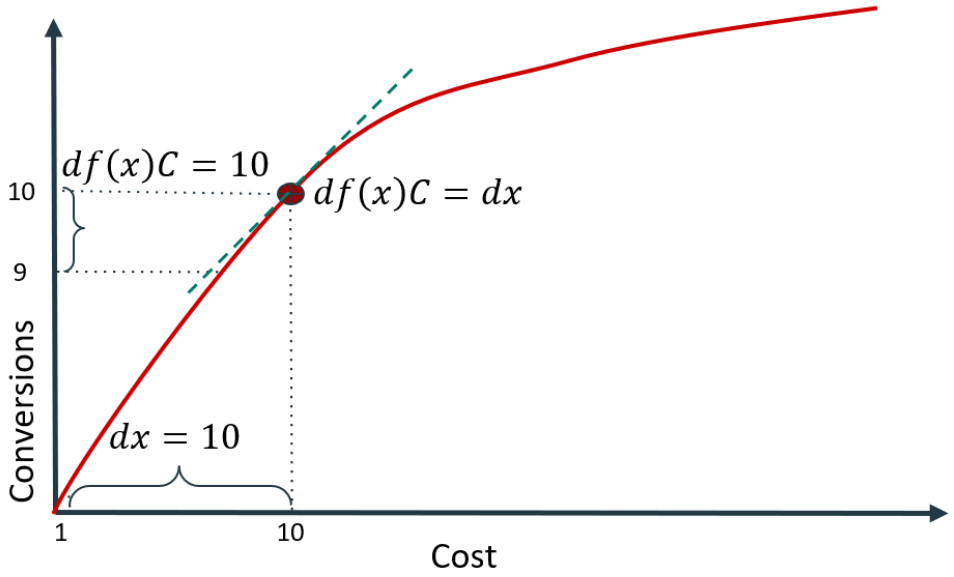

Figure 3.2: Example of optimal point with C=10 (value of a conversion)

Computing the point where $f'(x) = 1$ we are informed about which position in the curve every incremental cost unit will generate a marginal return smaller than the marginal cost added. However, the problem could also be approached in another way. We could face the problem leaving apart marginality, and taking into consideration the cost-conversion ratio on the whole curve. This approach is normally the simplest and most direct one to explain: we are no longer interested in always spending, for each conversion, at most the cost of its value, but we are interested that the average cost of all conversions does not exceed the

value of conversions (which is fixed, while the cost is variable and increasing). In this way, the goal is still to maximize conversions, but the constraint becomes $x \leq Cf(x)$. Under these conditions, the scenario doesn't change much: instead of finding the point beyond which each extra cost unit will yield less than the value of the cost, we want to find the point at which the entirety of the budget must equal the entirety of the returns generated by conversions. In this way, there may be conversions that individually cost more than their single value, but this extra cost is balanced by the other conversions whose return was greater than the cost. The balance between costs and returns will be neutral, but the maximum number of possible conversions will have been generated without losing money on the totality of the acquisitions realized. Following this second approach, it is clear that the possible budget to invest is greater than that of the first approach.

Now, let's check graphically what we have argued so far. Visualizing graphs can help us understand the advantages and disadvantages of the two approaches. Let's start considering the marginal approach for the variations of the constant in our function that is formulated as $y = \lambda x^{\alpha}$.

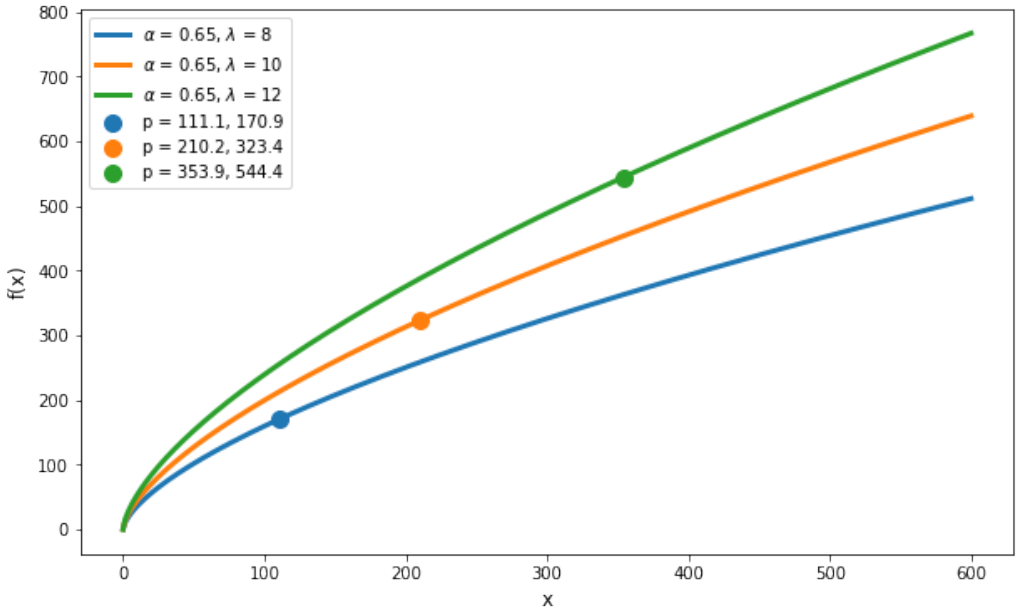

Figure 3.3: How the optimal point changes with the marginal approach over the constant λ

The constant and the exponent of the function are values that condition each other, but for

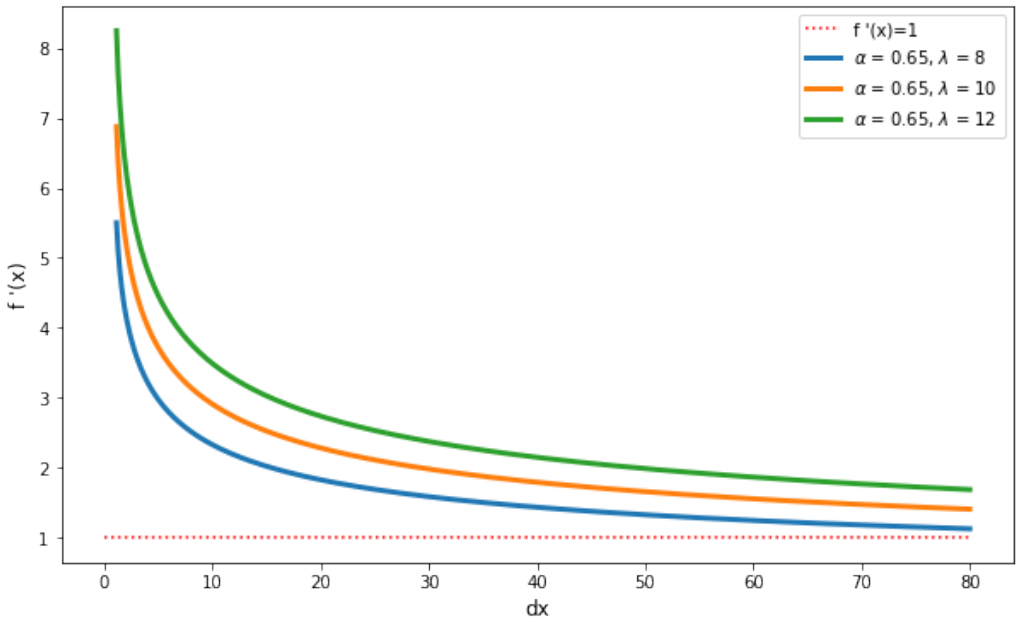

Figure 3.4: Plot of $f'(x)$. When $f'(x) = 1$, we spot the optimal point that is reported in $f(x)$ in Figure 3.3

totally demonstrative purposes we are going to consider the two elements in a disjoint way to show how some errors can generate more serious distortions than others. In this scenario, three curves were generated with the same exponent $\alpha = 0.65$ (a plausible value for real data that do not have particularly good payout curves), and the constant λ was varied at a rate of 2. We can observe that variations over the constant generate curves that are certainly different but maintain a similar order of magnitude. The constant tells us how high the first inflection point (first point of diminishing return) must be to the y-axis. It's like giving it the initial scale on which the data are placed. Once past that point, the curves radiate, growing steadily. This makes perfect sense since the inclination is given by the exponent, but it also tells us that an error in the choice of the constant grows more or less linearly with respect to the constant. We can afford, then, to make small errors in estimating the constant without making traumatic errors in calculating the budget. We can probably choose less informative and more exploratory priors, even when we have little data, without heavily affecting the result.

We said that this approach allows us to maximize $f(x)$ without ever spending more for one conversion than its value. In this way, it is clear that we also maximize the savings (or profits, depending on your point of view) to achieve the same number of conversions.

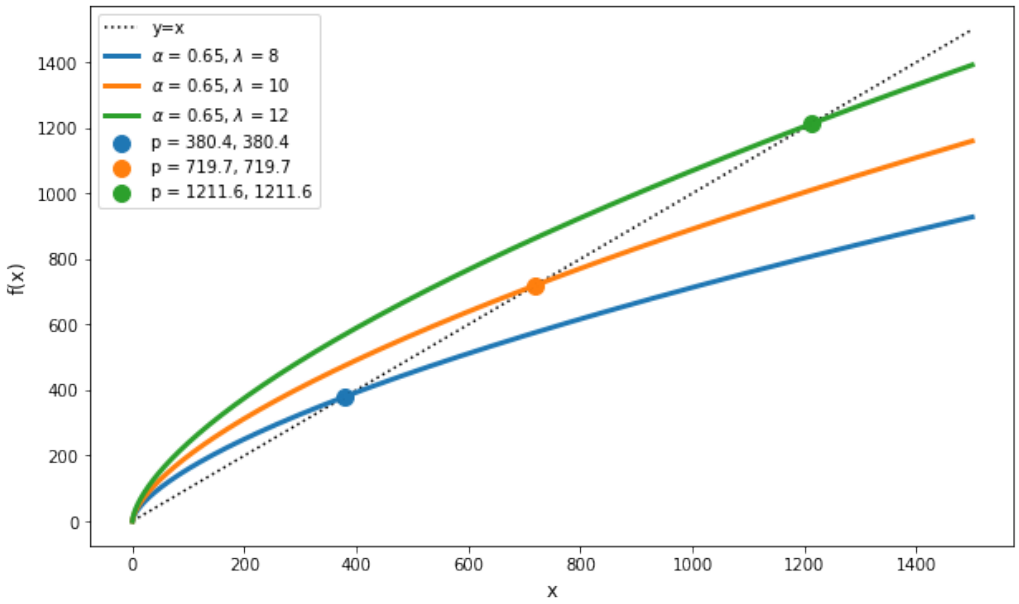

Figure 3.5: How the optimal point changes with the average approach over the constant λ

This is certainly a conservative approach, in the sense that one would prefer, on average, to have one less conversion if it were to cost me even a penny more than its value. In short: our desire to convert is less than our desire to spend an affordable amount per conversion. However, there are times when the desire to increase the number of conversions is greater even than the desire to not spend too much per conversion. Maybe we are conducting an aggressive awareness campaign, or we want to empty our warehouse as soon as possible. All in all, we can only worry about making as many conversions as possible without ultimately ever spending more than the total amount we earn. If we reasoned then no longer on the marginality of the single conversion but on the average of the total cost of the conversions that we obtain we can invest much more budget and obtain many more conversions. In our example we can compare Figure 3.5 and Figure 3.3: in the average approach the budget invested (x axes) is roughly triple, and the revenues obtained (y axes) are roughly doubling the marginal approach.

Now we can see how our curve changes its structure as α changes. We have said that the exponent is the supporting element of the curve, but it has not yet been made explicit when an error in these terms can affect the choice of the budget. Here is a graph in which the constant $\lambda = 6.5$ has been fixed and α has been made to vary with a rate of just 0.1.

Let's start also this time with the marginal approach. It is immediately evident how a variation of the exponent can affect the positioning of the budget in a much more substantial way than the constant does. One can observe how the error grows exponentially as the exponent increases. If an error between $\alpha = 0.4$ and $\alpha = 0.5$ generates a budget variation of little more than double (from 4.9 to 10.9), a variation of exponent between 0.7 and 0.8 generates a budget variation of more than 24 times (from 156 to 3802). Such a deviation can be the cause of the immediate death of a campaign and also of much dissatisfaction on the part of those who have to allocate a budget that can vary 24 times based only on a tenth of a unit of the exponent. When the exponents are high (a likely hypothesis), even a variation of a hundredth of a unit can generate substantial variations and this confirms that much more attention must be paid in the choice of priors, but also the selection of data. Especially when we have little data available even a few new points can make the curve change substantially. In these cases, it is considered important to subject the data to outlier detection processes to limit the inconvenience caused by random fluctuations.

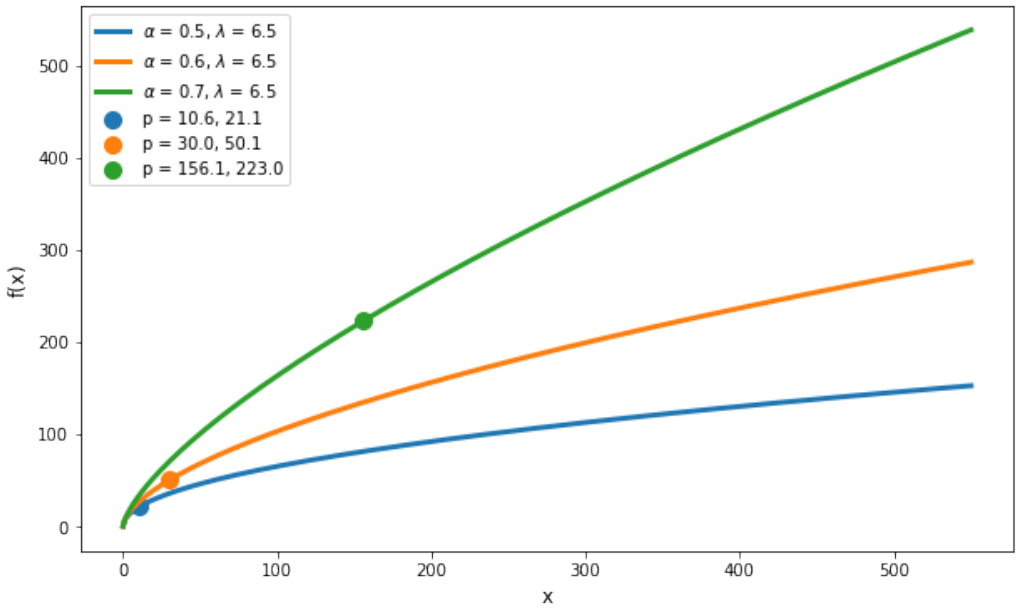

Figure 3.6: Positioning of the optimal point over the variation of α with marginal approach

From the perspective of the average conversion value approach, we find that the effects are similar. I don't think it's possible to determine that variations in one parameter will cause one approach to be preferred over another. As a general rule, it is sufficient to say

that we must be very cautious in the choice of the exponent, even on infinitesimal margins, especially when we go on values close to 1. In this regard, we could limit the exploration to a range rather narrow than the average to avoid going to explore the tails of the posterior probability that may present more extreme values. This kind of solution can be effective in limiting gross errors, but clearly, it also limits the explorative potential of Thompson-Sampling and makes the framework less sensitive to changes (both causal fluctuations or real changes in the data generation process).

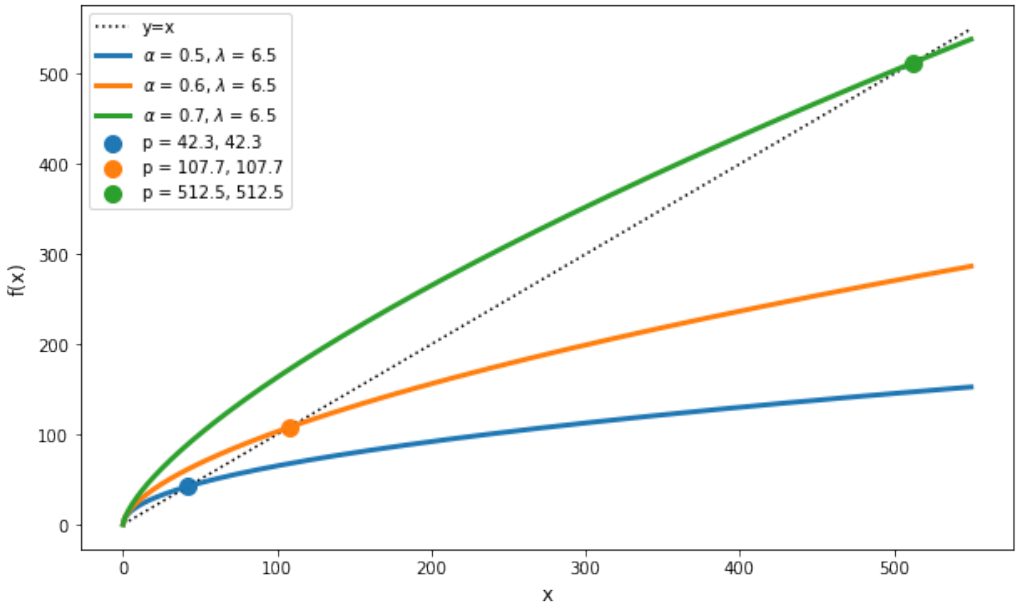

Figure 3.7: Positioning of the optimal point over the variation of α with conversion average value approach

Ultimately, it is possible to observe how the first approach is a much more sophisticated compromise: the marginalistic approach is not easy to understand and consequently to apply in real-life contexts, especially when one has to convince a media planner to spend less money, or a company to sell less (when it is not convenient). However, it is observable how in real contexts, some curves initially extracted can present very uncertain predictions with large variations of exponent or constant. These wide variations follow very fluctuating initial budgets and especially for the marginalistic approach the risk is to invest very small amounts of money that could lead to so few conversions and so little budget invested. Such an eventuality in practice becomes difficult to justify in the long run (although sometimes it would be enough to rethink the advertising campaign at the base), so the approach of the

average value of the conversion becomes the most justifiable on the marketing and business side.

3.5 Temporal Information

The standard implementation of the framework is not designed for handling temporal information. Data is introduced and used regardless of the time it was produced. In this way, a rather regular trend of the data is assumed, as well as a constant generative process. But this is not true in practice. The information about when the data was acquired, but also when the budget will be used is undoubtedly important for advertising campaign planning. It is well known that in an advertising market, regulated by bidding strategies, at least two factors can strongly influence the performance of campaigns: the predisposition of the market to receive advertisements and the size of the competition. The first element is quite evident: if the campaign has been prepared in a way that is not functional for the market, its failure is assured. There can be errors in the creative part, in the copy, in the target, and also in the timing. Many contextual elements can determine the failure of an advertising campaign, but certainly, the timing has a central role. If the timing is wrong, surely the campaign will not perform at its maximum. If I'm promoting a brand of coats, it is trivial to say that the payout curve of the advertisement will not behave the same in November, or in June.

Since many platforms have implemented smart bidding solutions, the aspect of the ads competition is less evident, but it is always present behind the curtain. If all of a sudden, without having made any changes to the advertising environment, the Cost Per Click inexplicably increases, it's because one or more competitors have entered the bidding process and they are spending money to acquire the same spaces that were previously being contended by fewer participants. If the competition is greater, the costs for having the same advertising space available will also increase. Generally, competition grows and intensifies when the market is more inclined to convert (the two phenomena are certainly correlated since advertisers will prefer to operate in a more receptive market, hoping to do better than their competitors), but certainly, this element is also conditioned by the temporal feature

and it can significantly affect the performance of a campaign. So it is clear that also the bidding strategy and the competition can affect the payout curve since costs and conversions will fluctuate a lot.

I don't think there is a one-size-fits-all solution to this type of problem, but we know that the Bayes theorem is based on prior probabilities and data. A solution, therefore, can require manipulating the data that we have available to make them as consistent as possible with what we expect. In this case, the previous knowledge of how the market might behave in certain periods of the year is fundamental and a campaign with a lot of historical data can help us a lot to have an idea of what could be the future trends. For example, we could use the historical data of periods of our interest in the previous years to increase their weight in the process of curve extraction. Or, in case historical data are not available, it is possible to interpolate new data with regressors or oversampling algorithms to increase the weight of some periods that we believe should be overestimated. Another more elegant technique could be to act on priors: the prior knowledge of the market can contribute to moving the posterior probability towards what we expect. Changing priors probabilities in the Bayesian inference process, and choosing more informative and denser priors, it is as if we give more weight to what we believe before we have seen the data. it is an interesting approach that allows us to work directly on the model, but it can also lead to disastrous consequences if our beliefs turn out to be false.

The manipulation of the data or priors enables the framework to be more "receptive" to changes in a certain direction: we apply an optimistic or pessimistic bias. It must be said that thanks to the exploratory Bayesian approach, the framework will always give a proportionate weight to the "anomalous" trends and if the anomalous trend becomes more and more present, the Thompson-Sampling algorithm would slowly learn to give it an adequate weight. However, if the historical data are very consistent, the learning process risks being too slow to grasp the particularity of the moment. Furthermore, often the periods of variation outside the norm are limited in time: we think of the weeks before Christmas or Saint Valentine. This means that once the period is over, one should return to a much more regular state. Again, this kind of framework is not prepared for changes and it risks

being too slow to get used to the exceptional nature of the period. For this reason, it is sometimes necessary to act artificially to ensure that the framework can be more reactive and functional.

3.6 Can the Framework Fail?

At this point, it is important to understand what the behavior of the framework will be with respect to the environment in which it is deployed. First of all we recall that the framework aims to optimize a process of budget allocation, so it is not possible to establish in advance how much a campaign can spend, but in practice this can be hard to cope with media planning expectations. We have already mentioned that in favorable conditions, i.e. when cost conversions ratio is fairly close to 1, the extracted curve could present an optimum budget point particularly far from the budgets normally invested. At this point, if the future data will confirm the positive trend forecast, then it would be possible to generate a much higher number of conversions than what has been done up to now (or at least generate the same number of conversions at a lower cost). If, on the contrary, new data will contradict historical data, the budget would be progressively adjusted to the new data generation process with an intermediate stage of greater uncertainty and exploration.

In an opposite condition, the identified budget may be too low compared to what a media planning could expect. This could happen, but we would not be surprised. The performance of the campaign could have worsened for increasing costs due to new competitors in the market, to incorrect changes in the context of the campaign that has led to a worse reception by the public, or to the saturation of the market for which each new conversion will become very expensive and not cheaper. It is worth recalling that, under a stable generative process of the data, the framework will converge to the optimal solution with the scope to get the highest number of conversions at the lowest cost possible. Problems arise when the generative process is not stable and, unfortunately, in general, this is the case. The generative process of data is not a phenomenon with clear boundaries, with a beginning or an end. It has fuzzy boundaries and it is sometimes difficult to determine what is a natural fluctuation in the data, or what is a clear paradigm shift. For this reason, a Bayesian ap-

proach supported by the Thompson-Sampling algorithm can guarantee, even in a context where the generative process is changing rapidly, not to repeat the same error over and over again.

In general, we know that the default framework is not designed to manage temporal information. This can inevitably create problems, for the reasons we have already described, but an extra example can help. Let's imagine that up to the day $d-1$ the data generation process is quite stable, and therefore the variance of the data is rather limited to daily fluctuations. At day d, an event deeply affects the process and the new data has a structure that greatly increases the variance of the data-set. At this point, it would be important to keep in mind that probably in this setting the most recent data should have a greater weight than the past data. However, even by doing nothing, if the new generative process is stable and durable enough, the framework will converge to an optimal solution. In the intermediate phase, however, when the exploration is at its maximum, money or possible conversions are being wasted (it is still learning!). If we could "inform" the framework about the order in which the data arrived, we would be able to have a faster and more painless adaptation. A possible solution could be to weigh the various instances based on their generation date. The regression would take into account the data as well as their timing. This could be a good point of improvement for the framework in the future. Another less elegant, but effective solution could be to fix a temporal sliding window of the period under analysis, to progressively abandon outdated data. The problem of this solution could be finding an adequately large sliding window to ensure consistency of the data, but at the same time guaranteeing the greater relevance of the most recently acquired data. Beyond the implementation choices, there are margins for improvement that also leave room for creative and non-trivial solutions.

3.7 Summary

In this chapter, we analyzed the theoretical part of the framework for budget allocation. Its strengths, the logic it follows, the various choices that can be made to address certain problems, and the situations in which it might work differently than we expect have

been considered. The chapter began by specifying the necessary premises and purpose, including the boundaries that it has to respect. We have seen the structure and the steps the framework follows, the structure of the model to extract, and the limits to which it must submit. Subsequently, we discussed in more detail the curve that will be extracted, its characteristics, how Bayesian learning can help us to bring out the structure, and how the Thompson-Sampling algorithm is necessary to ensure a fair proportion between exportation and exploitation of knowledge. Next, we determined the two viable approaches to determining the optimal point (the marginal approach and the total-average approach), pointing out how these two methods can behave with variations in the curve. Then, we analyzed how the temporal information can be handled in a framework that is not originally designed and built to handle temporal features, proposing several possible solutions, some more partial than others. Finally, we analyzed the answers that the framework could give us depending on the environment in which it operates, trying to identify those circumstances in which the framework could operate in a way that does not conform to our expectations. We have also tried to offer possible solutions or prospects for future growth that might refine the process more and more.

At this point that is the end of the chapter, we added few important pieces to the framework workflow: $a)$ We have seen the structure of the curve that has to be extracted through Bayesian inference, $b)$ We discovered how Thompson-Sampling can be a good trade-off between exploration and exploitation, $c)$ We discussed two different approaches for finding the optimal point in the curve, $d)$ We briefly discussed about the relationship between cost and budget.

Chapter 4

Implementation Choices and Development

This chapter will discuss the implementation and technical choices that were made in developing the framework. Some of the choices can be considered contextual to the data at hand, while others are to be considered standard for this type of work. It is not given that the solutions adopted in this context are the best in absolute: each context is different from the other, but also the needs are different according to the scope of application.

4.1 The Data

For this proposed implementation we will make use of freely available online data related to the e-commerce Google Merchandise Store. Google makes freely available for educational purposes the Google Analytics Demo account of its online e-commerce of Google-branded products. The dataset shows daily data related to conversions and costs of a Google advertising campaign, in the period December 2021- March 2022. The campaign has not been running continuously, so complete data is not available for the entire time-frame. The campaign under analysis is not targeting a single type of goal so the conversion value is not fixed. Therefore, the value of the single conversion is set at an indicative value of $75. Then, both Return and Cost are expressed in dollars.

At a first glance at the distribution of the data, we realize that there are clear fluctuations.

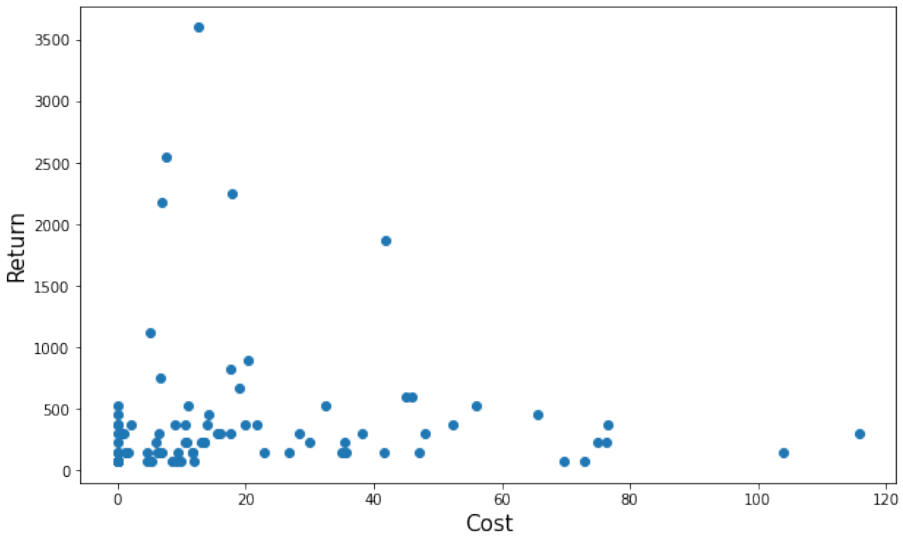

Figure 4.1: Campaign Return vs. Cost daily data plot

This noise could be very annoying for the extraction of the curve, so it is necessary to clean the instances that could disturb the inference. To accomplish this task was chosen the Local Outlier Factor algorithm[1] (LOF). LOF is based on comparing the relative density of a point with that of its neighbors. The only parameter to which the algorithm is subject is precisely the number of neighbors to be considered. Therefore it is worth comparing the number of outliers that are detected as the number of neighbors considered changes. The implementation chosen is offered by the popular Machine Learning Python library Scikit-Learn[2].

As can be seen from Figure 4.2 a definitive number of outliers is not defined, due to the unsupervised nature of the algorithm and the shape of the data. However, we chose to consider 10 neighbors: it is a substantial portion of the dataset (composed of 84 instances), but also represents the first visual local optimum in the plot in which a small variation in the number of neighbors does not change consistently the number of outliers. In this way, 11 outliers are eliminated from the dataset. The second point of local optimum would be around 30 neighbors, however, this would delete a too large portion of the dataset, including in the noise instances that are not part of this cluster. As good habit suggests, the data were

[1] Breunig et al., "LOF: Identifying Density-Based Local Outliers".

[2] Pedregosa et al., "Scikit-learn: Machine Learning in Python".

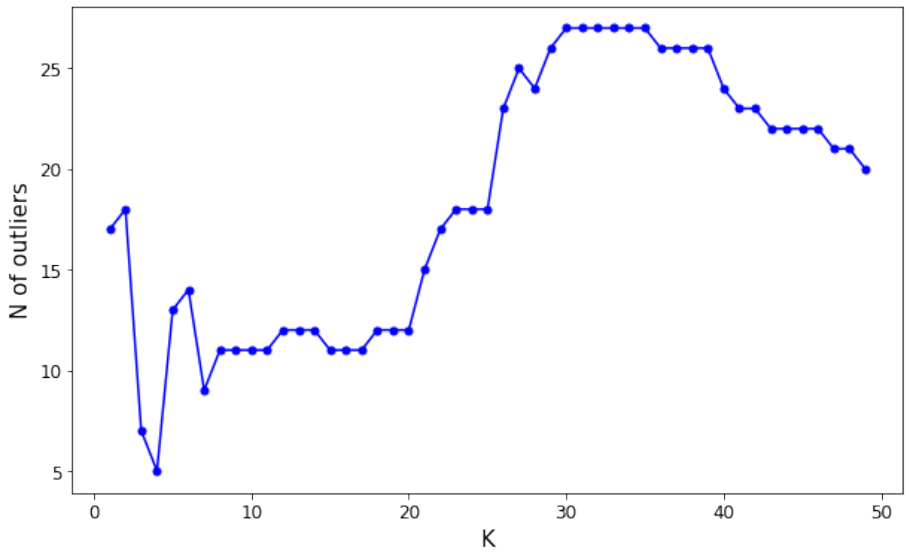

Figure 4.2: Number of Outliers vs. Number of Neighbors in LOF

normalized with the classical implementation of the Standard Scaler($z = \frac{x-\mu}{\sigma}$, where μ is the average of the samples and σ is their standard deviation) developed by Scikit-Learn to make LOF algorithm perform better. Once the data cleaning has been performed, the dataset looks like in Figure 4.3.

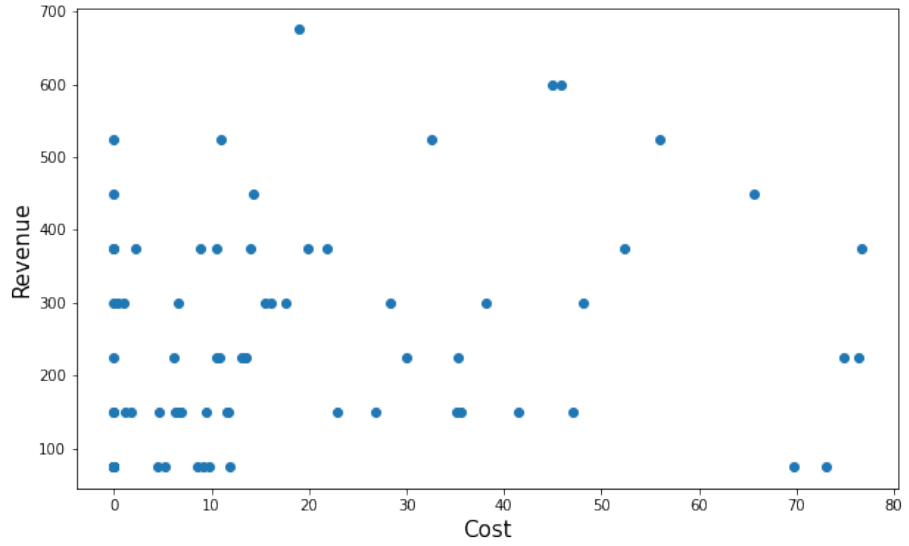

Figure 4.3: Dataset shape after outliers detection

Fortunately, the data is provided already pretty clean so it doesn't need much preprocessing. At this point, we can move on to the cornerstone of the chapter which is the definition

of the model. A final consideration: In such an environment, we are much more confident to work with consistent data. In general, applying this step to every round has to be considered important. This avoids heavy causal swings. Also, if we were to eliminate instances generated by a different process (and not outliers), in a relatively short time LOF would no longer consider them outliers but part of the core dataset. The process of this algorithm is very contextualized to the number of nearest neighbors of each instance and the relative proximity of its neighbors. From the author's perspective, older but established data is preferred over new data that may significantly increase the variance but then turns out to be just noise. Surely the risk is that so the learning process is slower, but probably slow learning, with appropriate adjustments, is to be preferred to fast learning, but very unstable.

4.2 The Model

PyMC3's architecture for instantiating a Bayesian probabilistic model is very straightforward. A model is instantiated, model priors must be defined, and then likelihood. This is the heart of Bayesian inference and is defined in a handful of lines of code.

```
import pymc3 as pm
with pm.Model() as model:

    alpha=pm.Beta('alpha', alpha=3, beta=2) ##prior
       probability of alpha parameter
    epsilon=pm.HalfNormal('epsilon', sigma=10) ##prior
       probability of likelihood sd
    lam=pm.HalfNormal('lam', sigma=20) ##prior probability of
       lambda parameter
    link=pm.Deterministic('link',  lam*cost**alpha) ##
       definition of link function
    likelihood=pm.Normal('conversions', mu=link, sd=epsilon,
       observed=conversion) #definition of the likelihood
```

Code 4.1: Model initialization with PyMC3

Since our function has a structure $y = \lambda x^{\alpha}$ the parameters we are interested in are λ and α on which we have placed two nice prior probabilities. In the Bayes formulation, those probabilities are $p(\theta)$, where θ is a vector of the parameter that are treated as random variables. What do we know about λ? We know that it must be positive since $f(x)$ must always be positive (we can't have negative conversions), and we know that it can assume any possible positive values, so we need a continuous distribution. Furthermore, we know that it cannot assume exaggeratedly large values, given the data we have available. The choice of its prior probability, therefore, falls on a $Half\ Normal$ distribution, sometimes also referred to as $|Normal|$. This function can only be positive, and has a behavior identical to the positive half of a $Normal$ curve: it decreases very fast. The $Half\ Normal$ is nothing more than an always-positive $Normal$ with $\mu = 0$, so it has a support $x \in [0, \infty)$. According to this, we put a prior with $sigma = 20$. This value makes the prior free enough, but not uninformative. We could have increased σ much further, but we want to construct curves that manage to be consistent, even with high variance data. This is why we prefer to give slightly more informative priors, rather than imposing manual boundaries later in the process. Other distributions can only take positive values. A family of functions very commonly used in these cases is the exponential one. However, in this context, after some experiments, the $Half\ Normal$ seems to work better than the other competitors. So, the first prior probability is defined as $\lambda \sim |N|(\sigma = 20)$[3].

The second prior probability is related to the α parameter. In this case, we have more information and the choice is definitely easier. We know that the value of the exponent must necessarily be between 0 and 1 to maintain a congenial structure. We know that it can certainly take any possible value within this range. Moreover, it would be useful to add a bias on what we think should be the slope of the curve. Beta distribution seems to fit the bill. Beta has two parameters $\alpha = n\ successes$ and $\beta = n\ failures$. In our context,

[3]We recall that the symbol $\sim$ should be read as $is\ distributed\ as$ or $follows\ a$. So in that case the notation is $\lambda\ follows\ an\ Half\ Normal\ distribution\ with\ \sigma = 20$.

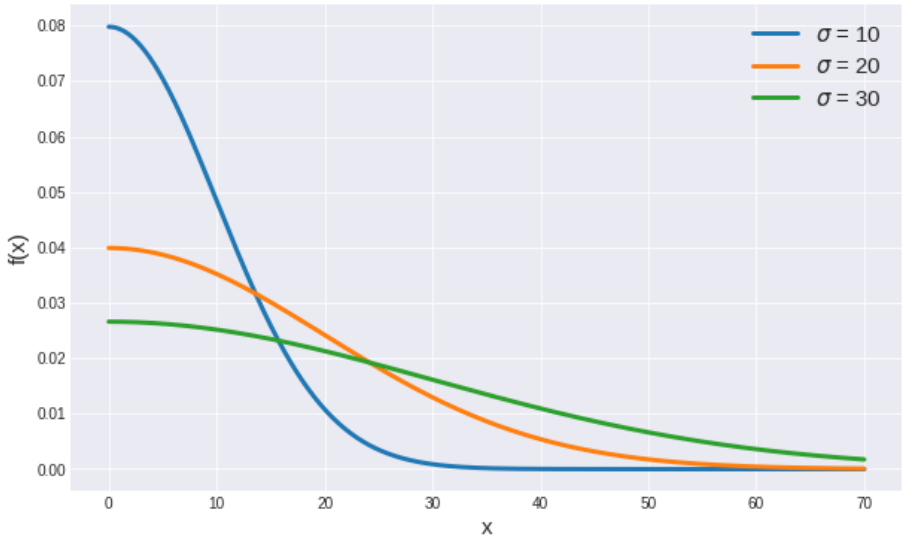

Figure 4.4: $Half\ Normal$ distribution with different σ values

we do not talk about successes or failures, but the fact is that the mean of the function is given by $\mu = \frac{\alpha}{\alpha+\beta}$. So, the greater is α and the smaller is β, the greater will be the mean. The mean can be equal to 0 or 1 just in the case when α or β are equal to 0, but, spoiler alert, they are never zero, and the distribution requires α and β values to be at least equal to 1 to make everything work fine. In general, increasing α means moving the means of the function toward 1 and increasing β means moving toward 0. Since we don't have an hypothesis about the slope of the extracted curve, it seems reasonable to put $\alpha\ and\ \beta$ equal to 2. Doing so we impose a very low informative prior over the exponent, but we just assume a greater density around the mean $\mu = 0.5$. The formalization of these assumptions is $\alpha \sim B(\alpha = 2, \beta = 2)$.

Next, we defined what is referred to by some as a link function. It is nothing more than the function we want to describe, it is the backbone of the likelihood. In the notation, we used $cost$ as a vector where the cost coordinates of our data are stored. We save the function in our model as a deterministic variable, so that it is recorded as an attribute, but in practice, we can easily write it without using pm.Deterministic. Finally, that leaves likelihood: $p(D|\theta)$. It must be said that if we were to assume that the data were perfectly distributed along the curve described by the function $y = \lambda x^{\alpha}$ the link function would correspond to the likelihood. However, we know that the data are naturally subject to a physiological error

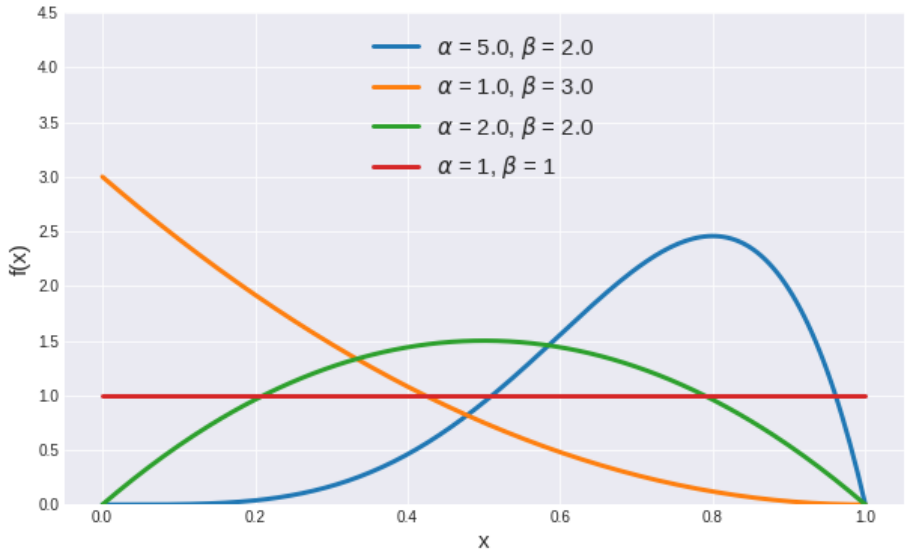

Figure 4.5: Beta distribution with different values of α and β

and generally it is a good approximation to assume that this error is Gaussian distributed. It may be very difficult to define the distribution of the noise since it is counter-intuitive to imagine a generative process behind something that has to be treated as noise, but since we have to manage it somehow, we hope that the error is normally distributed. At this point the function becomes $y = \lambda x^{\alpha} + error$. In this way, the likelihood becomes a $Normal$ distribution whose mean is the link function, but whose error is governed by the standard deviation ϵ. So the $likelihood \sim N(\mu = link, \sigma = \epsilon)$. In the implementation we used a parameter `observed=conversion`. Since the likelihood is the plausibility of the data, of course, we need to refer to the real data somehow: the link function contains the information about the cost, while the parameter `observed=conversion` gives the information about the dependent variable.

At this point, there is only one element to consider that we have purposely left out. The $Normal$ distribution is by the mean but it also needs a standard deviation. Defining the standard deviation of the error distribution, however, seems an arduous task, so the best idea is to place a prior on it! Since we don't have any specific information related to the distribution of the error, it may seem reasonable to place a prior that is not much informative. On the other hand, the distribution has to be consistent enough to generate curves that are somehow coherent and compact. Another important piece of information is that the stan-

dard deviation by definition can be only positive, so, also, in this case, an $Half\ Normal$ distribution can fit well. Now, we want consistency, but also something not much informative: a right middle ground to these two necessities seems a standard deviation $\sigma = 15$. Finally, we have a distribution like $\epsilon \sim |N|(\sigma = 10)$.

4.3 Sampling Process

Once we've set up the model structure the big work is done, but we have not still computed the posterior distribution. Since we make use of MCMC algorithms for the computation, we are going to sample a big number of draws from the posterior distribution. In this way, we avoid the indigestible integral that is required by the computation of $p(D)$. In our case, the sampling process will be processed by the default sampler of PyMC3, called NUTS (No-U-Turn Sampler)[4], an MCMC family algorithm, that is very efficient, usually requires no human intervention, and suits only continuous distributions. We are not going into details here, but NUTS is an extension of the Hamiltonian-Monte Carlo algorithm which is itself a variation of the Metropolis-Hastings algorithms that we described previously. But how can we choose the right number of samples? Well, there is no rule for that but, empirically: as many as possible. The sampling process is computationally expensive since it is repeated thousands of times to get a reliable distribution, so it requires a lot more time than many other machine learning algorithms, but it is not always necessary to ask the algorithms for one million draws. Of course, if we sample many times and if we use several chains the probabilities of divergences or stocking in local optimums is lower. From my own experience, if at least three chains are used, and all three agree in describing a similar distribution, we can rest assured that we are probably going in the right direction (it is very unlikely that three chains randomly initialized in space will end up stuck in the same wrong place). In such a case, we don't need hundreds of thousands of draws to get a consistent approximation, and we can save a lot of time.

In PyMC3 the sampling process is performed as follow:

[4] Hoffman and Gelman, "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo".

```
with model:
  trace = pm.sample(40000, chains = 3, cores = 3, tune=10000,
     target_accept= 0.90, return_inferencedata=True)
```

Code 4.2: Sampling process

We chose to sample 40000 (just to be sure), with 10000 tuning steps. The tuning phase is roughly what happens before the sampling. Since the initialization process is random, and the starting points of the random walk may be far away from the optimal point, we may risk keeping some initial draw that is not consistent with respect to the real distribution. The tuning phase tries to avoid this inconvenience: it draws an initial number of samples that are then discarded during the sampling phase. Those draws describe values that are evidently out of range, and they would be a waste of time during the sampling. We can figure the tuning phase as an initial and rough analysis that discards the least probable values of the distribution, to focus the sampling phase in the right spot. The other parameters are the following. `chains=3` is the number of Markov Chains we want to initialize. `cores=3` is the number of computer cores we want to make the chains run to (in this case one chain per core). `target_accept=0.90` is a parameter for the tuning of step size in NUTS (support $\in [0, 1]$), but the important this to know is that a higher value will work better for problematic posteriors. `return_inferencedata=True` is a parameter to make return an InferenceData object which is easier to manage for Arviz due to the integrations with ArviZ. ArviZ is a Python package extremely useful for the graphical support of the Bayesian analysis[5]. It helps us a lot when we have to understand what is going on under the hood and it is always a good idea to look for a good built-in plot in the API documentation.

4.4 Results and Debugging

The first reasonable thing to do once we have sampled is to check the distribution of the posterior distribution. Since we used three different chains to perform the process, we hope that the three chains did pretty much the same job. To do this we resort to the comfortable

[5]Ravin et al., "ArviZ a unified library for exploratory analysis of Bayesian models in Python".

ArviZ library. With just a short command we can plot the smoothed plot of three chains' sampling distribution, and the random walk through the possible values (we sampled 40000 times, so the random walk is not that understandable, but it can help to visualize the variations of draws). Remember that the trace is the Inference Data object where we stored the chains.

```
import arviz as az
fig = pm.plot_trace(trace, var_names=['alpha', 'lam'], figsize
   =(19.5, 7.5), legend=True, kind='trace')
```

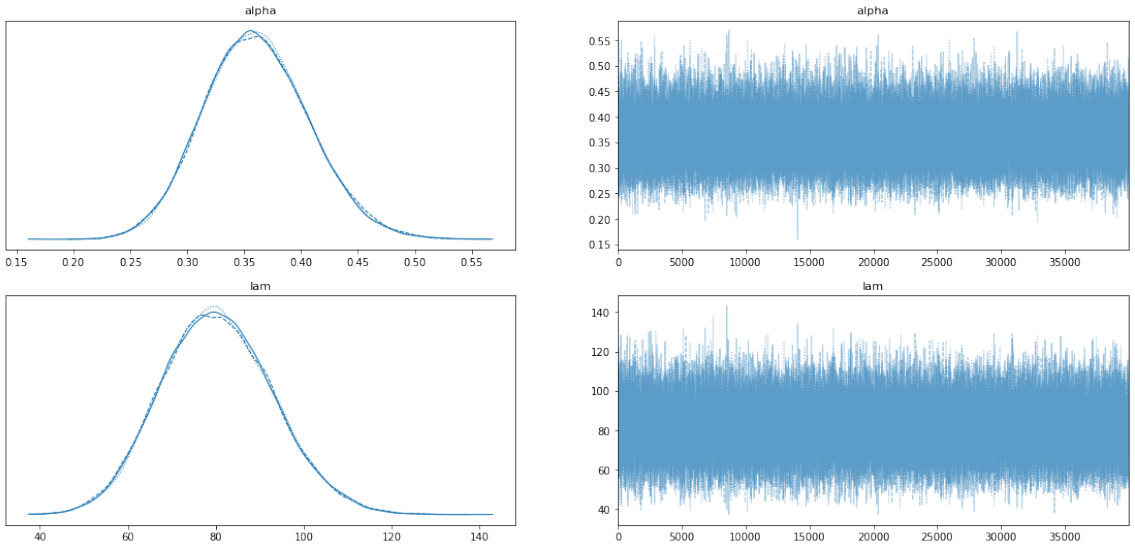

Figure 4.6: ArviZ trace plot. On the right there are the distribution, and the random walk on the left.

We choose to plot the two most interesting variables. Fortunately, the chains behaved in the same way and we can see it by the fact that in the distributions the three curves, each of which represents a different chain, overlap most of the time. Moreover, we did not encounter divergences, and this is another clue to the goodness of the sampling process. In addition, we notice that the variance of λ is quite relevant, while the variance of α seems much more balanced. We are going to see soon how this can impact the curve extraction, but for now let's see where most of the draws are concentrated, where the posterior probability is denser. Since we are not Frequentists, but convinced Bayesians, we don't speak of a confidence interval, but a highest density interval (HDI). This notation is much more understandable: we are sampling from the posterior probabilities, and we can analyze

how many draws fall inside a defined range. The range is custom-defined, and there are no boundaries to respect to have a statistical significance. Bayesian statistics does not set what is true or what is wrong, it just describes how the events are described: it's up to us to define what can be acceptable or what cannot be.

```
fig = az.plot_posterior(trace, var_names=['alpha', 'lam'],
   hdi_prob = 0.90)
```

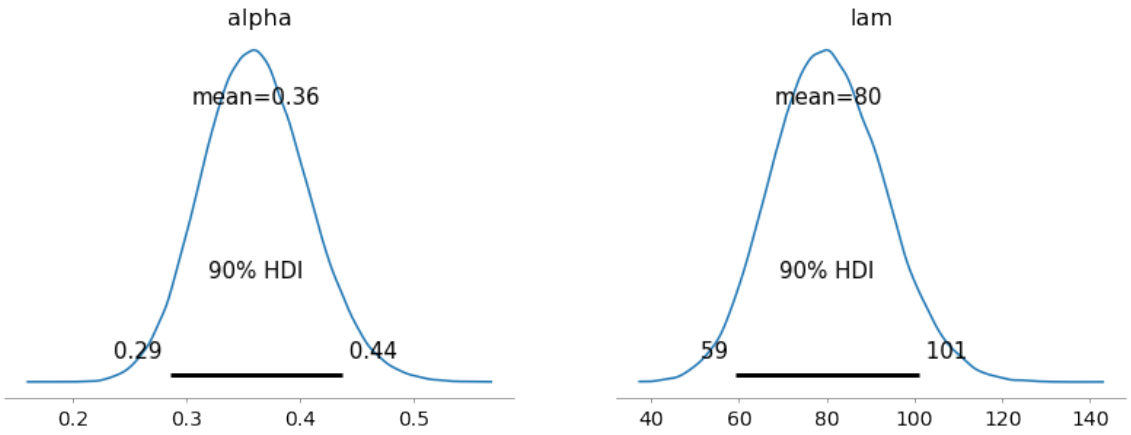

Figure 4.7: ArviZ trace plot. On the right there are the distribution, and the random walk on the left.

The ArviZ posterior plot describes how the variables behave inside a custom range: our range is set to be 90%. In such an environment, we are discarding the most 5% tailed samples. What can the plot tell us about the posterior distribution? The λ is not precisely defined, probably because the data show a greater disorder close to the origin, so there is no clear starting point of the power law. On the other hand, the distribution of α is denser around a narrower range of values, especially considering the low scale of the values[6]. At this point, it is important and meaningful to know what is the relationship between our parameters. In principle, we should have the intuition that α and λ could be somehow dependent if we have to extract a curve from some data. Since λ increases linearly, while α exponentially, there should be some kind of exponential function that bounds these two parameters. But, remains significant to see, thanks to the sampling process, how this rela-

[6] As we pointed out in Chapter 3, variations of decimals for low values in the exponent are significantly less impactful than same variations for higher values.

tionship fits our data. Well, luckily ArviZ comes with a specific plot for this purpose. Let's plot the variables in pairs, but this time considering also the ϵ: we may be interested also to know if some parameter is related to the standard deviation of our data.

```
ax = az.plot_pair(trace, var_names=["alpha", "lam", 'epsilon'
    ], kind=["scatter", "kde"], marginals=True, point_estimate=
    "median", figsize=(17, 9))
```

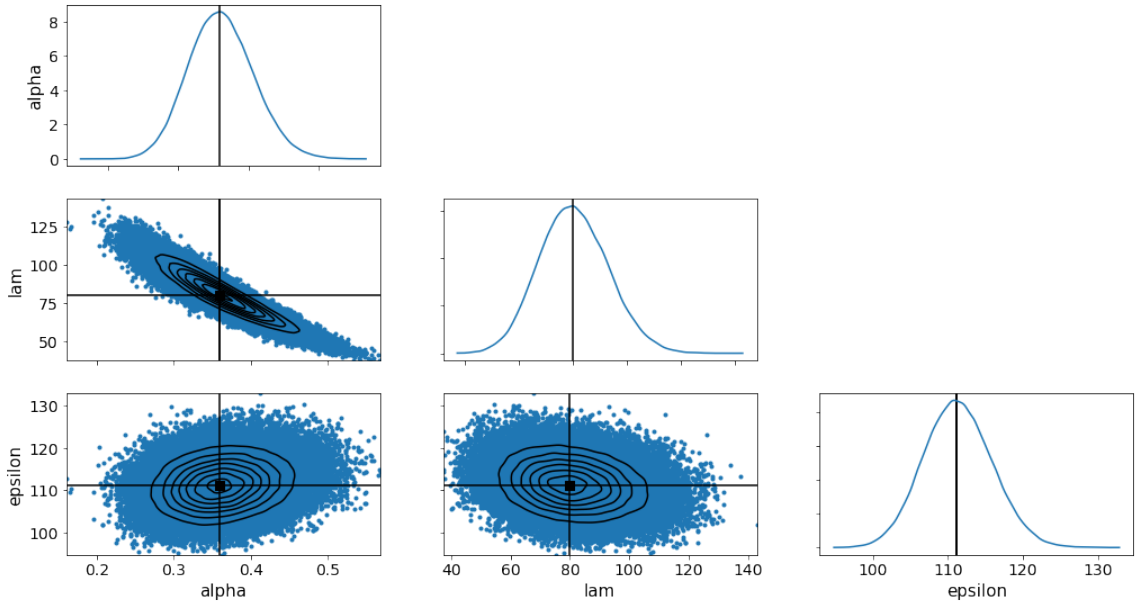

Figure 4.8: ArviZ pair plot. The scatter plots represent the draws with respect to a pair of parameters, while the density functions are the marginal probabilities of the single parameters.

The pair plot confirmed the intuition: α and λ are dependent and we have to maintain this dependency once we extract the curve. This is evident by the shape of the point cloud in the scatter plot[7]. It has a well-defined direction that expresses inverse proportionality: if α increases, λ decreases, and vice-versa. On the other hand, ϵ seems to be independent of the model parameters, and this seems to make sense: since the likelihood is described as a $Normal$ distribution, the μ encapsulates the information related to the model (α and λ in our case), while the standard deviation the information that describes the dispersion of the data we are dealing with, under the condition of the model.

[7]The cloud of points represents the joint posterior probability of the parameters. The concentric ellipses show the probability density.

One last thing about the debugging mode. Since the draws are not completely independent, sometimes we can face problems of autocorrelation. In MCMC the draw $n + 1$ will be dependent only by the position of the draw n, but in the long run, in some cases with variables that are subject to correlation, the dependency may be broadcast along the chain implying the factual non-independence of the samples. This can affect the results since the initial position is randomly initialized, but the initial conditioning should be lost soon. Fortunately, in the more advanced sampling methods, the eventuality of non-independence is quite rare, but it's worth checking if something went wrong, especially because our α and λ are inversely proportional.

```
autocorr = az.plot_autocorr(trace, var_names=['alpha', 'lam'],
    max_lag=50)
```

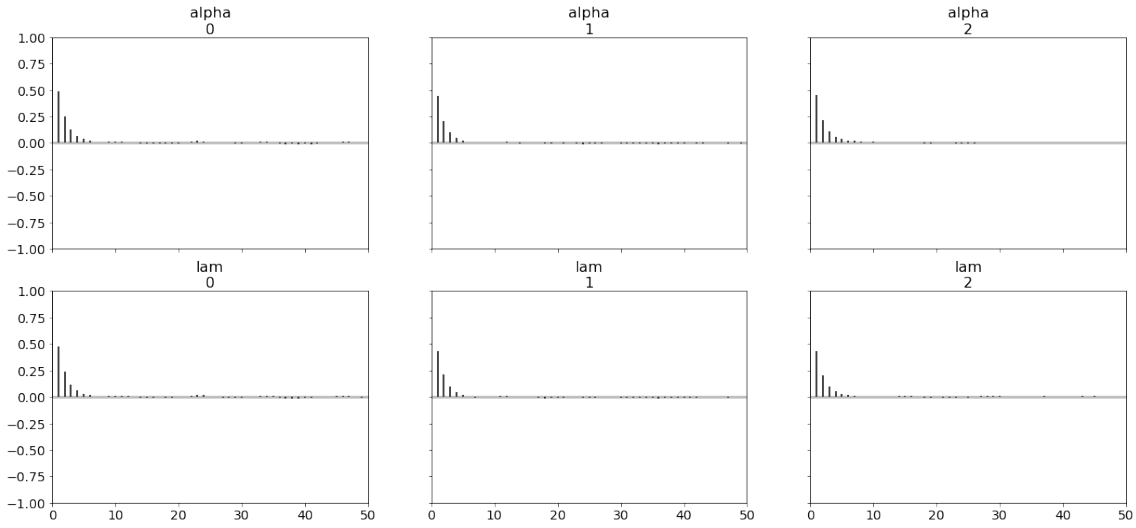

Figure 4.9: ArviZ autocorrelation plot.

Since the autocorrelation dramatically decreases after just a bunch of draws, it seems that everything is working well.

4.5 Curve Extraction

The time has come! Finally, we can see the results of the framework. Let's extract the curve. We are going to randomly select one specific draw. Every draw contains information about

all the priors since it is extracted from the joint posterior probability of all the priors we put. Extracting one draws at the time we are making compete one curve at every round, converging in the long run to the densest part of the marginal posterior distribution of the parameters. This is the Thompson-Sampling algorithm we have treated in Chapter 3, and the Figure 4.8 can help us understand where the densest part of the distributions are, and consequently where most of the draws (and the resulting curves) are concentrated. To visualize the boundaries of this distribution it is significant to plot a large number of curves randomly extracted from the joint posterior distribution so we can see how the model can behave in the space.

```
import random
import matplotlib.pyplot as plt
import numpy as np

fig, ax = plt.subplots(figsize=(12,8))
c = int(random.choice(trace.posterior.chain))
for i in range(300):
  d = int(random.choice(trace.posterior.draw))
  sample = trace.posterior.sel(chain=c, draw=d)
  lam = float(sample.lam)
  exp = float(sample.alpha)
  vector = np.linspace(1, 1000, 1000)
  conversions = lam*vector**exp
  ax.plot(vector.reshape(-1,1), conversions.reshape(-1,1),
     markersize= 1.5, color='green', alpha=0.1)
ax.scatter(cost, conversion, color='blue')
conversions_ad_mean=float(np.mean(trace.posterior.lam[c]))*
   vector**float(np.mean(trace.posterior.alpha[c]))
ax.plot(vector.reshape(-1,1), conversions_ad_mean.reshape
   (-1,1), markersize = 3.5,  color='red', label='Average
   Extracted Curve: y={0}x^{1}'.format(round(float(np.mean(
   trace.posterior.lam[c])), 2), round(float(np.mean(trace.
```

```
   posterior.alpha[c])), 2)))
ax.set_ylabel("Returns", fontsize=15)
ax.set_xlabel("Cost", fontsize=15)
ax.set_xlim([-1, 150])
ax.set_ylim([15, 900])
ax.legend(loc="best", fontsize=12)
fig.show()
```

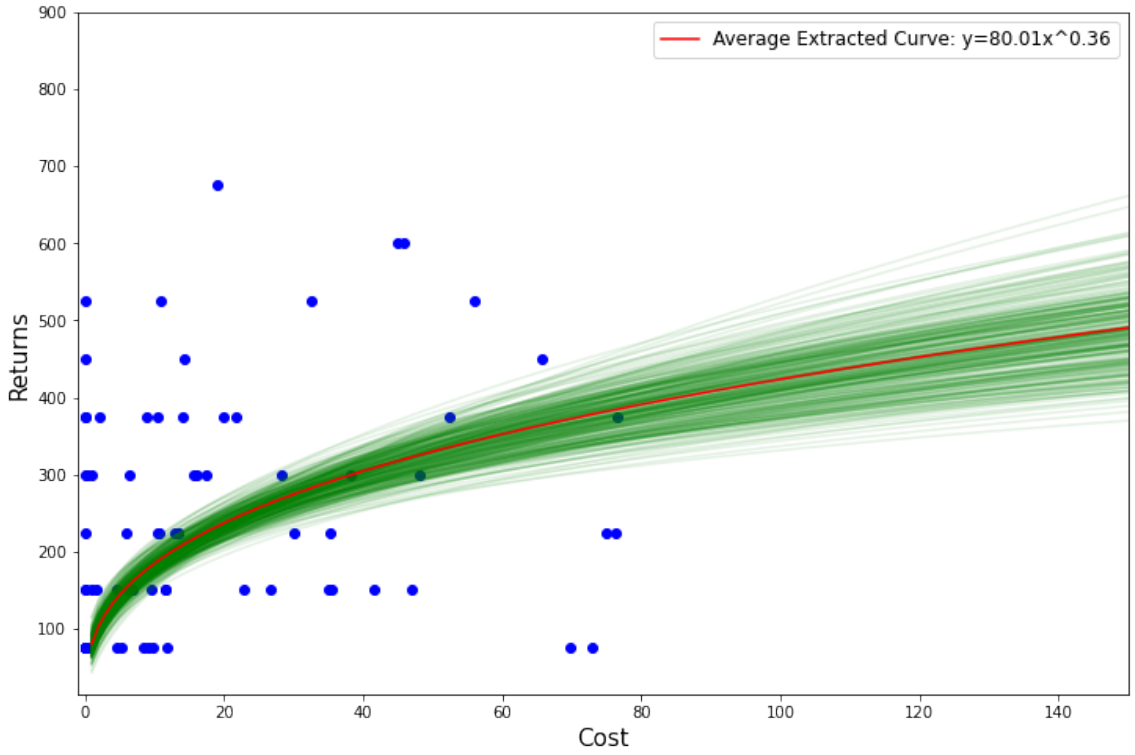

Figure 4.10: Curve random extrapolation of 300 draws

We can see from Figure 4.10 where the curve is denser. Since the three chains worked the same way, we randomly chose one of them to perform the extraction: the result would not change relevantly otherwise. We did not impose any manual boundary to the distribution so what is shown in Figure 4.10 is a true set of draws from the joint plot distribution. Since the data do not show a clear power-law shape, we have to deal with a great variance, especially for lower costs. This is probably why the λ suffers such a great fluctuation in its distribution, but we can see that in practice this doesn't affect a lot the curve extraction.

Even the α is quite stable in the representation we gave: of course, the variation is less evident while we are inside the point cloud, and the uncertainty increases for costs never experienced before. However, if we look at the densest region (the darkest green region) even for high Costs, we can assert that the largest part of the curve falls into a reasonable region. Recall that we wanted to represent the entirety of the distribution, but in practice, it is always a good idea to put some boundaries, even very broad, to discard the distribution heavy tails[8]. Finally, we also decided to plot the average curve extracted, whose parameters are, not surprisingly, pretty much the same as the densest of the scatter plots in Figure 4.8.

4.6 The Optimal Point

Now it's time to find the budget to allocate. We re going to show in the same plot both the approaches we discussed before: the marginal approach, and the average Cost Per Acquisition approach. In this way can be even more evident why the marginal approach is much more safe, profitable and cautious.

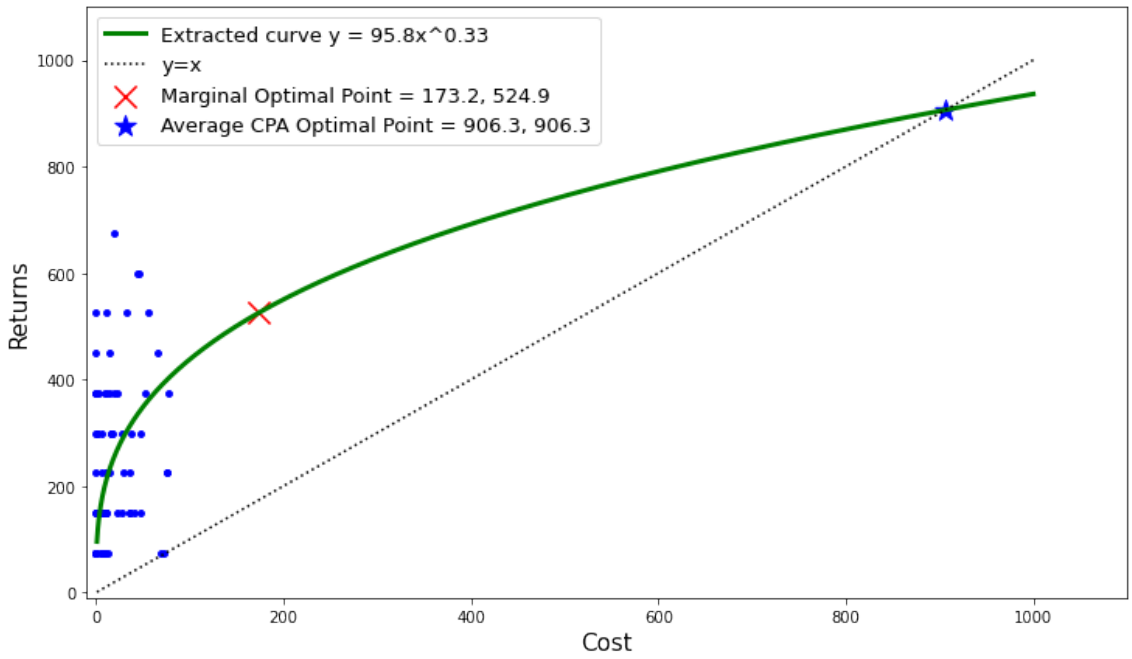

Figure 4.11: Optimal Point for Marginal and Average CPA approach.

[8]Remember also that we used a large number of tuning steps, that before the sampling process reduced the possible exploration area

We leave the code at the end of the section since it can be a bit tricky and heavy to post here. However, this is the resolution of the framework: we spot two optimal Costs following the two approaches. But why the marginal approach seems to be the best under these conditions? Well, in addition to the prescriptions we gave in Chapter 3, we can see from Figure 4.10 that naturally the uncertainty increases where data are not present. This means that the average CPA approach risks producing results that are subject to great uncertainty. Since we are spotting a single point, the uncertainty can be deleterious and we should try to reduce it as much as we can. This argument can be used once again to prefer the marginalistic approach over the other: it is more conservative, and it can help us to learn as much as we can about our present condition.

In our implementation we used the Sympy[9] library to solve the derivative necessary for the computation of the optimal point in the marginal approach, and the optimize.fsolve function from Scipy[10] library to solve the system of equations for the average CPA approach.

```
from scipy.optimize import fsolve
from sympy import *

## Function for defining the structure of the solver for
   system of equations
def solver(z):
  vector = z[0]
  conversions = z[1]
  f = np.empty(2)
  f[0] = vector-conversions
  f[1] = lam*vector**exp-conversions
  return f
## Inizialize the curve
vector = np.linspace(1, 1000, 1000)
c = int(random.choice(trace.posterior.chain))
d = int(random.choice(trace.posterior.draw))
```

[9] Meurer et al., "SymPy: symbolic computing in Python".

[10] Virtanen et al., "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python".

```
sample = trace.posterior.sel(chain=c, draw=d)
lam = np.round(float(sample.lam), 1)
exp = np.round(float(sample.alpha), 2)
## Plot the figure
fig, ax = plt.subplots(figsize=(12, 7))
## Plot the extracted curve
conversions = lam*vector**exp
ax.plot(vector.reshape(-1,1), conversions.reshape(-1,1),
   linewidth= 3, color='green', label="Extracted curve y = {}x
   ^{}".format(lam, exp))
## Plot the data
ax.scatter(cost, conversion, color='blue', s=15)
## Find the optimal point with marginalistic approach
x = Symbol('x')
y = lam*x**exp
yprime = y.diff(x)
x = solve(yprime-1)
x_1 = float(x[0])
ax.scatter(x_1, lam*x_1**exp, label=r'Marginal Optimal Point =
    {}, {}'.format(np.round(x_1, 1), np.round(lam*x_1**exp, 1)
   ), s=200, color='red', marker='x')
## Solve the system of equations for the optimal point in
   average CPA approach
z_guess = np.array([2000, 2000])
z = fsolve(solver, z_guess)
## Plot the y=x curve and optimal point
ax.plot(vector, vector, label=r'y=x', color = 'black',
   linestyle='dotted')
ax.scatter(z[0], z[1], label=r'Average CPA Optimal Point = {},
    {}'.format(np.round(z[0], 1), np.round(z[1], 1)), s=200,
   color='blue', marker='*')
plt.xlabel('Cost', fontsize=15)
```

```
plt.ylabel('Returns', fontsize=15)
ax.set_xlim([-10, 1100])
ax.set_ylim([-10, 1100])
plt.legend(loc='best', fontsize=13)
plt.show()
```

Code 4.3: Code for Figure 4.11

4.7 Cost and Budget

So far we treated the Cost and the Budget as the same entity. In practice, it is not true. It may happen that if the Budget is too high with respect to the bid, we may not spend all of the total Budget because we are not in the position to spend all the money. In the environment of a smart bidding platform, things are much more gray: we are not completely conscious of what is happening under the hood of a machine learning model that has the task to maximize a function. However, the problem is real and it is very important to understand that Budget and Cost are not the same. Budget is the quantity of money we don't want to overcome, on average. Cost is what in reality, under every possible contingency, we are going to spend. Sometimes media planners want to live in the limbo between the Budget as a maximum, and the Budget as a target. This view is not coherent, especially during an optimization process. Moreover, in an advertising platform, we can regulate the Budget, but the Cost is something not under our control. This can be a problem for our framework since we should deal with Costs and Revenues, but we can operate over Budget. So we should also study the relationship that lies between Costs and Budget. In this case, the real daily Cost would be the dependent variable, while the Budget would be the independent variable and the process of relationship extrapolation could be performed as well as we performed the previous function extrapolation. The function that links Budget and Cost is not trivial to identify and it is something complex to deduce reversely from the data, but on the other hand, if the data show a general path we can use much easier functions, instead. Sometimes a linear regression can be enough, and in this case, the extrapolation could be made with the following model, but the process would be the same for any kind

of underlying function:

```
with pm.Model() as model_fun:
  # Priors
  slope = pm.Normal('slope', 0, sd=5)
  inter = pm.Normal('inter', 0, sd=10)
  sigma = pm.HalfNormal('sigma', sd=5)
  # Link function
  link = pm.Deterministic('link', inter + avg_cost * slope)
  # Likelihood
  likelihood = pm.Normal('y', mu = link , sd = sigma, observed
      = budget)
  # Sampling
  trace_fun = pm.sample(10000, chains=2, tune=1000, cores=2,
     return_inferencedata=True)
```

4.8 Summary

In this chapter, we first looked at the features we should expect from the data we have available and any cleanup we need to do before making any inferences. Subsequently, we have analyzed the structure of the model, bringing motivations and arguments about the choices made (mainly on the choices of prior probabilities and likelihood) highlighting what may be the weaknesses and strengths of the system. In addition, we have seen how the sampling process is performed and which tricks to keep in mind in order not to fall into errors of divergence, or local optimum. Then, we have seen how to analyze the results we obtain, trying to study the trend of the chains, the posterior distributions, the variance of the draws, the proportionality between the parameters, and how to notice cases of collinearity. Thanks to this we were able to extract a large sample of curves to observe what can be the behavior of the extraction, noting also the areas of greater density and consistency with the posterior distributions. This led us to the identification of the optimal point with the two methods, visually explaining why once again the marginal method turns out to be the most appro-

priate. Finally, we have analyzed the relationship between Cost and Budget, highlighting the fact that they are not two stackable concepts and that they refer to different choices and implications, making it necessary to study even approximately their relationship.

Conclusions

At this point, it would be natural to wonder if the framework so far described and carefully documented has already been tested in some way. In fact, this has happened, however, the framework has not given the results that could have been expected. The reasons are many and the ”failure” of the case study can give us some food for thought on what can be good practices to keep in mind. The test advertising campaign aims to bring customers to a specific landing page of the site on which they have to act on their user profile related to the status of their food vouchers (some specific actions correspond to a conversion). Each conversion is calculated to be worth approximately €100.

Seasonality The campaign that has been used for testing suffers from high seasonality. This is because its objective is closely related to the working season. For this reason, at least 4 different periods of the year can be identified in which the data generation processes vary significantly. We have a first segment in the second half of January-April, in which performance remains around the average. In the May-August period, conversions collapse, due to the seasonal vacations. It returns around an average in the period September-October, while in the period November-mid January performance is much more consistent than the average. This high and evident seasonality has not allowed us to accumulate a sufficient amount of data to be able to refine the allocation process. Even if the framework had found, at a certain point, a balance, after a short time everything would have changed and it would have been necessary to start the learning process all over again. Furthermore, given the pandemic situation still in place, the eventual use of historical data is virtually impossible.

Covid-19 pandemic We link back to the previous point through the problems caused by the Covid-19 pandemic. This extreme and destabilizing event generated consistent

disruption in data organization. In fact, the benchmark for performance evaluation of the optimization was made in the period February-March 2021, when the vaccination campaign was still in its beginning and we were living, in Italy, the third pandemic wave. In addition, many people were still in smart work, which greatly reduced the need to check their meal tickets. This made it difficult to assess the actual success of the framework: the benchmark itself refers to a period that cannot be taken as a reference. In this case, the ideal would be to conduct an A/B test on two identical campaigns, but this was beyond our possibilities.

Event Shortage Another element that negatively affected the success of the experiment was the very low amount of conversion events. This makes it very complicated to determine when a consistent and momentary variation of conversions can be determined by a daily fluctuation determined by the context and exogenous causes (and therefore to be classified as an outlier), or if the variation was actually determined by the optimization of the framework. In the campaign in question, after outlier detection, the average daily conversions were less than 2 units and events above 4 conversions were very rare. In such a context, real optimization can be safely confused with random fluctuations. If a campaign converts so little on a daily basis, there are probably upstream issues that make any optimization process difficult. In fact, even though built-in smart bidding strategies are active, the results are still poor and unsatisfactory. An ideal context for the testing campaign involves a number of daily conversions that would not be disturbed by variations in individual conversions.

Weak Purpose Closely related to the previous point is a more general discussion about the nature of the campaign. The framework works to its full potential when we are in presence of a real e-commerce campaign, or something very similar to it. The objective should be much clearer than it is in the campaign in analysis and it should be oriented to the sale of products, without lateral purposes. The idea of the testing campaign is that of an environment where the objective is both to convert, but also increase awareness and have a solid online presence. Furthermore, the purpose of conversion seems difficult to determine, since it does not seem to address a need that

would not be fulfilled in the same way by a simple online search. The campaign's target audience is those who are already customers of the company, not new customers to be acquired. For this reason, if the customer wants to know their status, they will not need an ad hoc advertising campaign but will carry out targeted research at the time of need. This is probably why the campaign struggles to acquire new conversions and often the CPA is higher than the conversion value itself. Furthermore, the value set for each conversion (quantification that has been provided to us) represents a quantitative datum that is difficult to estimate for the very structure of the campaign. For this reason, one of the pillars on which the theoretical structure of the framework is based is weakened.

From these issues, it becomes clear that working in a setting where temporal information is not directly handled can be exhausting and problematic. We have devoted a section to this topic, and I believe it may be a fertile ground for the future development of this work. Efficiently managing data-generating processes that can be determined a priori by specific temporal events can clearly determine brilliant success, from dismal failure. In addition, it might be interesting to test the performance of the framework by varying the sampling algorithms: it is not certain that NUTS is the final choice for our case. In fact, we have presented a model with with a small number of parameters and perhaps even less sophisticated algorithms can arrive at even more satisfactory solutions.

At this point a question may arise: what was the purpose of the experiment if many of these limits were clear from the start? Certainly, to be able to carry out an all-around analysis it was necessary to concretely develop the framework and this test was an excellent opportunity to do so. From the point of view of implementation, the design process was successful, also thanks to the great support of the generous and competent community that revolves around PyMC3. Moreover, in an attempt to find the right combinations between prior probabilities and likelihood, various probability distributions (such as a more long-tailed Student-t distribution) and also hierarchical models have been tested. However, the solution adopted and shown seems the most functional, essential, and convenient. For this reason, it is to be considered an interesting and enriching experience. In practice, it is very difficult for an advertising agency and for its staff, engaged in everyday life and in everyday

problems, to understand and to cope with the full potential of such a framework without a solid basis of statistics and Machine Learning. For this reason, we can be satisfied even in this way and we trust in the skills of some enlightened and multifaceted marketing directors to give confidence to this structure.

Even the author of this essay was quite ignorant about all the fascinating implications of Bayesian inference before starting to work directly on this task. Furthermore, I hope that during the treatment of the thesis it emerged how much the Bayesian inference tools and a Bayesian view of life can also be useful for the humanities. We have not dealt with the subject head-on, but there is a lot to be said about how Bayesian inference offers a much more realistic, and complete view of the reliability of statistical tests. This is especially true in the case of sparsely populated samples, very rare events, or poorly informative conditions. In any case, Bayesian inference offers a huge amount of food for thought and PyMC3, at least from my point of view, was a great starting point to get closer to this world. Initially, especially if you do not have a basis for probabilistic programming, it could be hermetic and obscure, but over time it turns out to be an excellent key to reading that glosses over the mathematical details of the individual algorithms to get to the really interesting part of the question.

Bibliography

Barabási, A. and Albert R. “Emergence of Scaling in Random Networks”. In: *Science* 286 (1999), pp. 509–512.

Bloomberg.com. *Google Delays Phaseout of Advertising Cookies Until 2023*. URL: https://www.bloomberg.com/news/articles/2021-06-24/google-delays-phase-out-of-advertising-cookies-until-2023 (visited on 02/24/2022).

— *Meta Renews Warning to EU It Will Be Forced to Pull Facebook*. Feb. 7, 2022. URL: https://www.bloomberg.com/news/articles/2022-02-07/meta-may-pull-facebook-instagram-from-europe-over-data-rules (visited on 02/15/2022).

Breunig, M. M. et al. “LOF: Identifying Density-Based Local Outliers”. In: *SIGMOD Rec.* 29.2 (May 2000), pp. 93–104. URL: https://doi.org/10.1145/335191.335388.

Britannica.com. *diminishing returns*. URL: https://www.britannica.com/topic/diminishing-returns (visited on 03/11/2022).

Broido, A. D. and A. Clauset. “Scale-free networks are rare”. In: *Nature Communications* 10.1017 (2019). URL: https://doi.org/10.1038/s41467-019-08746-5.

Brue, S. L. “The Law of Diminishing Returns”. In: *Journal of Economic Perspective* 7.3 (1993), pp. 185–192.

Chater, N., J.B. Tenenbaum, and Yuille A. “Probabilistic models of cognition: Conceptual foundations”. In: *Cognitive Sciences* 10.7 (July 2006), pp. 287–291.

Chib, S. and E. Greenberg. “Understanding the Metropolis-Hastings Algorithm”. In: *The American Statistician* 49.4 (1995), pp. 327–335. URL: http://www.jstor.org/stable/2684568.

CNIL.fr. *Use of Google Analytics and data transfers to the United States: the CNIL orders a website manager/operator to comply*. Feb. 10, 2022. URL: https://www.cnil.fr/en/use-google-analytics-and-data-transfers-united-states-cnil-orders-website-manageroperator-comply (visited on 02/15/2022).

Corner, A., J.L. Harris, and U. Hahn. "Conservatism in Belief Revision and Participant Skepticism". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 32.32 (2010), pp. 1625–1630.

Encyclopediaofmath.org. *Beta-function*. URL: https://encyclopediaofmath.org/index.php?title=Beta-function (visited on 03/06/2022).

Eur-lex.europa.eu. *CHAPTER V*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%5C%3A32016R0679%5C&from=FR#d1e4227-1-1 (visited on 02/24/2022).

Everitt, B.S. and A. Skrondal. *The Cambridge Dictionary of Statistics*. IV. Cambridge: Cambridge University Press, 2010.

GDPR.eu. *What is GDPR, the EU's new data protection law?* URL: https://gdpr.eu/what-is-gdpr/ (visited on 02/15/2022).

Gelman, A. et al. *Bayesian Data Anlysis*. III. CRC Press, 2013, p. 675.

Goldstein, M. L., S. A. Morris, and G. G. Yen. "Problems with fitting to the power-law distribution". In: *The European Physical Journal B - Condensed Matter and Complex System* 41 (2004), pp. 255–258.

GroupM. *This Year Next Year. Global End-Of-Year Forecast. December 2021*. Tech. rep. New York, 2021.

Han, B. and C. Arndt. "Budget Allocation as a Multi-Agent System of Contextual & Continuous Bandits". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021), pp. 2937–2945.

Han, B. and J. Gabor. "Contextual Bandits for Advertising Budget Allocation". In: *The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20)* (2020).

Hoffman, M. D. and A. Gelman. “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo”. In: *Journal of Machine Learning Research* 15 (2014), pp. 1351–1381. URL: http://www.stat.columbia.edu/~gelman/research/published/nuts.pdf.

IAB.Europe. *Adex Benchmark Study 2020*. Tech. rep. Brussels, 2020.

Kong, D. et al. “A Combinational Optimization Approach for Advertising Budget Allocation”. In: *Companion Proceedings of the The Web Conference 2018* (2018), pp. 53–54. URL: https://doi.org/10.1145/3184558.3186925.

Kruschke, J. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. II. Cambridge: Academic Press, 2014.

LinkedIn.com. *LinkedIn is Leaving China!* URL: https://www.linkedin.com/pulse/linkedin-leaving-china-sam-maiyaki/ (visited on 02/24/2022).

Martin, O. *Bayesian Analysis with Python: Introduction to Statistical Modeling and Probabilistic Programming Using PyMC3 and ArviZ*. II. Kindle Edition. Birmingham: Packt Publishing Ltd, 2018.

Meurer, A. et al. “SymPy: symbolic computing in Python”. In: *PeerJ Computer Science* 3 (2017). URL: https://doi.org/10.7717/peerj-cs.103.

Newman, M. E. J. “Power laws, Pareto distributions and Zipf’s law”. In: *Contemporary Physics* 46.5 (Sept. 2005), pp. 323–351. URL: http://dx.doi.org/10.1080/00107510500052444.

Nuara, A. et al. “Online joint bid/daily budget optimization of Internet advertising campaigns”. In: *Artificial Intelligence* 305 (2022).

Pedregosa, F. et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

Politico.eu. *French privacy regulator rules against use of Google Analytics*. Feb. 10, 2022. URL: https://www.politico.eu/article/french-privacy-regulator-rules-against-use-of-google-analytics/ (visited on 02/15/2022).

Privacysandbox.com. *Building a more private, open web*. URL: https://privacysandbox.com/ (visited on 02/15/2022).

PwC, IAB. *Internet Advertising Revenue Report*. Tech. rep. London, 2021.

Ravin, K. et al. "ArviZ a unified library for exploratory analysis of Bayesian models in Python". In: *Journal of Open Source Software* 4.33 (2019), p. 1143. DOI: 10.21105/joss.01143. URL: https://doi.org/10.21105/joss.01143.

Salvatier, John, Thomas V Wiecki, and Christopher Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: *PeerJ Computer Science* 2 (2016).

Statista.com. *Digital Advertising - Worldwide*. URL: https://www.statista.com/outlook/dmo/digital-advertising/worldwide (visited on 02/24/2022).

Strens, M. "A Bayesian framework for reinforcement learning". In: *ICML* 2000 (2000), pp. 943–950.

Tversky, A. and D. Kahneman. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment". In: *Psychological Review* 90.4 (1983). Ed. by Inc American Psychological Association, pp. 293–315.

— "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157 (Sept. 1974), pp. 1124–1131.

— "Subjective Probability: A Judgement of Rapresentativeness". In: *Cognitive Psychology* 3 (1972). Ed. by Academic Press, pp. 430–454.

Virtanen, P. et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. URL: https://rdcu.be/b08Wh.

Wang, X., F. Li, and f. Jia. "Optimal Advertising Budget Allocation across Markets with Different Goals and Various Constraints". In: *Complexity* (May 2020).

Wasserman, L. A. *All of Statistics: A Concise Course in Statistical Inference*. Berlin: Springer, 2004.

Zenith. *Advertising Expenditure Forecasts. December 2021*. Tech. rep. London, 2021.

Aknowledgements

I would like to thank InTarget as a whole for the precious time they granted me while I was studying topics that not so easy to digest. Specifically, I want to say thank you to Ilaria and Luca who trusted me and had all the patience I needed. Their presence has made my entry into the workplace much more comfortable than I ever expected. Moreover, thanks to the advertising team, in particular Michelangelo, who accepted to give confidence to me and to my experiment, although the results were not visible.

Thanks especially to Professor Nicola Ciaramella who allowed me to deepen matters so interesting and full of implications. Thanks for the total availability and the long office hours thanks to which there was the opportunity to discuss very fascinating topics without watching the clock. Finally, thank you for your trust and support, despite my area of study and purely humanistic background. Without this thesis opportunity, I probably would not have discovered the enormous potential of Bayesian statistics.

Thanks also to my university mates with whom I had the opportunity to confront for months on such a non-trivial subject, perhaps even boring them a bit. The Digital Humanities faculty is a mine of curious, hungry people with unique adaptability skills. I am sure the future will prove us right.

Hunc igitur terrorem animi tenebrasque necessest
non radii solis neque lucida tela diei discutiant sed
naturae species ratioque. Principium cuius hinc nobis
exordia sumet, nullam rem e nihilo gigni divinitus
umquam.

Lucrezio